Horse genomics and the Dorothy Russell Havemeyer Foundation

Special Issue Co-ordinator
E. Bailey
Horse genomics and the Dorothy Russell Havemeyer Foundation

E. Bailey
MH Gluck Equine Research Center, University of Kentucky, Lexington, KY, USA
E-mail: ebailey@uky.edu

In 1995, the Dorothy Russell Havemeyer Foundation (DRHF) conducted the First International Equine Gene Mapping Workshop in Lexington, Kentucky. Scientists working on horses needed information about the equine genome to better perform their research. The scientific advances of the Human Genome Project, begun in 1990, were providing tools and information that made this goal attainable. The DRHF conducted a series of horse genome workshops between 1997 and 2009 in San Diego, USA; Uppsala, Sweden; Brisbane, Australia; Pretoria, South Africa; Dublin, Ireland; Lake Tahoe, USA; and Newmarket, UK. Over 200 scientists from 20 countries around the world participated in these workshops, sharing ideas and resources, collaborating and creating a common genomics framework. In 2006, the National Human Genome Research Institute (NHGRI) in the USA chose to sequence the horse as a consequence, in part, of both the work done by this community of scientists and their research interests going forward. As a result, the complete DNA sequence of the horse is available online (Wade et al. 2009) and has enabled powerful new research methods for equine science.

This special issue of Animal Genetics was funded by the DRHF to underscore the importance of the genome to horse research. In this issue, scientists report research and discoveries made possible using the new genomic information. Indeed, the work includes gene discoveries and genetic characterization of horse breeds and sheds light on hereditary conditions that affect performance of horses. But the genome information is also useful to understand non-hereditary diseases and traits as well. Several reports in this issue address gene expression in connection with exercise and laminitis. Importantly, this is just the beginning. As scientists become more familiar with using genomic information for equine studies, we can anticipate more discoveries and the development of new diagnostic tests, therapeutic products and management approaches to improve the health and well-being of the horse.

The impact of these studies and the availability of genome information for the horse will be significant at many levels. Good health is a high priority for horses. Owners and breeders train and select horses for a wide variety of characteristics including conformation, athleticism, temperament, intelligence, speed, stamina and coat colour. Poor health can undo years of work on part of the horse owner. Furthermore, the horse industry contributes significantly to economies around the world. Although we no longer depend on horses for power and transportation, horses continue to be an important agricultural product. A 2005 study in the United States suggested that the horse industry had a $102 billion (indirect) impact and provided employment (indirect and induced) to more than 1.4 million people (American Horse Council, 2005). The worldwide economic impact of the horse is obviously much larger.

Use of the genome sequence in research will have far reaching effects on the health of horses and the health of the economy, worldwide. This should be remembered as a legacy from the Dorothy Russell Havemeyer Foundation.

Conflicts of interest
The author has declared no potential conflicts.

References
Genome-wide SNP association–based localization of a dwarfism gene in Friesian dwarf horses


*Section of Veterinary Clinical Studies, School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Dublin, Ireland. †Animal Genomics Laboratory, School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Dublin, Ireland. ‡Department of Equine Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands. §Department of Surgery and Anaesthesiology of the Domestic Animals, Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium. ¶School of Medicine and Medical Sciences, University College Dublin, Dublin, Ireland. **Animal Breeding and Genomics Centre, Wageningen University and Research Centre, Wageningen, The Netherlands. ††Conway Institute of Biomedical and Biomolecular Research, University College Dublin, Dublin, Ireland.

Summary

The recent completion of the horse genome and commercial availability of an equine SNP genotyping array has facilitated the mapping of disease genes. We report putative localization of the gene responsible for dwarfism, a trait in Friesian horses that is thought to have a recessive mode of inheritance, to a 2-MB region of chromosome 14 using just 10 affected animals and 10 controls. We successfully genotyped 34,429 SNPs that were tested for association with dwarfism using chi-square tests. The most significant SNP in our study, BIEC2-239376 ($P_{2df} = 4.54 \times 10^{-5}$, $P_{rec} = 7.74 \times 10^{-6}$), is located close to a gene implicated in human dwarfism. Fine-mapping and resequencing analyses did not aid in further localization of the causative variant, and replication of our findings in independent sample sets will be necessary to confirm these results.

Keywords: dwarfism, Friesian, gene, GWAS, horse, SNP association.

Introduction

The recent completion of a draft sequence of the horse genome (Wade et al. 2009), combined with technological and methodological advances in the analysis of complex genetic traits in humans (Steemers & Gunderson 2007), has provided veterinary genetic epidemiologists with an array of tools for the study of diseases with a heritable component. Specifically, the advent of high-density single nucleotide polymorphism (SNP)-based genotyping arrays and the concomitant growth in knowledge of the haplotypic structure of mammalian genomes has led to the adoption of genome-wide association studies (GWAS) for disease gene mapping (Frazer et al. 2007; Orr & Chanock 2008). These studies, which have recently been successful in the identification of large numbers of genetic loci contributing to disease in humans, exploit the high degree of correlation between genetic variants at any particular region of a chromosome to efficiently interrogate the entire genomes of large numbers of samples using a minimally redundant set of the so-called tagging-SNPs (Howie et al. 2006).

The GWAS paradigm in human studies has been to genotype hundreds of thousands of SNPs in many thousands of affected and unaffected individuals. These large numbers are a result of the hypothesis that multiple loci with a range of effect sizes contribute to the aetiology of complex diseases; large numbers of samples are thus required to attain the requisite statistical power for unambiguous detection of such loci. Charlier et al. (2008) and Karlsson et al. (2007) have demonstrated that association-based mapping can be effective in the detection of single locus recessive traits in animals using much smaller numbers of individuals.
Within the Friesian horse breed, congenital dwarfism has been recognized for many years and occurs at a frequency of 0.25% (Osinga 2000; Back et al. 2008). We recently reported a full phenotypic characterization for Friesian dwarfs, and we refer readers to that article for full details of the features of the trait (Back et al. 2008). Briefly, the Friesian dwarf phenotype results from physeal growth retardation in both limbs and ribs, reflected in a characteristic disproportional growth disturbance. The potential for post-natal growth in these animals, albeit at a reduced rate, is responsible for mature dwarfs having a head of the same size as unaffected animals, a broader chest with narrowing at the costochondral junction, a disproportionally long back and abnormally short limbs. Furthermore, radiographs reveal a dysplastic metaphysis of the distal metacarpus and metatarsus. Light microscopy of growth plates at the costochondral junction demonstrates an irregular transition from cartilage to bone, and thickening and disturbed formation of chondrocyte columns, which is similar to findings in osteochondrodysplasia (Back et al. 2008).

The molecular mechanism underlying the growth disturbance in Friesian foals has yet to be determined, though the trait is heritable and appears to follow an autosomal recessive pattern of transmission (Osinga 2000; Back et al. 2008). Here, we report on the mapping of the genetic determinants of dwarfism in the Friesian horse using GWAS methodologies in a small number of Friesian dwarfs and normal controls.

Materials and methods

Horses

Ten Friesian dwarf horses were selected by the investigators over an 8-year period (2001–2008) based on their unique phenotype (Back et al. 2008) and availability of DNA. Ten normal Friesian horses (controls) were selected in 2008 based on a normal phenotype. In both cases and controls, first- and second-degree relatives were excluded from the study. Pedigrees were verified by microchip and/or passport of the Royal Friesian Horse Studbook (KFPS). Ten millilitres of heparinized blood samples were taken and stored at –80 °C. DNA isolation was performed using QIAamp DNA blood maxi kit from Qiagen according to the manufacturer’s instructions. DNA samples were quantified using Quant-iT PicoGreen dsDNA kits (Invitrogen) according to the manufacturer’s instructions, and the DNA concentrations were adjusted to 20 ng/µl.

Genotyping

Samples were genotyped using EquineSNP50 Genotyping BeadChips (Illumina). This array contains approximately 54 000 SNPs ascertained from the EquiCAB2 SNP database of the horse genome (http://www.broadinstitute.org/mammals/horse) and has an average spacing of 43.2 kb between adjacent variants. Genotyping was performed in-house on an Illumina Bead Station according to the manufacturer’s recommended protocol. The samples that were genotyped for this study were a subset from 96 equine DNA samples that were genotyped in the same batch, thus ensuring an adequate number of samples for genotype cluster seeding. We included 21 pairs of duplicate samples for QC purposes and observed 100% concordance between genotype calls for 19 duplicate pairs and 99.9% concordance in the remaining two pairs. No samples were excluded on the basis of low assay completion rates. We successfully genotyped 54 633 loci; the overall genotype completion rate was greater than 99.9% (no-call rate = 7.7 × 10⁻⁴). We omitted from further analysis 20 204 SNPs that were monomorphic, as well as those that had minor allele frequencies (MAF) less than 5% in our samples, leaving 34 429 SNPs in our working build of the data.

Fine-mapping

We attempted to localize the most significant association by genotyping an additional 319 evenly spaced SNPs from EquiCAB 2.0 across the region spanning from 3.16 to 5.7 MB on chromosome 14. We included an additional 55 SNPs that were present on the EquineSNP50 array for confirmation of genotype calls from the GWAS. Genotyping for fine-mapping was performed by Sequenom and was conducted using iPLEX chemistry on the MassArray platform. Twenty-two SNPs failed to genotype, and the overall call rate of the remaining loci was 94%.

Resequencing

PROP1 was identified in the GWAS as a strong candidate gene for dwarfism. The exons and UTRs were resequenced to search for a possible causative variant. Regions of interest were amplified by polymerase chain reaction using Platinum Taq DNA Polymerase kits (Invitrogen) in the same case–control series as used for the genotyping experiments, according to the manufacturer’s instructions and using M13 sequence tagged target-specific primers (sequences available upon request) designed using Primer3 (Rozen & Skaletsky 2000). DNA sequencing was outsourced to Macrogen Inc and was performed using Applied Biosystems ABI3730XL capillary sequencers. Raw sequence data was processed using Phred/Phrap, version 030415 (Ewing et al. 1998) and was screened for polymorphisms using PolyPhred, version 6.15 (Nickerson et al. 1997). Traces were visualized and manually inspected using CONSED, version 17.0 (Gordon et al. 1998). SNPs were mapped to EquiCAB 2.0 using BLAT (Kent 2002) and were classed as novel if they did not appear in the EquiCAB 2.0 SNP collection.
Statistics

We compared genotype frequencies in cases and controls, testing for disease association using chi-square tests with two degrees of freedom. Where appropriate, we modelled genetic effects by minor allele using one degree of freedom chi-square tests. SNPs in the fine-mapping analysis were analysed using the Cochran–Armitage test to identify trends. As the study was exploratory in nature, we report P-values uncorrected for multiple testing, as is conventional in GWAS. For reference, however, the Bonferroni corrected α-level (corresponding to a study-wide α-level of 5%) for the current study is 1.72 × 10^-6. All statistical analyses were conducted using PLINK, GLU and R (Purcell et al. 2007; R Development Core Team 2009). Plots of regional association were generated in R using a modified version of code available at http://www.broadinstitute.org/mpg/snap/index.php.

Results

The GWAS results, sorted by chromosome, are shown in Fig. 1. We observed a peak of association on chromosome 14, the best SNP being BIEC2-239376 (P = 4.54 × 10^-5). We further investigated the association at this locus by fitting specific genetic models. The significance of association at BIEC2-239376 was greatest under a recessive model (P = 7.74 × 10^-6). All 10 of the dwarfs in the GWAS were TT homozygotes, while the controls were comprised of four CC homozygotes and six heterozygotes. BIEC2-239376 was the most significant SNP in a cluster of SNPs at the p-arm of chromosome 14. We assessed the independence between BIEC2-239376 and other chromosome 14 loci by calculating the correlation coefficient, r^2, between the SNP with the strongest signal and all other SNPs on the chromosome (Fig. 2). We reanalysed each chromosome 14 locus using logistic regression, conditioned on BIEC2-239376, and concluded that linkage disequilibrium with BIEC2-239376 is the main driver of association at the other loci.

We genotyped 374 SNPs, of which 319 were not on the GWAS array, in the region of association on chromosome 14. Of these, 184 were either monomorphic or had low MAF in the study population. From the remaining 190 SNPs, BIEC2-250663 (P = 4.94 × 10^-5) was more significantly associated with dwarfism than BIEC2-239376 (P = 6.12 × 10^-5) (Fig. 3); dwarfs were fully homozygous for both markers, and the homozygous genotype observed in dwarfs was not present in controls. An additional SNP, BIEC2-249929 was in perfect linkage disequilibrium with BIEC2-239376.

BIEC2-239376 is located 34 kb from the gene encoding the equine homolog of the prophet of PIT1, paired-like homeodomain transcription factor protein, PROP1 (Fig. 3). Inactivating mutations in PROP1 are known to cause dwarfism in humans via combined pituitary hormone deficiency. We therefore selected the PROP1 gene for targeted resequencing of each exon, intron–exon boundary and both UTRs in our sample set. We detected a total of nine SNPs (Table 1), all of which we believe to be novel based on their absence from the EquCab 2.0 SNP database. Each novel SNP was rare, and many were singletons; none were associated with the dwarfism phenotype. Thus, resequencing

![Figure 1](image1.png)  
**Figure 1** Manhattan plot of P-value in the Friesian horse dwarfism GWAS. Association of 34 429 SNPs with dwarfism represented by −log_{10} P-values from a two degree of freedom chi-square test plotted by chromosome and sorted by chromosomal position. No SNPs in the GWAS remained statistically significant after correction for multiple testing.

![Figure 2](image2.png)  
**Figure 2** Association of chromosome 14 SNPs with Friesian horse dwarfism. Association plot of chromosome 14 SNPs with dwarfism, with chromosomal position on the x-axis and −log_{10} P-value on the y-axis. The most significant chromosome 14 SNP, BIEC2-239376, is indicated by a red diamond. Each square on the plot represents a single SNP; the colour of each square represents the strength of correlation between that SNP and BIEC2-239376, with dark red indicating an r^2 of 1 and white representing an r^2 of 0.
analysis of the protein-coding components of PROP1 failed to detect a putative functional variant.

**Discussion**

High-density SNP-based genotyping arrays have been successfully used to identify genetic variants contributing to both common and rare diseases in humans and animals (Charlier et al. 2008; Manolio et al. 2008). Here, we report the application of SNP genotyping arrays for mapping dwarfism in horses. Using a small panel of affected and unaffected Friesian horses, we have identified a region of association on chromosome 14 that is strongly associated with the dwarf phenotype.

GWAS exploit the underlying correlations between genetic variants that initially arise via mutation on a common genetic background that is gradually eroded by recombination events through successive generations. In humans, complex population structure means that correlation between loci may diminish relatively quickly and thus exist only over short distances. In many breeds of animals, however, selective breeding strategies and small effective population sizes mean that pairwise correlations between loci may be strong over large regions and on less complex backgrounds. Thus, a much smaller number of genetic markers should be required to effectively capture information on untyped variants. The drawback of long range linkage disequilibrium in the context of disease mapping is that localization of causal mutations is difficult, because the associated regions are often very large. In this study, the associated region on chromosome 14 exceeds two megabases in length, peaking with BIEC2-239376 at 3 761 355 bp. Fine-mapping using a denser panel of SNPs failed to resolve the situation: we identified two additional SNPs, BIEC2-250663 and BIEC2-239376, which were located at 3 761 254 and 5 418 619 bp, respectively. No other genotyped SNP in the region was fully homozygous, as would be expected for a recessively inherited trait, in the dwarfs. This study therefore serves as an example of the difficulty in overcoming LD for fine-mapping in inbred populations.

**Table 1** Sequences of nine novel SNPs identified by resequencing PROP1.

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>5’ flanking seq</th>
<th>Alleles</th>
<th>3’ flanking seq</th>
<th>SNP Position (bp)</th>
<th>Location</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROPI_1</td>
<td>AAGGACACCG</td>
<td>C/T</td>
<td>CAACCACAAA</td>
<td>3773707</td>
<td>5’ UTR</td>
<td></td>
</tr>
<tr>
<td>PROPI_2</td>
<td>CACAAAAAAAA</td>
<td>A/C</td>
<td>CACATCCCAAG</td>
<td>3773692</td>
<td>5’ UTR</td>
<td></td>
</tr>
<tr>
<td>PROPI_3</td>
<td>AAAAGGGGGA</td>
<td>C/T</td>
<td>GCTGCCCTCCT</td>
<td>3773583</td>
<td>Exon 1</td>
<td>Arg/Cys</td>
</tr>
<tr>
<td>PROPI_4</td>
<td>CTTGAGAGGC</td>
<td>C/G</td>
<td>AGAGGCCCCG</td>
<td>3773447</td>
<td>Intron 1</td>
<td></td>
</tr>
<tr>
<td>PROPI_5</td>
<td>CTTTGTTTCCC</td>
<td>G/T</td>
<td>CGCAAGGGCC</td>
<td>3725488</td>
<td>Exon 2</td>
<td></td>
</tr>
<tr>
<td>PROPI_6</td>
<td>GGAGTGCCG</td>
<td>C/T</td>
<td>GACCCCGTCTG</td>
<td>3725604</td>
<td>Exon 2</td>
<td>Arg/His</td>
</tr>
<tr>
<td>PROPI_7</td>
<td>GTCCTCCCTTG</td>
<td>C/T</td>
<td>GGGGAAGGCC</td>
<td>3725624</td>
<td>Exon 2</td>
<td></td>
</tr>
<tr>
<td>PROPI_8</td>
<td>CTCCTTCTCCT</td>
<td>T/C</td>
<td>CAGCCCTCTCA</td>
<td>3724500</td>
<td>Exon 4</td>
<td></td>
</tr>
<tr>
<td>PROPI_9</td>
<td>CTGTCGCCCCC</td>
<td>C/T</td>
<td>GTAAAACACA</td>
<td>3724306</td>
<td>3’ UTR</td>
<td></td>
</tr>
</tbody>
</table>

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 2–7
dwarfism in humans. We identified the equine homolog of \textit{PROP1} as being a strong positional candidate gene for dwarfism in Friesians. Inactivating mutations in \textit{PROP1} cause disruption of the growth hormone (GH) axis and are manifested as a dwarfism phenotype. Central pituitary underproduction of GH leads typically to proportional dwarfism, as seen in the German Shepherd dog (Andresen & Willeberg 1976; Hanson et al. 2006). In the Friesian dwarf horses, however, a disproportional growth disturbance is seen, which would imply a local defect or disturbance in one of the regulatory systems for growth plate development. Growth plate development is under the control of many autocrine and paracrine factors (Kronenberg 2003). Recently, de Graaf and colleagues investigated the functioning of the hypothalamic-pituitary growth axis in three Friesian dwarfs. No evidence of hypothalamic-pituitary dysfunction or failure of IGF-1 production was found, suggesting that the cause of the congenital growth abnormality was located distal or peripheral to the level of the GH receptor in the liver and may have been a defect in a peripheral IGF-1 or GH receptor or may not involve the GH-IGF-1 axis at all (de Graaf-Roelfsema et al. 2009). Concomitant with this observation, we did not detect coding polymorphisms in \textit{PROP1} that were associated with dwarfism in our Friesian horse samples.

When the normal bone remodelling process is disturbed in horses, abnormal defects in the growth plate may result. A local defect or disturbance in one of the regulatory systems for growth plate development potentially can result in a disproportional growth disturbance as is typically seen in Friesian dwarf horses (Vaughan 1976; Jeffcott & Henson 1998; Gee et al. 2005). Further screening of the OMIM database for genes in the chromosome 14 region that could be linked to a distorted bone remodelling process in abnormal growth plate development identified \textit{ZNF346}, \textit{COL23A1} and \textit{B4GALT7} as candidate genes. \textit{ZNF346} is proposed to play a role in apoptosis (cell death), a process crucial for the normal transition of cartilage into bone seen during normal physeal growth (Gibson 1998; Ballock & O’Keefe 2003). It could be speculated that disturbed apoptosis plays a role in the physeal growth retardation typically seen in Friesian dwarfism. Both \textit{COL23A1} and \textit{B4GALT7} are proposed to play a role in collagen network formation. Although a disturbed collagen network is known to effect calcification and subsequent transformation of cartilage into bone (Wassen et al. 2000), the specific role of both genes in collagen formation is highly speculative and largely unclear. \textit{B4GALT7} is also proposed to play a role in connective tissue disorders and has been related to disturbed fibril organization and proteoglycan synthesis. Both processes could play a role in the abnormal development of bone and subsequent retardation of growth plate growth that are observed in Friesian dwarfism (Kvist et al. 2006; Burdan et al. 2009).

Finally, based on the equine build from UCSC Broad Institute (UCSC Genome Browser: http://genome.ucsc.edu), both \textit{FGFR1} and \textit{FGFR2} are included within the critical region of chromosome 14 flanked by BIEC2-249929 and BIEC2-250663. Both fibroblast growth factor receptors play key roles in skeletal development, and mutations have been related to skeletal dysplasia and dwarving syndromes (Esvarakumar et al. 2002; White et al. 2005). Profound effects on bone elongation have been shown through supposed suppression of chondrocyte and osteoblast function. However, in contrast to the relative normal head proportions seen in the Friesian dwarf syndrome, \textit{FGFR1} and \textit{FGFR2} mutations seem to have a significant effect on flat bone growth and skull formation.

Using an agnostic genome-wide approach, we have identified a putative region that may harbour a gene for dwarfism in Friesian horses. Validation of this finding in a larger group of animals and segregation analysis in known pedigrees is warranted before further localization of the causative mutation is conducted. This may prove extremely challenging given the strong linkage disequilibrium observed in the region. This study suggests that, with the advent of new genomic tools, studies of equine diseases may yield important new insight into pathogenesis and may be translatable to orthologous human traits.

Acknowledgements

We would like to thank the equine practitioners for submitting the dwarf cases and the studbook (http://www.fps-studbook.com) for their technical assistance. This material is based upon works supported by the Science Foundation Ireland under the Research Frontiers Programme (Foundation Grant No. R10526).

Conflicts of Interest

The authors have declared no potential conflicts.

Web URLs

\texttt{GLU: http://code.google.com/p/glu-genetics/}
\texttt{PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/}
\texttt{SNAP: http://www.broadinstitute.org/mpg/snap/}
\texttt{EQUCan2.0: http://www.broadinstitute.org/ftp/distribution/horse\_snp\_release/v2/}
\texttt{UCSC Genome Browser: http://genome.ucsc.edu/}

References


Linkage disequilibrium and historical effective population size in the Thoroughbred horse

L. J. Corbin*, S. C. Blott†, J. E. Swinburne†, M. Vaudin†, S. C. Bishop* and J. A. Woolliams*

*Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin Biocentre, EH25 9PS, UK. †Animal Health Trust, Newmarket, CB8 7UU, UK

Summary

Many genomic methodologies rely on the presence and extent of linkage disequilibrium (LD) between markers and genetic variants underlying traits of interest, but the extent of LD in the horse has yet to be comprehensively characterized. In this study, we evaluate the extent and decay of LD in a sample of 817 Thoroughbreds. Horses were genotyped for over 50 000 single nucleotide polymorphism (SNP) markers across the genome, with 34 848 autosomal SNPs used in the final analysis. Linkage disequilibrium, as measured by the squared correlation coefficient ($r^2$), was found to be relatively high between closely linked markers (>0.6 at 5 kb) and to extend over long distances, with average $r^2$ maintained above non-syntenic levels for single nucleotide polymorphisms (SNPs) up to 20 Mb apart. Using formulae which relate expected LD to effective population size ($N_e$), and assuming a constant actual population size, $N_e$ was estimated to be 100 in our population. Values of historical $N_e$, calculated assuming linear population growth, suggested a decrease in $N_e$ since the distant past, reaching a minimum twenty generations ago, followed by a subsequent increase until the present time. The qualitative trends observed in $N_e$ can be rationalized by current knowledge of the history of the Thoroughbred breed, and inbreeding statistics obtained from published pedigree analyses are in agreement with observed values of $N_e$. Given the high LD observed and the small estimated $N_e$, genomic methodologies such as genomic selection could feasibly be applied to this population using the existing SNP marker set.

Keywords effective population size, horse, linkage disequilibrium, population history, single nucleotide polymorphism, Thoroughbred.

Introduction

Linkage disequilibrium (LD) describes the non-random association of alleles at different loci and can result from processes such as migration, selection and genetic drift in finite populations (Wang 2005). The efficacy of genomic techniques such as genome-wide association studies (GWAS), marker-assisted selection (MAS) and genomic selection is dependent on the extent of LD and its rate of decline with distance between loci within the population under study. The recent release of the Illumina Equine SNP50 Genotyping BeadChip has increased the potential for such techniques to be applied to the horse. Researchers have already begun to make use of the single nucleotide polymorphism (SNP) chip in GWAS (Bannasch et al. 2009; Blott et al. 2009; Lykkjen et al. 2009), and the opportunity exists to use validated loci for MAS in the future, as some success has already been seen in the localization of QTL for simple diseases (Drögemuller et al. 2009; Eberth et al. 2009; Gabreski et al. 2009). However, as has been shown in human studies, when applied to complex diseases, the outcomes of GWAS are generally less successful (Manolio et al. 2009), and thus the genomic selection techniques of Meuwissen et al. (2001) may become more attractive. The opportunities will depend on the extent of LD, and therefore the characterization of LD exhibited with the current SNP50 BeadChip will assist in...
planning future studies of complex traits and in the development of genomic tools.

Linkage disequilibrium structure can also provide insights into the evolutionary history of a population. The strength of LD at different genetic distances between loci can be used to infer ancestral effective population size ($N_e$), where $N_e$ is the number of individuals in an idealized population that would give rise to the same rate of inbreeding as observed in the actual breeding population (Falconer & Mackay 1996). Deterministic equations derived by Daetwyler et al. (2009) show that, once the $N_e$ for a population is known, the accuracy of genomic selection for a range of scenarios can be calculated. The pattern of historical $N_e$ in domestic livestock populations can also help us to understand the impact of selective breeding strategies on the genetic variation present in populations and can provide an insight into the level of inbreeding in populations for which pedigrees are incomplete or unavailable.

The pattern of LD in the Thoroughbred has yet to be comprehensively characterized, and predictions of $N_e$ are limited to those inferred from pedigree data, which itself may be inaccurate (Hill et al. 2002). An early study by Tozaki et al. (2005), based on 300 horses, concluded that useful LD in the Thoroughbred extends up to 7 cM, but this study covered only one small region of the genome. More recently, Wade et al. (2009) investigated LD across ten 2-Mb regions in a number of different horse breeds, using sample sizes of 24 horses per breed. In contrast, genome-wide LD in livestock populations has been the focus of numerous studies (McRae et al. 2002; Heifetz et al. 2005; Khatkar et al. 2008). Studies have also been carried out to evaluate the historical $N_e$ of a variety of cattle breeds, all of which suggest a continuous decrease in $N_e$ since the time of domestication (Thévenon et al. 2007; de Roos et al. 2008; Qanbari et al. 2009).

The objective of this study was to characterize LD in a large sample of Thoroughbred horses using data generated from the Illumina Equine SNP50 BeadChip and to consider the results in the context of genomic methodologies. The decline of LD over distance is used to predict the effective population size both assuming a constant population size and assuming linear growth. These results are considered in the context of current knowledge of the establishment of the Thoroughbred breed.

Materials and methods

Genotypic data

The data for this study originated from two disease association studies, and the dataset comprises genotype data for 817 UK Thoroughbreds. Whilst the original data collection required horses to be categorized as cases or controls for the diseases of interest, for the purpose of this study, the horses were treated as a single population sample. Blood samples were collected in EDTA, and DNA was extracted either by Tepnel (http://www.tepnel.com/dna-extraction-service.asp) or at the AHT using Nucleon BACC DNA extraction kits (http://www.tepnel.com/dna-extraction-kits-blood-and-cell-culture.asp). A small dilution of each sample was prepared at 70 ng/ul and submitted for genotyping to Cambridge Genomic Services (http://www.cgs.path.cam.ac.uk/services/snp-genotyping/services.html). The Illumina Equine SNP50 Genotyping BeadChip (http://www.illumina.com/documents/products/datasheets/datasheet_equine_snp50.pdf) was used. This comprises 54 602 single nucleotide polymorphisms (SNPs) located across all autosomes and the X chromosome. These were selected from the database of over one million SNPs (http://www.broadinstitute.org/ftp/distribution/horse_snp_release/v2/) generated during the sequencing of the horse genome (http://www.broadinstitute.org/mammals/horse).

Genotyping data was analysed using the Illumina GenomeStudio genotyping module, and a series of quality control metrics were used to identify poorly performing SNPs. Quality control (QC) at this stage led to the removal of 7.1% ($n = 3895$) of SNPs from the analysis owing to poor genotyping quality (see Table S1). Further QC carried out as part of this study led to the removal of an additional 21 SNPs, which were genotyped in less than 95% of samples. The genotyping rate once these exclusions had been made was greater than 99%, with no individuals having more than 10% of SNPs missing. Markers which deviated significantly from Hardy–Weinberg equilibrium (HWE) ($P < 0.0001$) were excluded (Purcell et al. 2007; Purcell 2009). Previous studies have demonstrated that including markers with low minor allele frequencies (MAF) can bias LD estimates (Goddard et al. 2000; Qanbari et al. 2009; Toosi et al. 2010), therefore a MAF threshold of 0.10 was imposed on the data. Outcomes of the HWE and MAF screening are given in the results. This study used only autosomal markers.

Linkage disequilibrium

The measurement of LD used throughout this study is the squared correlation coefficient between SNP pairs ($r^2$) (Hill & Robertson 1968), computed as:

$$r^2 = \frac{D^2}{p_A p_B p_a p_b},$$

where $D = p_{AB} p_a p_b$ and $p_A$, $p_B$, $p_a$ and $p_b$ are the frequencies of alleles A, a, B and b, respectively. An EM algorithm (Weir 1996) was implemented to estimate haplotype frequencies. $r^2$ was calculated (to four decimal places) for all syntenic marker pairs. Individuals with a missing genotype for a given marker were excluded when calculating LD for that marker. Details of the physical position of the markers can be found in Illumina product literature (http://www.illumina.com/documents/products/marker_lists marker_list_equineSNP50.xls). To accommodate the large range of
marker distances observed and to enable clear presentation of results showing LD in relation to physical distance between markers, SNP pairs were divided into three distance classes and subsequently put into 87 distance bins, with bin ranges dependent on the class (see Table S2). The mean $r^2$ for each of the distance bins was then plotted against the median of the distance bin range (Mb). This analysis was carried out on a chromosome by chromosome basis; the pooled results are presented here. $r^2$ was also calculated for a random selection of non-syntenic markers. Thirty SNPs per autosome were randomly selected, and $r^2$ values were calculated for all non-syntenic markers, resulting in a total of 418 500 pairwise comparisons.

Modelling of decline of linkage disequilibrium with distance

Under the assumption of an isolated population with random mating, Sved (1971) derived an approximate expression for the expectation of $r^2$:

$$E(r^2) = (1 + 4Nc)^{-1},$$

where $N$ is effective population size, and $c$ is the recombination frequency. In this paper, as in previous studies (Hayes et al. 2003; Tenesa et al. 2007; Thévenon et al. 2007; de Roos et al. 2008; Qanbari et al. 2009; Villa-Angulo et al. 2009), $c$ is replaced by map distance in Morgans. This is justified by the approximation of the more precise equation for $E(r^2)$ given by Sved (1971), where $(1-c/2)(1-c)^{-2}$ replaces $c$. This function is a reasonable approximation to both Haldane and Kosambi map distance for $0 \leq c \leq 0.5$. Based on this formula, a non-linear least squares approach to statistically model the observed $r^2$ was implemented within R (R Development Core Team 2009) using the following model:

$$y_i = 1/(a + 4bd_i) + e_i,$$

where $y_i$ is the value of $r^2$ for SNP pair $i$, at linkage distance $d_i$ in Morgans. Parameters $a$ and $b$ were estimated iteratively using least squares. Chromosome-specific megabase-to-centimorgan conversion rates were calculated based on total physical chromosome length, as stated on the NCBI website (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9796) and total chromosome genetic length from the equine linkage map (Swinburne et al. 2006) (see Table S3). Marker pairs with <100 bp between them were excluded from this analysis, as it has been suggested that Sved’s (1971) model is not appropriate for very small values of $c$ (Hill 1981; de Roos et al. 2008), and at small distances gene conversion contributes to the breakdown of LD (Frisse et al. 2001; Ardlie et al. 2002; Tenesa et al. 2007). The minimum MAF threshold of 0.10 was also applied here, as Eq. 2 may be a poor approximation when allele frequencies are close to zero (Hill 1981; Hudson 1985). This model was applied to data for each autosome in turn, and parameter estimates were combined by meta-analysis in R (R Development Core Team 2009) using an inverse variance method for pooling and a random effects model based on the DerSimonian–Laird method (DerSimonian & Laird 1986) (see Appendix S1 for further details).

Ancestral effective population size estimation

Rearrangement of Eq. 2 allows the prediction of effective population size at a given point in time, expressed as generations in the past (Hayes et al. 2003; de Roos et al. 2008):

$$N_T(t) = (4c)^{-1} \left[ (r_T^2)^{-1} - 1 \right],$$

where $N_T$ is the effective population size $t$ generations ago, $c$ is the distance between markers in Morgans, $r_T^2$ is the mean value of $r^2$ for markers $c$ Morgans apart, and $c = (2c)^{-1}$ when assuming linear growth (Hayes et al. 2003). As previously mentioned, marker pairs with less than 100 bp between them and SNPs with MAF <0.10 were excluded from this analysis. To compute $N_T$, the number of prior generations was selected and a suitable range for $c$ was calculated (see Table S4). The binning process was designed to ensure sufficient SNP pair comparisons within each bin to get a representative estimate of $r^2$. The mean distance and mean $r^2$ between marker pairs in each bin was then computed. This process was carried out for each chromosome in turn and also for markers pooled across chromosomes, as is suggested by Hayes et al. (2003) to reduce the variability of estimates of $N_T$ caused by finite population size.

**Results**

Genotypic data

Of the 52 603 autosomal SNPs genotyped, 34 848 (66.2%) remained after filtering, resulting in more than 20 million pairwise comparisons. Of those SNPs excluded, 173 SNPs were excluded for not being in HWE and a further 13 372 for having a MAF <0.10 (4086 of these were in fact monomorphic in our sample). The number of SNPs per autosome remaining after exclusions ranged from 416 to 2,760 and was closely related to chromosome length, as

![Figure 1](image-url)
shown in Fig. 1. The average distance between adjacent markers (±SD) was 64.05 ± 86.84 kb, with the distance between adjacent SNPs ranging from 1 bp to 2849 kb. The MAF of remaining SNPs followed a uniform distribution and averaged (±SD) 0.30 ± 0.12.

Linkage disequilibrium

Linkage disequilibrium declined with increasing distance between SNP pairs, as shown in Fig. 2a, b and c. The most rapid decline was seen over the first 0.2 Mb, with the mean \( r^2 \) decreasing by more than half over this period. The mean \( r^2 \) then decreased more slowly with increasing distance, and the decline in LD was almost linear with log-transformed distance (Fig. 3). The coefficient of variation of \( r^2 \) increased from 0.6 at 5 kb to a maximum of 2.2 at 15 Mb, subsequently decreasing and remaining below 1.5 for distances greater than 50 Mb. A total of 10 130 SNP pairs were in complete LD (\( r^2 = 1 \)); 5139 of these were adjacent pairs. The mean (±SD) \( r^2 \) between random non-syntenic markers was 0.0018 ± 2.49 × 10^{-3} and was similar to that observed between syntenic markers at distances greater than 100 Mb (Fig. 2c).

Modelling of decline of linkage disequilibrium with distance

The non-linear regression modelling of the decline of LD with distance resulted in both \( a \) and \( b \) being significantly different from zero. The mean estimate and 95% confidence interval by meta-analysis across autosomes for \( a \) was 2.25 [2.18; 2.33] and for \( b \) was 103.1 [95.8; 110.3]. The line of predicted \( r^2 \) from the non-linear regression equation only approximately follows that of the mean observed \( r^2 \), with the greatest discrepancy occurring at distances less than 0.03 Mb, as shown in Fig. 3. Parameter \( b \) showed greater variability between chromosomes than parameter \( a \), although estimates for both parameters showed an approximately symmetrical distribution about the median. A significant negative correlation \( (P < 0.01) \) was observed between estimates of \( b \) and chromosome length (cM), but there was no such relationship between estimates of \( a \) and chromosome length (cM) (Fig. 4a, b). The interpretation of \( b \) as an estimate of effective population size is considered in the discussion.

Ancestral effective population size

We observed an initial pattern of decreasing \( N_e \) with values of over 3000 estimated in the distant past (see Figure S1) and values closer to 100 estimated at 20 generations ago (Fig. 5). Our results suggest that an increase in \( N_e \) has occurred over the past ten generations, with a maximum of approximately 190 observed two generations ago. Variation in predicted \( N_e \) across chromosomes was greatest for estimates corresponding to the most recent ten generations and

---

Figure 2 Average linkage disequilibrium (solid line) and the 25th and 75th percentiles (dashed lines) (measured by \( r^2 \)), plotted against the median of the distance bin range (Mb). Percentiles Calculated for each bin in turn. (a) Distance range from 0 to 0.5 Mb. \( r^2 \) values averaged using bins of 0.01 Mb and pooled over autosomes (minimum of 4900 SNP pairs per distance bin). (b) Distance range from 0.5 to 20.5 Mb. \( r^2 \) values averaged using bins of 1.0 Mb and pooled over autosomes (minimum of 196 000 SNP pairs per distance bin). (c) Distance range from 20.5 to 190 Mb. \( r^2 \) values averaged using bins of 10.0 Mb and pooled over autosomes (minimum of 4375 SNP pairs per distance bin).

Figure 3 Predicted \( r^2 \) versus observed \( r^2 \) against mean distance between markers (cM) (on a log scale). Predicted \( r^2 \) calculated using Eq. 3 with \( a = 2.25 \) and \( b = 103 \).
Discussion

This study provides an overview of LD in the Thoroughbred using a high density SNP panel. Validation work by Khatkar et al. (2008) on their cattle data suggests that our sample size of more than 800 horses is more than sufficient to obtain an unbiased picture of LD in our population. The pattern of decline of LD with distance in this population is consistent with that reported by Wade et al. (2009) in a sample of 24 Thoroughbreds, with both data sets exhibiting a decrease in \( r^2 \) from \(~0.6\) to \(0.2\) when the distance between markers is increased to \(0.5\) Mb. The LD observed is higher at short distances and more extensive than that observed in human populations (Shifman et al. 2003). Linkage disequilibrium declines more slowly in our population than in the range of cattle populations studied by de Roos et al. (2008), with \( r^2 \) remaining above \(0.3\) for distances up to \(185\) kb in our data, compared with a maximum distance of \(35\) kb in the cattle data.

The mean value of \( r^2 \) between non-syntenic SNPs was \(0.0018\), and this provides an approximation of the LD that can be expected by chance, assuming that the markers used have not undergone simultaneous selection. The value observed here is lower than, but of a similar magnitude to, that observed by Khatkar et al. (2008) in a sample of over \(1500\) cattle \((0.0032)\). The mean non-syntenic \( r^2 \) value reflects both sampling of animals and genetic sampling (drift), and hence may be expected to decrease with increases in both sample size and \(N_e\). Therefore, the larger non-syntenic value in Australian Holstein–Friesian cattle may more reflect a lower \(N_e\) in this cattle population. The low LD seen between non-syntenic SNPs in our population suggests that the LD created by admixture during breed formation (Hill et al. 2002) has declined to negligible levels for these markers. A similar decline of LD between non-syntenic markers was observed in Coopworth sheep approximately ten generations after the foundation of the breed through crossing (McRae et al. 2002). At distances greater than \(100\) Mb, average \( r^2 \) between syntenic SNPs is reduced to non-syntenic or background levels, and is no longer a function of distance. This is expected, as the recombination rate at such distances approaches \(0.5\).

By using Sved’s (1971) formula for the expectation of \( r^2 \), a non-linear regression model was fitted to the data to describe the relationship between linkage distance and LD. Without making any assumptions about the value of \( r^2 \) at the intercept, estimates of \(a\) and \(b\), as predicted using Eq. \(3\) and averaged over all autosomes, were \(2.25\) and \(103\), respectively. Parameter \(a\) determines the value of expected \( r^2 \) when the line crosses the \(y\)-axis (i.e. when the distance between markers is effectively zero). Our estimate of \(a\) supports an alternative version of Sved’s (1971) equation, derived by Tenesa et al. (2007), which takes into account mutation and puts \(a\) equal to two, whilst at the same time raising the question of whether fixing \(a\) to unity in the model (as in Abasht et al. (2009), Toosi et al. (2010) and Zhao et al. (2005)) is appropriate. The impact of such model assumptions are explored in Corbin et al. (2010). The heterogeneity of variance associated with the observed \( r^2 \), such that the variance of \( r^2 \) declined with increasing distance between markers, may also have impacted on our results. We observed a significant negative relationship between chromosome length \((cM)\) and estimates of \(b\) from the non-linear model, suggesting LD is higher in longer chromosomes. This contrasts with the findings of Tenesa et al. (2007), who observed a positive relationship, but is in
keeping with the observations of Khatkar et al. (2008) and Muir et al. (2008) in domestic livestock species.

Our estimate of $b$ (103) is an estimate of $N_e$ assuming constant population size. However, this assumption is difficult to sustain, and therefore, $b$ more likely represents a conceptual average $N_e$ over the period inferred from the marker distance range, for example see Toosi et al. (2010). For this reason, Fig. 5 shows the results following the approach of Hayes et al. (2003) by calculating historical $N_e$, assuming linear population growth. The pattern observed shows a decrease in $N_e$ up until around 20 generations ago, followed by an increase until one generation ago. The interpretation of such trends is difficult, with the observed dip in $N_e$ potentially representing any one of a number of scenarios, including a founder event, an immigration event, a hybridization event or any combination of these (Wang 2005). Therefore, it is useful to consider our observation in the context of what is known about the Thoroughbred’s demographic history.

Documentary evidence suggests that the Thoroughbred was derived from a cross between sires originating from the Mediterranean Middle East and British native breeds, and the breed was established during the seventeenth century (Hill et al. 2002). It is not clear from published literature what effects an admixture like this would have on patterns of estimated $N_e$ prior to the crossing event, although clues may be observed in Toosi et al. (2010). However, what may be predicted is that such a crossing event would appear as a bottleneck in the population, creating an initially high level of LD in the beginning. Therefore, one might infer from our results that the lowest point of the curve reflects the point at which the breed was formed; this approximately coincides with the findings of Mahon & Cunningham (1982) that Thoroughbreds born in the 1960s were separated from seventeenth century founders by an average of 21.5 generations. Cunningham et al. (2001) also found evidence for a population bottleneck at the time of breed formation.

The reliability of this method depends both on the technical implementation (Corbin et al. 2010) and, as discussed above, on the demographic history of the breed. Some calibration of the accuracy of the $N_e$ profile presented can be obtained by comparison with values obtained from pedigree analyses. For example, Cunningham et al. (2001) calculate the effective number of studbook founders of the Thoroughbred to be 28.2. As this relies on calculating the long-term contributions of the founders, quantitative genetic theory (Woolliams & Bijma 2000) suggests that the $N_e$ for this generation is twice this value if in HWE, providing an estimate of 56 soon after breed formation. This may be compared with the minimum $N_e$ of 88 obtained in this analysis, which gives fair agreement. A further estimate of reliability can be obtained by comparing the mean inbreeding of 0.125 (SE 0.005) obtained by Mahon & Cunningham (1982) for the 21.5 generations from breed foundation to 1964 with the accumulated inbreeding for generations four to 25 (assuming four generations since 1964) using $1 - \prod_{i=1}^{25} (1 - 1/2N_e)$, with $N_e$ being estimated from Fig. 5. The value obtained of 0.112 is remarkably close. Therefore, our minimum of $N_e = 90$ is of the correct magnitude, and the increase in $N_e$ over the last ten generations may be explained by an increase in actual population size. In Thoroughbreds, with low reproductive rate of the mare and the ban upon use of artificial insemination, there is a greater likelihood that increases in census size will be translated into effective population sizes. The trend in $N_e$ observed in the most recent generations should be interpreted with caution because of the technical limitations of the methods.

Implications for genome-wide association studies, marker-assisted selection and genomic selection

The extent of LD in a population can be used to estimate the SNP density required for GWAS studies to be effective, as well as giving some indication as to the likely precision with which the QTL region will be located. The required sample size is said to be inflated by $1/r^2$, when it is necessary to rely on marker-QTL LD, rather than on the QTL itself (Du et al. 2007), and this has prompted authors to propose thresholds for useful LD. The term ‘useful LD’ has been described as the proportion of QTL variance explained by a marker (Zhao et al. 2005), and the consensus is that an average $r^2 > 0.3$ will permit reasonable sample sizes to be employed for GWAS (Ardlie et al. 2002; Du et al. 2007; Khatkar et al. 2008). In this dataset, markers 185 kb apart achieve an average LD of $r^2 = 0.3$, and this corresponds to approximately 14 500 evenly spaced markers across the genome. However, because markers with $r^2 = 1$ will likely be excluded in genomic selection, and given the high variability of $r^2$ values at small distances, this is likely to be an underestimation of the actual number of SNPs needed. Indeed, in this study, whilst markers separated by less than 250 kb had a mean $r^2$ of 0.32 (after the exclusion of those pairs in complete LD), less than half the SNP pairs exhibited $r^2$ values of greater than 0.3. With MAS also relying on close and consistent linkage between markers and QTL, the high LD observed here is promising. Genomic selection (GS) appears to be effective at lower average $r^2$ than that required for GWAS, with simulation results demonstrating accuracies of up to 0.65 with an average $r^2$ between adjacent SNPs as low as 0.2 and a trait heritability of 0.1 (Calus et al. 2008). Deterministic equations derived by (Daetwyler et al. 2009) demonstrate that the accuracy of GS can be expressed as a function of the effective number of loci ($M_e$) in a population. $M_e$ relates to the number of independent chromosome segments and, given our current $N_e$ estimate of ~180 and assuming a random mating population, the $M_e$ for our population is ~1500 (Meuwissen 2009). Thus, we
are now able to predict the potential accuracy of GS in this population for a range of scenarios.

In summary, we used dense SNP genotype data to characterize LD and make inferences regarding ancestral Ne for a large sample of Thoroughbred horses. In the population studied, LD extended for long distances, reaching baseline levels at around 50 Mb. From the decay in LD with distance, we inferred ancestral Ne and observed a decrease in Ne since the distant past, which reached a minimum of ~90 generation ago, followed by an increase until the present time. Such an approach could be used to investigate the demographic histories and rates of inbreeding of horse breeds with less extensive pedigree records than the Thoroughbred. The results indicate that genomic methodologies which are reliant on LD between markers and QTL have the potential to perform well within Thoroughbred populations genotyped for the 50-K SNP chip.

Acknowledgements

The authors are financially supported by the British Equestrian Federation, the Biosciences Knowledge Transfer Network and the Biotechnology and Biological Sciences Research Council (BBSRC). Genotyping was funded by the Horserace Betting Levy Board and the Thoroughbred Breeders’ Association. LJC would like to thank W.G. Hill and I.M.S. White for helpful discussions.

Conflicts of interest

L.J. Corbin received a bioscience CASE Studentship PhD. S.C. Blott received grants from Horserace Betting Levy Board, The Kennel Club. S.C. Bishop received grants from Quality Meat Scotland, EBLEX. J.A.W. received grants from Aviagen, Cobb, Quality Meat Scotland, EBLEX. The remaining authors have no conflicts to declare.

References

Abasht B., Sandford E., Arango J. et al. (2009) Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations. BMC Genomics 10(Suppl. 2), S2.


Linkage disequilibrium in the Thoroughbred horse


Supporting information

Additional supporting information may be found in the online version of this article.

Figure S1 Average estimated effective population size plotted against generations in the past, truncated at 10,000 generations.

Figure S2 Boxplot representing estimated effective population size plotted against generations in the past, truncated at 1,000 generations.

Table S1 Quality control criteria implemented on genotype data and the number of SNPs discarded at each step.

Table S2 Distance classes and bin ranges for the linkage disequilibrium summary.

Table S3 Chromosome-specific centimorgan to megabase (cM/Mb) conversion ratios.

Table S4 Description of generation binning process.

Appendix S1 Details of Meta-Analysis.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
RT-qPCR Comparison of mast cell populations in whole blood from healthy horses and those with laminitis

S. A. Brooks* and E. Bailey†
*Department of Animal Science, 129 Morrison Hall, Cornell University, Ithaca NY 14853, USA. †108 Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, KY 40546-0099, USA

Summary

Inflammatory damage to the digital laminae, a structure responsible for suspension of the distal skeleton within the hoof capsule, results in a painful and often life-threatening disease in horses called laminitis. There can be many diverse causes of laminitis; however, previous work in the horse has suggested that in each case, the inflammation and resulting tissue damage is consistent with the action of mediators released from mast cells (MC), as well as the downstream consequences of their activation. The recent development of molecular genetics tools to characterize cells based on their transcriptional activity makes a new approach for measuring MCs possible. Healthy thoroughbred horses from a variety of age groups were used to assess the amount of variation in \( \text{KIT} \) (encoding mast cell growth factor receptor) and \( \text{TPSB2} \) (encoding mast cell tryptase beta 2) gene expression present in the population and to establish “normal” values. Horses (\( n = 9 \)) with a wider range of body condition scores (3–8), because of a more lax management setting that could predispose them to laminitis, had significantly higher \( \text{KIT} \) expression in circulating peripheral blood cells than horses under individualized management conditions (\( n = 10 \)) that produced ideal body condition scores (4–6) (mean 2.573-fold, \( P < 0.0005 \)). Likewise, horses affected with acute laminitis (\( n = 11 \)) had elevated expression of \( \text{TPSB2} \) (2.760-fold, \( P = 0.0011 \)) relative to control horses (\( n = 15 \)). These data suggest that investigation of MC-related genes \( \text{KIT} \) and \( \text{TPSB2} \) may be effective to assay MC population and activity. More work is needed to refine the diagnostic criteria to better describe at what point MC activation occurs and illustrate the use of gene expression assays in clinical cases of laminitis. Additionally, MC activation is associated with inflammatory disease in several mammalian species and may prove a valuable therapeutic target in the horse.

Keywords horse, inflammation, laminitis, mast cell, RT-qPCR, whole blood biomarker.

Introduction

Laminitis is one of the most common causes of death in the horse. The precise aetiology of the disease is unknown, although it can occur following a wide variety of inflammatory stimuli. It is characterized by a breakdown of the laminae, a complex structure that suspends the distal phalanx within the hoof capsule. This inflammation destabilizes the foot, leading to rotation and/or displacement of the distal phalanx. Accompanied by bruising and oedema, it is extremely painful and debilitating. Vascular changes are associated with the disease, in the form of both increased regional blood flow and regional hypoperfusion, especially in the laminar capillary bed, leading to vascular damage and poor perfusion of the tissues (Moore et al. 2004). The diverse mechanisms for initiation of laminitis are varied and poorly understood. They can include obesity, equine metabolic syndrome, carbohydrate overload (over-consumption of grain or lush pasture), black walnut toxicity, endotoxemia, sepsis, mechanical injury and glucocorticoid administration, to name but a few (Hood 1999). These varied triggers are frequently grouped in to three broad categories, endocrinopathic, toxic and traumatic laminitis. Regardless of the triggering event, be it physical trauma, obesity/endocrine imbalance or bacterial/toxic assault, the end result is the same, lamellar breakdown.
Mast cells are sensitive to the types of biological situations responsible for the development of laminitis and play a central role in inciting damage caused by the types of triggers mentioned above. Mast cells (MCs) are a specialized subset of immune cells of bone marrow origin. Specifically, connective tissue mast cells act as ‘gatekeepers of the microvasculature’, are located in vascularized connective tissues and are often a first defence against attack. Mast cells respond to a variety of stimuli, just a few of which include endocrine imbalance, bacteria and mechanical injury (Gruber & Kaplan 2005). Once activated, they quickly degranulate, releasing powerful pre-formed mediators including histamine, heparin, superoxide, tryptase and chymase as well as other Matrix metalloproteinases (MMPs), platelet-activating factor, TNF-α and many inflammatory cytokines (Gruber & Kaplan 2005). The mediators produced by MCs lead to rapid inflammation in their resident tissues followed by the recruitment of leucocytes and additional MCs. Although equipped to protect the body from a variety of assaults, MC activation, particularly in a confined area, can cause more damage than good (Lazarus et al. 2000).

Because of their broad tissue distribution and migratory development, in vivo study of MCs is difficult in large mammals where invasive procedures must be limited. Identification of cells in tissue is traditionally carried out by morphology, histochemical staining and immunological staining. Identification of MCs by histology is often dependent on subjective quantification of heterochromatic staining of the granules. In humans, strategies have been developed to take advantage of the accessibility of whole blood and the transient presence of MC products in the circulation. These techniques measure the presence of mediators including tryptase and histamine by ELISA and have proven useful for conditions ranging from mastocytosis to anaphylaxis and heart attack. The recent completion of the equine genome sequence and the resulting novel molecular genetics tools specific to the horse are useful to characterize cells based on their transcriptional activity and thus make a new approach possible. Using the genome sequence, primers and probes can be quickly and easily generated for our target of interest. We have chosen RT-qPCR for mast cell-specific genes, including that encoding mast cell growth factor receptor (KIT, expressed exclusively by mast cells in the circulation, aka CD117) and TPSB2 (encoding mast cell tryptase beta 2, a mediator produced by maturing and activated MCs) (Liu et al. 2006). The KIT gene has been shown to be expressed specifically by circulating MC progenitors, the precursor cells for mature MCs that appear to be similar to basophils (Scherthanher et al. 2005). Measuring changes in KIT gene expression in blood will illustrate changes in the circulating population of MC precursors. Finally, tryptase (encoded by TPSB2) is a mast cell-specific product that is induced during maturation and activation (Liu et al. 2006). Presence of TPSB2 mRNA in the peripheral circulation will reflect the presence of maturing mast cell precursors as they migrate to their destination tissues (Welker et al. 2005). We therefore hypothesize that the detection of these genes by qRT-PCR could allow their use as biomarkers for determining the potential aetiologies and clinical severity of laminitis.

Materials and methods

Horses

Experiment 1 Thirty thoroughbred horses were selected from a single farm to establish reference values for relative gene expression and evaluate changes relative to age and body condition. Fifteen horses were between 6 months of age and 2 years. The remaining 15 individuals were 5 to 23 years. All horses were healthy and on a carefully managed diet designed to maintain optimal body condition. The mean body condition score (BCS) of this group was six (Henneke et al. 1983). As defined by Henneke and co-authors, body condition scores are based on assessment of adiposity at five anatomical regions that is combined in to a single score that ranges from one to ten. Scores between four and six are ideal, while scores above seven are associated with obesity, and scores below three are associated with poor condition. Samples from this group of horses were obtained over a 2-day period in late September.

Experiment 2 Two groups of thoroughbred mares (‘A’ and ‘B’) on two different farms with two different management situations were identified to compare the effect of environment on MC-related genes (Table S1-a). During these months, lush grazing is plentiful in well-managed pastures, thus exposing our study horses to one of the most common risk factors for laminitis. Mature, age-matched Group A mares were selected from among the horses used in Experiment 1. These ten mares resided primarily at pasture but were kept indoors overnight during inclement weather, including early spring. On farm A, a major focus of management was to maintain a BCS of five by restricting access to lush pasture and feeding supplemental grain when necessary. At the time of sampling, group A mares were free from lameness, had body condition scores ranging from five to six and were on average 15.3 years (range 10 to 23 years). At farm A, the incidence of laminitis was reported by management to be approximately 2%. Group B’ was comprised of nine mares residing at pasture, including during the lush spring and fall seasons associated with pasture laminitis. Supplemental hay was used to maintain body condition during winter months. Individuals were removed from pasture and stall-kept only for veterinary treatment (including for development of lameness). At the time of sampling, mares were free from lameness, had body condition scores ranging from three to eight and were on average 16.7 years (range 11 to 23 years). This broad range in body condition was likely due to the lack of
individualized nutritional management and free access to lush spring pasture. Incidence of laminitis at farm B was reported by management to be approximately 20%, providing evidence that the environment on this farm was more favourable to the development of laminitis than that on farm A. Samples from farm B were obtained over a 2-day period in late September.

**Experiment 3** Eleven additional thoroughbred mares diagnosed by the attending veterinarian with chronic or acute laminitis were obtained from farm B (Table S1-b). The majority of these mares developed laminitis as a result of obesity and/or a sudden change in pasture/diet. Disease severity among affected horses ranged from mild lameness treated with supportive shoes to Obel grade IV lameness and eventual euthanasia (individual scores provided in Table S1-b). These samples were collected in the summer months (from May to August). A set of 15 older, laminitis-free mares from Experiment 1 were used as controls, as they matched this group in breed as well as mean age.

**RNA isolation**

Blood was collected by jugular venipuncture into RNA-stabilizing PAXgene blood collection tubes. Total RNA was isolated and DNase-treated using the PAXgene Blood RNA kit (PreAnalytiX/Qiagen). RNA samples were quantified using a NanoDrop spectrophotometer (NanoDrop Technologies).

**RT-qPCR**

Reverse transcription was carried out using the TaqMan Reverse Transcription reagents and random hexamers [Applied Biosystems (ABI)] with incubation for 45 min at 45 °C in a thermocycler. Exon-spanning primers and MGB-binding probes were designed using the Primer Express software program (ABI) for the following genes: β-actin (ACTB, accession #AF035774), mast cell growth factor receptor (KIT, #AY874543) and mast cell tryptase beta 2 (TPSB2, aka equine MTC-1, accession #AJ515902) (Table 1). Exon boundaries were determined by BLAT search (UCSC genome browser, genome.ucsc.edu) and alignment of the mRNA sequence against the equine genome (EqCabv2.0). All primers produced a single band of the correct size when products were visualized by electrophoresis on an agarose gel. Although there are several known actin pseudogenes, DNase treatment eliminated detectable transcripts in non-RT controls. The primers and probes were used in a reaction including the equivalent of 40 ng cDNA from the RT step and TaqMan Gene Expression Master Mix (ABI) as per the manufacturers’ protocol. Thermocycling and detection were performed on the 7500 Fast Real-Time PCR System using the standard parameters (60°C annealing, ABI). Appropriate negative and non-RT controls were included for each run.

**Data analysis**

Assay efficiency was calculated for each sample using the LinRegPCR software (Ramakers et al. 2003). Individual wells were excluded and replaced if their calculated efficiency (E) exceeded two standard deviations from the mean or if the $R^2$ for the linear fit estimating efficiency for that well was below 0.99. Relative target gene expression was adjusted for mean reaction efficiency and calculated relative to a reference gene using the following equation: relative expression $(RE) = E_{(Reference \text{ gene})}^{C_t \ (Reference)} / E_{(Target \text{ gene})}^{C_t \ (Target)}$ (Schefe, 2006). There was no significant difference in raw ACTB $C_t$ among groups, indicating that multiple references genes were not necessary. Fold-change was calculated as a ratio of relative gene expression in one group versus the other group in the same experiment.

Graphs, descriptive statistics and tests for significance were calculated using the JMP 7.0 software package (SAS Institute Inc., 2007). Coefficient of fit was calculated for the data from experiment 1, including linear and polynomial fits. A Shapiro–Wilk test for goodness-of-fit to the normal distribution revealed that the data from experiments 2 and 3 were not normally distributed. For these experiments, a Wilcoxon test was used to calculate P-values.

**Results**

**Experiment 1**

No trend in expression of either gene (KIT and TPSB2) was observed owing to age among the group I horses (best fit $R^2 = 0.050$ and 0.053 for a polynomial curve, Fig. 1a and b, respectively). Additionally, no effect was observed based on BCS category, although the limited amount of variation within this group is not a good measure of this type of

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Primers and probes used in this study.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>Forward Primer 5′–3′</td>
</tr>
<tr>
<td>ACTB</td>
<td>CGA GGC CCC CTG AAC</td>
</tr>
<tr>
<td>KIT</td>
<td>GCG TCC TGC TTC TCC TGT T</td>
</tr>
<tr>
<td>TPSB2</td>
<td>CGG CGG CGG ACT GT</td>
</tr>
</tbody>
</table>

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 16–22
association. Mean values were established for the group as a whole of 2.303–3 for KIT/ACTB and 3.70e-6 for TPSB2/ACTB.

Experiment 2

Expression of KIT, a cell surface marker for MCs, was increased 2.573-fold (P = 0.0005) in those horses from farm ‘B’, the facility with the higher rate of laminitis, relative to farm ‘A’ (KIT/ACTB, Fig. 1a). An increase of similar magnitude was also observed in TPSB2/ACTB expression between farm ‘B’ and ‘A’ (2.633 fold, P = 0.0071, Fig. 1b). There was no significant difference in age between groups A and B, (mean 15.3 and 16.7 years, P = 0.76). Although the means were not significantly different for BCS (5.4 and 5.56, P = 0.49), the range of BCS values was slightly different between groups (P = 0.004 by a two-sided F-test) (Table S1). However, the sample size was insufficient to assess any direct correlation between this broad range in BCS and gene expression.

Experiment 3

KIT/ACTB expression did not change significantly between groups with laminitis and control horses (Fig. 3a). However, a mean 2.760-fold increase (P = 0.0011) in TPSB2/ACTB was found in the group with laminitis relative to the group with no previous diagnosis of the disease (Fig. 3b). The range of expression within the laminitic group was quite large (0.8442–12.27-fold increase over the control group) in contrast to the more narrow range among control horses. No direct association with Obel score or total days lame and TPSB2 expression was found, although this may in part be attributable to the small sample size (Table S1-b).

Discussion

The KIT and TPSB2 genes are expressed almost exclusively in whole blood by mast cells and their precursors. Consequently, these data demonstrate that mast cell gene expression can be measured in whole blood despite the transient nature of their presence in circulation. Although this could also be accomplished using flow cytometry and antibodies for these mast cell products, antibody development can be a lengthy process. Additionally, blood samples collected in PAXgene tubes for RNA analysis are far more stable and easy to handle than fresh blood for flow cytometry. Side-by-side comparisons of traditional flow cytometry and RT-qPCR using the same targets for the quantification of human peripheral blood cell populations have shown that RT-qPCR performs well at this task and is exceptionally sensitive (Pennington et al. 2001).

Age did not have an effect on gene expression for TPSB2 or KIT (Fig. 1). However, statistically significant differences were observed for management (Fig. 2) and the presence of lameness (Fig. 3). Horses in a more relaxed management setting, with unrestricted access to spring pasture, had 2.573-fold higher expression (P = 0.0005) of KIT. This difference in KIT/ACTB expression should, as supported by previous comparisons of flow cytometry and gene expression methods of quantifying cellular populations (Pennington et al. 2001), reflect an increase in the proportion of MC precursors in circulation. This increase could contribute to a predisposition to inflammatory conditions. A proportionate increase in TPSB2/ACTB expression is likely due to this change in the population of TPSB2-carrying cells rather than an increase in the transcription of TPSB2 per cell. This agrees with our hypothesis, given that both groups were free from disease or lameness and therefore MC activation, at the time of sampling. Further work is needed to determine specifically which conditions contribute to these changes in circulating, KIT-expressing cells and what their implications are for disease. As certain breeds are commonly believed to be more predisposed to laminitis, future work should also include the use of these assays in both predisposed and resistant breeds.

In contrast, changes in KIT/ACTB expression between the laminitic and healthy groups were not noted. Theoretically, once MC activation has occurred, all suitably mature MCs have been recruited to their target tissues, leaving behind a

![Figure 1](https://example.com/figure1.png)
baseline level of MC precursors. This may explain why a difference was not detected following an inflammatory event. However, significant changes (2.760-fold on average, \( P = 0.0011 \)) were noted in TPSB2/ACTB expression in the laminitic group. This likely reflects ongoing activation and maturation of MCs (Welker et al. 2005). The broad range of up-regulation observed in the laminitic group correlates well with the broad diagnostic criteria used in this study. This group contained everything from reasonably sound, Obel grade 0–1, chronically affected individuals to extremely severe acute cases. No relationship could be found between TPSB2 expression and the severity or duration of lameness or acute vs. chronic status. Thus, additional work on larger numbers of horses, including well-documented chronic, acute and developmental cases, with more detailed clinical histories is needed to investigate gene expression changes relative to more detailed diagnoses and a variety of trigger factors. Although not feasible for this study, it would be advantageous in future work if examination and diagnosis could be performed by a single clinician. However, objective quantification of the damage within the hoof remains a difficult task and is not always well correlated with lameness.

Previously, plasma histamine levels in chronically laminitic horses were reported as nearly three times that of unaffected horses (Rautschka et al. 1991). More recently, chronically laminitic horses were shown to have significantly higher responses to intradermal allergen tests compared to unaffected horses (Wagner et al. 2003). MCs are vital to the immediate response to allergens, and their degranulation and release of histamine is responsible for wheal formation (Gruber & Kaplan 2005). Thus, the exaggerated response to allergens may reflect an increase in the number or activity of MCs.

Current research in laminitis has provided additional evidence that MCs could be a pivotal part of the disease
process. Emigration of leucocytes is a significant early characteristic of the development of laminitis (Black et al. 2006; Hurley et al. 2006). MCs release several chemotactic factors that direct the infiltration of leucocytes into the site of inflammation (Gruber & Kaplan 2005). Notably, pharmacological stabilization of MCs reduces the expression of adhesion molecules on circulating leucocytes that is necessary for their movement through vessel walls (Zhao et al. 2005). Additionally, neutrophil infiltration into wounds is significantly reduced in mast cell-deficient mice, although wound healing was otherwise normal (Egozi et al. 2003).

Recent work in humans has provided new insight into the importance of the mast cell in vascular and inflammatory disease, particularly when secondary to obesity or metabolic syndrome (Chaldakov et al. 2001; Deliargyris et al. 2005; Kounis 2006). Although the vascular component of the pathological course of laminitis is well documented and preliminary work has suggested systemic mast cell involvement (Rautschka et al. 1991; Wagner et al. 2003), the specific role of the mast cell in laminitis has yet to be investigated. Our work examines the relationship between laminitis in the horse and circulating mast cells. If a method can be devised to reduce the number or reactivity of mast cells in tissues, then it may be possible to prevent laminitis in susceptible individuals and reduce the recurring cycle of tissue damage and inflammation, potentially allowing the effective pharmacological treatment of early and recurrent cases of the disease. Current treatments focus on reducing pain and damage and can do little to prevent reoccurrence. Although removing known trigger factors can help, there is a desperate need for a way to gain specific control of the inflammatory processes within the hoof. Several drugs are known to be effective in blocking mast cells in humans, and adapting these compounds for equine use may prove beneficial. However, we cannot devise an effective treatment protocol in the absence of a method to assess mast cell number and function. This work has shown that RT-qPCR with mast cell-specific genes could fill that role.

Acknowledgements

The authors thank the many horse owners and clinicians who supplied samples and clinical information for this study. S. Brooks was supported by the Paul Mellon Post-Doctoral Fellowship.

Conflicts of interest

The authors have declared no potential conflicts.

References


Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Relative expression data for individual horses.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Transcription of *LINE*-derived sequences in exercise-induced stress in horses

S. Capomaccio*, A. Verini-Supplizi*, G. Galla†, N. Vitulo‡, G. Barcaccia†, M. Felicetti*, M. Silvestrelli* and K. Cappelli*

*CSCS-DPDCV, University of Perugia, Via San Costanzo 4, 06126 Perugia, Italy. †DAAPV, University of Padua, Viale dell’Università 16, 35020 Legnaro, Italy. ‡CRIBI, University of Padua, Complesso Vallisneri, Viale G. Colombo, 3, Italy

Summary

A large proportion of mammalian genomes is represented by transposable elements (TE), most of them being long interspersed nuclear elements 1 (*LINE*-1 or *L1*). An increased expression of *LINE*-1 elements may play an important role in cellular stress–related conditions exerting drastic effects on the mammalian transcriptome. To understand the impact of TE on the known horse transcriptome, we masked the horse EST database, pointing out that the amount is consistent with other major vertebrates. A previously developed transcript-derived fragments (TDFs) dataset, deriving from exercise-stimulated horse peripheral blood mononuclear cells (PBMCs), was found to be enriched with *L1* (26.8% in terms of bp). We investigated the involvement of TDFs in exercise-induced stress through bioinformatics and gene expression analysis. Results indicate that *LINE*-derived sequences are not only highly but also differentially expressed during physical effort, hinting at interesting scenarios in the regulation of gene expression in relation to exercise.

Keywords exercise, exonization, horse, stress, transposable elements.

Complex genomes have a great advantage in transposable element (TE) insertion when new genes and splicing variants are required to cope with adaptation: exonization, together with retrogene formation, is one of the major sources of this variability (Lin et al. 2008). TE insertions may also alter the expression of nearby genes with new splicing sites, adenylation signals, promoters and transcription factor binding sites building new transcription modules that can generate diversity among transcriptomes (Goodier & Kazazian 2008).

The most widespread family of autonomous TEs is *long interspersed nuclear element L1* (*LINE*-1 or *L1*), an LTR-less (long terminal repeats) retrotransposon, which constitutes about half of the total number of TEs. Approximately 75% of mammalian genes contain at least one *LINE*-1 segment in their transcription unit, mainly within intronic regions and in poorly expressed genes (Han et al. 2004). Compelling evidence has implicated *LINE*-1 elements in the regulation of genome-wide gene expression by acting as a molecular fine-tuner of the RNA and microRNA biology (Ramos 2009).

In physiological conditions, *L1* elements are predominantly found in germ cells but they also seem to be abundantly transcribed in differentiated cells exposed to cellular stress (Li & Schmid 2001), suggesting their functional role as an integral component of global genomic response to environmental stress and in cellular stress–related disease aetiology, exerting drastic effects on the mammalian transcriptome (Schulz 2006). In addition, it is known that physical stress promptly activates neuroendocrine and immune responses with the production of pro-inflammatory mediators that alter gene expression, especially in peripheral blood cells (Kawai et al. 2007).

In this work, we investigated the involvement of several *LINE*-derived sequences in exercise-induced stress through bioinformatics and gene expression analysis, starting from an already available EST dataset derived from peripheral...
blood mononuclear cells (PBMCs) of endurance horses (Genbank accessions from CO508721 to CO598770, published in Cappelli et al. 2007; Genbank accessions from FG341833 to FG341843 and from GH986483 to GH986492, unpublished) containing 74 transcript-derived fragments (TDFs) with a different expression patterns at three time points defined as: before the race, at the end of the race, and 24 h after the race.

As the horse genome retroelement content is consistent with other mammals (Wade et al., 2009), we assessed the mean level of TE expression in Equus caballus. Specific analysis with RepeatMasker was conducted on the horse ESTs database, showing that repetitive elements are as low as 4.76% of the sequences, with LINEs equal to 2.17%, as expected. Conversely, RepeatMasker analysis of our EST dataset revealed that 23 of 74 sequences (equal to 31.0% or 26.8% of the total length of sequences in bp) belong to repetitive elements, the category of LINE being the one most represented, with 82.5% of the total. In particular, L1 plays the major role, with about 95% of the total (Table 1). Moreover, intersecting L1 repeat genomic coordinates with the EquCab2.0, we found that 72% of these elements fall in intergenic regions, while 28% are included in at least one gene. We also calculated the coverage of this repeat within the coding sequence of L1-containing gene: more than 82% of these sequences were inside introns (Table S1).

A similar behaviour is apparent for our TE-derived TDFs mapped in the horse genome using BLAT (Kent 2002): 17 of 23 sequences were comprised in predicted or known genes (Table 2).

An automated gene ontology (GO) analysis of our EST dataset was performed with BLAST2GO (Conesa et al. 2005) to annotate all sequences. Several ontological terms related to signal transduction and stress response were identified for the TE-derived TDFs (Tables S2–S4). Statistical analysis of GO data distribution using the Fisher’s exact test supported the enrichment of terms associated with TE-derived TDFs, such as response to external stimulus (P = 0.012) and GTPase activator activity (P = 0.038).

To verify the TE-derived TDFs modulation of expression during exercise, qRT-PCR was carried out as described elsewhere (Cappelli et al. 2008) on PBMCs of six horses chosen from the high-level participants in national and international endurance races (90–160 km). Samples were collected at the three different time points described previously. Primers were designed on TE-derived TDFs using Primer3 software (Table 2). Ten of the 23 tested sequences significantly increased their expression in horses at the end of the race. The fold-change, calculated as the ratio of race/basal values and expressed as log2 of the real value, ranged from 1.605X of CO508722 to 3.523X of CO508756 (Table 2).

RACE experiments were successfully conducted for most of the TDFs showing an increased expression. BLAT mapping of the extended cDNA sequences (GenBank accessions GU797237, GU797238 and GU797239) confirmed the genomic position of original TDFs and revealed an exercise-induced transcription of large segments of the LINEs contained in known or predicted gene introns.

Our results are consistent with the accepted concept that stressful environments increase TE retrotransposition as well as alternative splicing (Teneng et al. 2007), hence, indicating that physical stress may play a role in modulating the activity of LINE-derived and/or LINE-linked sequences in strenuous exercise in horses. For these sequences, it is possible to hypothesize that they represent the relic of LINEs embedded in modern genes modulated by exercise and stimulated to exonize. The cellular context appears to be a crucial factor in TE exonization with identified tissues and disease-specific transposon-derived cDNA sequences (Mersch et al. 2007).

These integrated molecular and bioinformatics data reveal new insights and shed some light on the genomic factors and regulatory mechanisms involved in strenuous exercise in horses. They also offer intriguing scenarios that

Table 1 RepeatMasker output on different groups of sequences.

<table>
<thead>
<tr>
<th>Endurance ESTs (% on 12589 bp)</th>
<th>Equus caballus ESTs (% on 1976073 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINEs</td>
<td>3.83</td>
</tr>
<tr>
<td>Alu/B1</td>
<td>0</td>
</tr>
<tr>
<td>MiRs</td>
<td>1.61</td>
</tr>
<tr>
<td>Line</td>
<td>22.15</td>
</tr>
<tr>
<td>Line1</td>
<td>20.94</td>
</tr>
<tr>
<td>Line2</td>
<td>1.21</td>
</tr>
<tr>
<td>L3/CR1</td>
<td>0</td>
</tr>
<tr>
<td>RTE</td>
<td>0</td>
</tr>
<tr>
<td>LTR elements</td>
<td>0</td>
</tr>
<tr>
<td>ERVL</td>
<td>0</td>
</tr>
<tr>
<td>ERVL-MaLRs</td>
<td>0</td>
</tr>
<tr>
<td>ERV_classl</td>
<td>0</td>
</tr>
<tr>
<td>ERV_classll</td>
<td>0</td>
</tr>
<tr>
<td>DNA elements</td>
<td>0.87</td>
</tr>
<tr>
<td>hAT-Charlie</td>
<td>0</td>
</tr>
<tr>
<td>TcMar-Tigger</td>
<td>0</td>
</tr>
<tr>
<td>Unclassified</td>
<td>0</td>
</tr>
<tr>
<td>Total interspersed repeats</td>
<td>26.84</td>
</tr>
<tr>
<td>Small RNA</td>
<td>2.22</td>
</tr>
<tr>
<td>Satellites</td>
<td>0</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>0</td>
</tr>
<tr>
<td>Low complexity</td>
<td>0.82</td>
</tr>
</tbody>
</table>

1 Downloaded from GenBank, December 2009
The query species was Equus caballus, RepeatMasker version open-3.2.7, sensitive mode. Run with blastp version 2.0MP-WashU RepBase Update 20090120, RM database version 20090120.
<table>
<thead>
<tr>
<th>GenBank Accession</th>
<th>Primers (Forward and Reverse)</th>
<th>Putative containing gene (ENSEMBL or N-SCAN naming)</th>
<th>Repeat Type (class/family)</th>
<th>Real Time (basal vs race)</th>
<th>Exp. status</th>
<th>FC</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO508721</td>
<td>GGCAATCTCTAATCTTGTGGTACTG</td>
<td>chr5.772.1 Dihydropyrimidine dehydrogenase</td>
<td>LINE/L1</td>
<td>Up-regulated</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508722</td>
<td>CTTGAGTGTGGCTATACTCTGGGC</td>
<td>ENSECAG00000016412 Brefeldin A-inhibited guanine nucleotide-exchange protein</td>
<td>LINE/L1</td>
<td>Up-regulated</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508730</td>
<td>TTAGATGCTCGGGGCAATGGGC</td>
<td>ENSECAG00000015342 Interleukin-8 Precursor (IL-8)</td>
<td>DNA/hAT</td>
<td>Up-regulated</td>
<td>0.0009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508733</td>
<td>TTCGGAAAGGGGAGTG</td>
<td>–</td>
<td>LINE/L1</td>
<td>Up-regulated</td>
<td>0.0213</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508740</td>
<td>ATATAATGCCCCACGCCATG</td>
<td>chrUn.282.1 Olfactory receptor 5an1 (olfactory receptor or11-244)</td>
<td>LINE/L1</td>
<td>– no amplification –</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508741</td>
<td>AGAGCAGTGAGGAGAAGATG</td>
<td>ENSECAG0000000381 DmX-like protein 2 (Rabconnectin-3)</td>
<td>LINE/L1</td>
<td>Up-regulated</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508743</td>
<td>AAGAAACAGCCCTAAGATG</td>
<td>chrUn.021.1 Trigger transposable element derived 1</td>
<td>LINE/L1</td>
<td>No change</td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508751</td>
<td>CCAATGCATGGGTGGAATGTC</td>
<td>–</td>
<td>LINE/L1</td>
<td>No change</td>
<td>0.756</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508752</td>
<td>AGAATATCCTACACACCCAT TAG</td>
<td>–</td>
<td>LINE/L1</td>
<td>No change</td>
<td>0.1158</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508753</td>
<td>ATCCAGCCCTCTGCAGATCTC</td>
<td>–</td>
<td>SINE/MIR</td>
<td>No change</td>
<td>0.3705</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508755</td>
<td>CCTTTATGGTTTTTGCTCCAG</td>
<td>ENSECAG00000005540 Putative RNA-binding protein 16</td>
<td>LINE/Cr1</td>
<td>Up-regulated</td>
<td>0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508756</td>
<td>AATGGGCAATCAGAGGTG</td>
<td>ENSECAT00000011120 Syntaxin binding protein 5</td>
<td>LINE/Cr1</td>
<td>Up-regulated</td>
<td>0.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508760</td>
<td>TAAAGGCACTGTGGTTATGCAAGAC</td>
<td>ENSECAG00000020278 Rap1 GTPase-activating protein 2 (Rap1GAP2)</td>
<td>LINE/L1</td>
<td>No change</td>
<td>0.5731</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508763</td>
<td>CCCTTACAAGCTGTAGCTATCC</td>
<td>chr11.298.1 Connexin 45</td>
<td>LINE/L1</td>
<td>No change</td>
<td>0.091</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508765</td>
<td>CTTTACTTCTTGGGCATTCCTTCC</td>
<td>–</td>
<td>LINE/L1</td>
<td>No change</td>
<td>0.955</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO508768</td>
<td>AGGCAATGAGGAGGAAAGTGG</td>
<td>chr11.974.1 Riken cdna a530088h08 gene</td>
<td>LINE/L1</td>
<td>– no amplification –</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FG341837</td>
<td>AAGGACATATAACCCCAAGGAC</td>
<td>ENSECAG0000001583 Aminopeptidase O (AP-O)</td>
<td>LINE/L1</td>
<td>– no amplification –</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FG341839</td>
<td>TGAGGTGTGATTCGTTCTCTCTTG</td>
<td>chr10.271.1 Zinc finger protein 420</td>
<td>LINE/L1</td>
<td>– no amplification –</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
might explain the transcriptional processes of L1-type retrotransposons in exercise-induced stress.

Acknowledgements

The authors thank Prof David Adelson for providing an improved version of LINE-1 annotation and Mr Gianluca Alunni for his valuable technical support. MIPAF SelMol supported this work.

Conflicts of interest

The authors have declared no conflicts of interest.

References


Supporting information
Additional supporting information may be found in the online version of this article.
Table S1 L1 coverage distribution within Horse genes.
Table S2 Annotation according to the cellular component.
Table S3 Annotation according to the biological process.
Table S4 Annotation according to the molecular function.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
A genome-wide association study for racing performances in Thoroughbreds clarifies a candidate region near the MSTN gene

T. Tozaki*, T. Miyake†, H. Kakoi‡, H. Gawahara‡, S. Sugita§, T. Hasegawa§, N. Ishida§, K. Hirota* and Y. Nakano¶

*Department of Molecular Genetics, Laboratory of Racing Chemistry, Utsunomiya, Tochigi. †Comparative Agricultural Sciences, Graduate school of Agriculture, Kyoto University, Sakyo, Kyoto. ‡Department of Genetic Analysis, Laboratory of Racing Chemistry, Utsunomiya, Tochigi. §Equine Research Institute, Japan Racing Association, Utsunomiya, Tochigi. ¶Department of Pharmacogenomics, School of Pharmaceutical Sciences, Showa University, Shinagawa, Tokyo, Japan

Summary

Using 1400 microsatellites, a genome-wide association study (GWAS) was performed to identify genomic regions associated with lifetime earnings and performance ranks, as determined by the Japan Racing Association (JRA). The minimum heritability ($h^2$) was estimated at 7–8% based on the quantitative trait model, suggesting that the racing performance is heritable. Following GWAS with microsatellites, fine mapping led to identification of three SNPs on ECA18, namely, g.65809482T>C ($P = 1.05E-18$), g.65868604G>T ($P = 6.47E-17$), and g.66539967A>G ($P = 3.35E-14$) associated with these performance measures. The haplotype of these SNPs, together with a recently published nearby SNP, g.66493737C>T ($P = 9.06E-16$) in strong linkage disequilibrium, also showed a very clear association with the performance ($P < 1E-05$). The candidate genomic region contained eight genes annotated by ENSEMBL, including the myostatin gene ($MSTN$). These findings suggest the presence of a gene affecting the racing performance in Thoroughbred racehorses in this region on ECA18.

Keywords GWAS, heritability, horse, racing performance, Thoroughbred.

Introduction

Thoroughbreds are a relatively recent horse breed derived from a small number of Arabian stallions and native British mares approximately 300 years ago (Cunningham et al. 2001; Hill et al. 2002). Various measures for the evaluation of racing performances in Thoroughbreds, such as earnings, racing times, and handicap ratings were used to estimate heritability (Tolley et al. 1985; Gaffney & Cunningham 1988; Oki et al. 1995; Williamson & Bellharz 1998; Mota et al. 2005; Sobczynska 2006). The racing time showed moderate heritability in the range of 0.1–0.2, with higher values for shorter distances. The reported heritabilities for the handicap and earning measures were generally higher than those for the racing time, in the range of 0.3–0.4.

Thoroughbreds have a very high muscle mass ratio to body weight (55%) (Gunn 1987). Recent studies proposed several genes related to muscle strength, which might be selected in horse breeding or evolutionary processes of Thoroughbreds (Gu et al. 2009; Hill et al. 2010). A candidate gene approach by Hill et al. (2010) identified a genetic variant associated with the horse myostatin ($MSTN$) gene, which was also associated with the best race distance (BRD) in Thoroughbreds from Ireland and New Zealand. The variant was identified as a SNP, g.66493737C>T, in the first intron of the $MSTN$ gene. Based on this SNP, Thoroughbreds with the C/C genotype were suitable for fast speeds at short-distance races; C/T genotype horses competed favourably in middle-distance races; and T/T genotype horses had greater stamina and represented most of the winning horses at the longer distance races.

Many significant advances have recently been achieved in horse genome mapping. Half- and full-sibling linkage maps of horses have been created (Penedo et al. 2005;
Swinburne et al. 2006). Horse–human comparative maps were published in 2007 and 2008 (Tozaki et al. 2007a; Raudsepp et al. 2008). With the complete horse genome sequence available to the equine genome community, we anticipate acceleration in discoveries of genetic traits of interest to horse breeders (Wade et al. 2009).

Here we describe the investigation of racing performances in Thoroughbred horses using an approach that we have described in earlier reports. In our previous studies, a genome-wide association study (GWAS) for complex traits in Thoroughbreds was proposed by Tozaki et al. (2005, 2007b), where the GWAS can be systematically performed in a two-step process. (i) initial screening with pooled DNA samples for genome-wide markers and (ii) screening for candidate markers implicated from the genome-wide scan.

In this study, we aimed to estimate heritability and to identify susceptible loci for racing performances in Japanese Thoroughbred populations by performing genetic and genomic analyses including GWAS for lifetime earnings (LE) on Japan Racing Association (JRA). The results of this study corroborate those reported by Hill et al. (2010).

Materials and methods

Estimation of heritability

To investigate the heritability and genetics on racing performances of the Japanese Thoroughbred racehorses registered by the JRA, we used their earnings on the racetrack. The JRA tabulates and reports earnings using two methods. First, total prize money won during a racing career is recorded. The top five racehorses in each race are usually awarded the prize money. This is designated as LE. Second, a record is kept of money won only for first place finishes. This metric basically counts money for the winning horse in each race. There are five ranks in the total amount. Rank 1 (0 million JPY class) is a rank for zero-winning horses. Then, the rank 2 (under 5 million JPY class), rank 3 (under 10 million JPY class), rank 4 (under 16 million JPY class), and rank 5 (over 16 million JPY class) are determined. Horses in the rank 1, 2, 3, and 4 can acquire 4, 5, 6, and 9 million JPY if they won their ranking races, and then they will be upgraded to the next rank. Horses in the rank 5 can get credit of the half of earning money if they won the rank 5 races. This second metric will be referred to as lifetime rank (LR). The distributions of LE and LR are shown in Fig. 1.

For estimating the heritability of LE and LR, 3927 Thoroughbreds were used, which were born between 1993 and 2000 and registered as JRA racehorses in October 2002. All the animals had retired by December 2006. For scoring LE and LR, all lifetime performance records were collected from the computerized database for JRA’s racehorses (JARIS: Japan Racing Information System) in January 2008. To estimate heritability, the pedigrees of the 3927 Thoroughbreds were traced back to ancestors for three generations; the total number of pedigree animals was 14 706.

The traits (LE and LR) were analysed using the quantitative trait model with the MTDFREML program (Boldman et al. 1995) and the GSTM program (Miyake et al. 1999), where the latter was mainly used for the computation of the standard errors and 95% confidence intervals (CI) of the heritability estimates. The SAS GLM procedure (SAS Institute Inc. 2000) revealed that the sex, retired year, and linear regression of the periods (years) of athlete life were significant. Then, the model was:

\[ Y_{ijk} = \text{SEX}_i + \text{YR}_j + \alpha (P_{ijk} - P^*) + u_k + e_{ijk}, \]

where \( Y_{ijk} \) is the value of LE or LR or their log 10 transformations, \( \text{SEX}_i \) is the effect of the \( i \)th sex, \( \text{YR}_j \) is the effect of the \( j \)th retired year, \( P_{ijk} \) is the period (years) of athletic life \( (P^* \) is the average of \( P_{ijk} \)), \( u_k \) is the breeding value of the \( k \)th individual \( (\text{i.e., it follows normal distribution } [N(0, A\sigma_u^2)]) \), \( e_{ijk} \) is the residual effect \( [N(0, \sigma_e^2)] \). \( A \) is numerator relationship matrix (Henderson 1973), \( \sigma_u^2 \) is additive genetic variance, and \( \sigma_e^2 \) is residual variance. Then, heritability \( (h^2) \) is defined as \( \sigma_u^2 / (\sigma_u^2 + \sigma_e^2) \). Based on the Box-Cox regression (lambda was 0.05710 and \(-0.09345 \) for LE and LR, respectively), the log 10 transformations were adapted. The total number of Gibbs samplings with GSTM program was 1 050 000, with a burn-in of 50 000.

Classification of Thoroughbred subpopulations for association study

For a GWAS, eight subpopulations were generated from the 3927 Thoroughbreds based on LE, gender, and pedigree. LE was used as an indicator of performance value, as described

---

**Figure 1** Frequency distribution for lifetime earnings and lifetime rank.
earlier. Gender was considered because distributions of earnings in males and females are different. Pedigree was used to avoid selecting offspring of the same sires and increase representation of different genetic backgrounds. Two groups of male horses with high LE were selected as described below and designated MG01 and MG02. Likewise, two groups of male horses with low LE were selected and designated MP01 and MP02. The same process was used to designate comparable groups for female horses.

The groups were designed as follows (see Fig. S1): First, male horses with the highest LE were chosen from each paternal half-sibling family (approximately 400 Thoroughbreds, because the number of sires in Japan is 400–450). Second, from the approximately 400 horses, 96 males were chosen by selecting those with the highest values of LE. This subpopulation of the 96 individuals was denoted as the Male-Good performance population 01 (MG01). For the purpose of replication, the second highest subpopulation was generated and designated MG02. For the male horses with lower LE (MP01 and MP02), a similar procedure was used, in which those with the lowest and second lowest LE from each family were chosen. If the lowest zero ties appeared in the same family, the horse with the most racing experience was chosen. In the same way, the subpopulations of female horses were chosen for the study and designated FG01, FG02, FP01, and FP02. These subpopulations were used for case–control association studies.

**Extraction of DNA**

The blood samples for DNA isolation were stored at −40 °C. The genomic DNAs were extracted from the stored blood by the MagExtractor System MX-2000 (Toyobo). All the DNA samples were measured twice (DU®7500 Spectrophotometer; Beckman Coulter) and diluted to a concentration of 20 ng/µl. These samples were measured again before they were combined under the case–control association studies. Finally, the DNA samples were combined into their respective pools. All the pooled DNA samples were adjusted to a concentration of 20 ng/µl.

**GWAS using microsatellites: first and second genomic screening**

For the first genomic screening, 1400 microsatellites, which cover 31 autosomal and X chromosomes, were used for GWAS. Following the initial tests, additional microsatellites were selected near those microsatellites showing statistical significance, and these were used in the second genomic screening. All the microsatellites used in this study were selected from the Animal Health Trust (AHT) linkage map (Swinburne et al. 2006), the workshop linkage map (Penedo et al. 2005), the radiation hybrid map (Raudsepp et al. 2008), and the human–horse comparative map (Tozaki et al. 2007a).

**Microsatellite genotyping of pooled DNAs**

For genotyping microsatellites, we prepared the following three primers: a sequence-specific forward primer conjugated with a 5′-TGA CCG GCA GCA AAA AAA TTG-3′ tail at its 5′ end, a sequence-specific reverse primer, and a FAM-labelled 5′-TGA CCG GCA GCA AAA ATTG-3′ primer (Applied Biosystems). The underlined sequences were used for fluorescence detection. PCR was performed as described in Tozaki et al. (2001). The reaction products were analysed using an ABI 3130 genetic analyser (Applied Biosystems), and genotyping data were applied to the Genotyper software (Applied Biosystems) to measure the peak height of alleles for each microsatellite.

**Fine mapping using SNPs**

Seventeen and 30 single nucleotide polymorphisms (SNPs) in ECA1 and ECA18, respectively, were used for fine mapping. They were genotyped by multiplex assays with the iPLEX using the MassARRAY® Assay Design software (Sequenom Inc.).

Four SNPs on ECA18, g.65809482T>C, g.65868604G>T, g.66493737C>T, and g.66539967A>G, were identified as SNPs associated with racing performance in both this study and that of Hill et al. (2010). The four SNPs were genotyped by high-resolution melting and unlabelled probe methods using LightScanner (Idaho Technology, Inc.). The primers and probes for the SNP genotyping are listed in Table S1.

**Statistical analysis for association study**

Differences in allele frequencies between subpopulations were analysed by chi-squared tests as P-values and odds ratios with Hardy–Weinberg Equilibrium (HWE). The calculations were carried out using the computer program SNPAlalyze ver. 7.0 Standard (Dynacom). The haplotype frequencies of three marker pairs (g.65809482T>C, g.65868604G>T, g.66493737C>T having strong linkage disequilibrium [LD] \( r^2 > 0.8 \) in this study) were compared between the case–control subpopulations. The significance of association between the haplotypes and the traits was tested using a permutation test with 10 000 permutations, performed by SNPAlalyze. The LD scores of the two marker pairs were evaluated with \( r^2 \), which has a range from 0 (linkage equilibrium) to 1 (complete LD). The LD scores were also calculated by SNPAlalyze.

**Results and discussion**

**Estimation of heritability**

The records of 3927 JRA horses were considered as a representative sample of the racehorse population in JRA at
that era. The heritability estimates were 0.12 and 0.11 for LE and LR, respectively (Table 1). The point estimates of heritability by the MTDFREML program were identical with those by the GSTM. With the log transformation, rather higher estimates were obtained (Table 1). The lower 95% confidence intervals (CI) were 0.08 for LE and 0.07 for LR, indicating that the racing performance in terms of LE is heritable.

**Genome-wide association study: first and second genomic screening**

In the first screening, pooled case–control populations (Table 2) were used for a GWAS using 1400 microsatellites: case (pooled MG01, MG02, FG01, and FG02) and control (pooled MP01, MP02, FP01, and FP02). Candidate microsatellites showing significant association with LE were then selected based on the characteristic trend of the allele frequency distributions in case–control subpopulations, as below. The selection criteria were as follows: (i) within the case groups and within the control groups, comparable genetic marker frequencies were observed and (ii) the distribution of genetic markers was different when comparing case and control subpopulations, as shown in Fig. S2. In the first and second screenings, two regions, one on ECA1 (185.10–185.35 Mb) and one on ECA18 (62.61–66.84 Mb), were selected. In particular, three microsatellites, TKY101, TKY016, and TK3080, spanning 65.55–65.80 Mb on ECA18 showed clearly different distributions between the case and control populations. No other markers achieved significance when applying these criteria.

**Fine mapping on ECA1 and ECA18**

Seventeen and 30 SNPs polymorphic (MAF ≥ 0.1) in Thoroughbreds, spanning 184.02–185.76 Mb on ECA1 and 61.64–67.55 Mb on ECA18, were used for the fine mapping using individuals in MG01 (96 horses) and MP01 (96 horses). No SNPs on ECA1 showed statistical significance in this assay. However, three SNPs on ECA18 showed strong positive signals for a candidate region near the *myostatin* gene (*MSTN*), supporting a recently published study (Hill et al. 2010). Hill et al. (2010) reported that one SNP, namely, g.66493737C>T, in the first intron of *MSTN*, was strongly associated with racing performance traits based on the BRD, such as sprinting ability and racing stamina. They suggested that the *MSTN* gene must be an essential genetic factor for racing performance in Thoroughbreds, although their study did not show statistical significance (*P* = 0.764) between the elite (TBE) and other (TBO) Thoroughbreds. Our candidate region spanning the three SNPs (g.65809482, g.65868604, and g.66539967), 65.80–66.54 Mb on ECA18, included the genomic region coding for *MSTN*. Therefore, with the addition of Hill’s information, four SNPs, namely g.65809482>T, g.65868604>C, g.66493737>C, and g.66539967>A, were used to evaluate the detailed association between these SNPs and racing performance. This was in terms of LE by the following various case–control populations; MG01 vs. MP01, MG02 vs. MP02, FG01 vs. FP01, FG02 vs. FP02, Male Good (MG01 + MG02) vs. Male Poor (MP01 + MP02), Female Good (FG01 + FG02) vs. Female Poor (FP01 + FP02) and Both Good (MG01 + MG02 + FG01 + FG02)

### Table 1

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Lifetime Earning (LE)</th>
<th>Lifetime Rank (LR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>h</em>^2^</td>
<td>95% CI</td>
</tr>
<tr>
<td>Original</td>
<td>0.12 ± 0.03</td>
<td>0.08–0.19</td>
</tr>
<tr>
<td>Log-transformed</td>
<td>0.21 ± 0.05</td>
<td>0.14–0.31</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Population</th>
<th>Average</th>
<th>Median</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Average</th>
<th>Median</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG01</td>
<td>230</td>
<td>170</td>
<td>1069</td>
<td>113</td>
<td>79</td>
<td>54</td>
<td>464</td>
<td>18</td>
</tr>
<tr>
<td>MG02</td>
<td>129</td>
<td>97</td>
<td>437</td>
<td>69</td>
<td>41</td>
<td>29</td>
<td>183</td>
<td>10</td>
</tr>
<tr>
<td>MP01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MP02</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FG01</td>
<td>97</td>
<td>79</td>
<td>483</td>
<td>42</td>
<td>30</td>
<td>23</td>
<td>233</td>
<td>4</td>
</tr>
<tr>
<td>FG02</td>
<td>51</td>
<td>38</td>
<td>163</td>
<td>18</td>
<td>14</td>
<td>12</td>
<td>59</td>
<td>4</td>
</tr>
<tr>
<td>FP01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FP02</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>All</td>
<td>38</td>
<td>15</td>
<td>1069</td>
<td>0</td>
<td>12</td>
<td>4</td>
<td>464</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>45</td>
<td>20</td>
<td>1069</td>
<td>0</td>
<td>14</td>
<td>4</td>
<td>464</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>25</td>
<td>9</td>
<td>535</td>
<td>0</td>
<td>8</td>
<td>4</td>
<td>233</td>
<td>0</td>
</tr>
</tbody>
</table>

Unit = 1 million JPY.

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 28–35
Table 3 SNPs tested for racing performance susceptibility.

<table>
<thead>
<tr>
<th>Markers</th>
<th>MG01 vs. MP01</th>
<th>MG02 vs. MP02</th>
<th>*Male Good vs. Male Poor</th>
<th>FG01 vs. FP01</th>
<th>FG02 vs. FP02</th>
<th>*Female Good vs. Female Poor</th>
<th>*Both Good vs. Both Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>g.65809482T&gt;C Allele model</td>
<td>P-value</td>
<td>1.11E-04</td>
<td>0.002</td>
<td>7.55E-07</td>
<td>9.10E-09</td>
<td>8.83E-07</td>
<td>4.75E-14</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>2.27</td>
<td>1.93</td>
<td>2.10</td>
<td>3.38</td>
<td>2.80</td>
<td>3.07</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.51–3.46</td>
<td>1.28–2.96</td>
<td>1.58–2.81</td>
<td>2.26–5.15</td>
<td>1.90–4.24</td>
<td>2.33–4.13</td>
</tr>
<tr>
<td>Dominant model</td>
<td>P-value</td>
<td>1.06E-04</td>
<td>0.004</td>
<td>1.77E-06</td>
<td>1.94E-07</td>
<td>1.30E-05</td>
<td>1.32E-11</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>3.31</td>
<td>2.43</td>
<td>2.84</td>
<td>5.63</td>
<td>4.23</td>
<td>4.88</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.80–6.16</td>
<td>1.31–4.71</td>
<td>1.87–4.35</td>
<td>3.02–12.20</td>
<td>2.33–8.53</td>
<td>3.07–8.01</td>
</tr>
<tr>
<td>recessive model</td>
<td>P-value</td>
<td>0.031</td>
<td>0.022</td>
<td>0.002</td>
<td>1.58E-04</td>
<td>5.94E-04</td>
<td>3.41E-07</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>2.41</td>
<td>2.71</td>
<td>2.54</td>
<td>4.05</td>
<td>3.50</td>
<td>3.76</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.00–6.20</td>
<td>1.17–7.65</td>
<td>1.38–5.25</td>
<td>2.01–9.48</td>
<td>1.70–8.20</td>
<td>2.39–6.75</td>
</tr>
<tr>
<td>g.65868604G&gt;T Allele model</td>
<td>P-value</td>
<td>1.19E-04</td>
<td>0.002</td>
<td>7.99E-07</td>
<td>1.69E-07</td>
<td>6.59E-06</td>
<td>5.79E-12</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>2.25</td>
<td>1.93</td>
<td>2.09</td>
<td>3.01</td>
<td>2.56</td>
<td>2.77</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.49–3.43</td>
<td>1.28–3.00</td>
<td>1.56–2.90</td>
<td>2.02–4.57</td>
<td>1.69–3.97</td>
<td>2.11–3.72</td>
</tr>
<tr>
<td>Dominant model</td>
<td>P-value</td>
<td>4.89E-05</td>
<td>0.004</td>
<td>9.19E-07</td>
<td>7.86E-07</td>
<td>4.38E-05</td>
<td>1.71E-10</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>3.57</td>
<td>2.43</td>
<td>2.94</td>
<td>5.18</td>
<td>3.89</td>
<td>4.49</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.92–6.75</td>
<td>1.36–4.41</td>
<td>1.83–4.62</td>
<td>2.78–10.16</td>
<td>2.14–8.40</td>
<td>2.86–7.73</td>
</tr>
<tr>
<td>recessive model</td>
<td>P-value</td>
<td>0.0528</td>
<td>0.0224</td>
<td>0.003</td>
<td>0.001</td>
<td>0.002</td>
<td>5.42E-06</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>2.16</td>
<td>2.71</td>
<td>2.40</td>
<td>3.34</td>
<td>3.18</td>
<td>3.26</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.00–5.41</td>
<td>1.00–8.54</td>
<td>1.29–4.73</td>
<td>1.59–7.52</td>
<td>1.64–7.88</td>
<td>2.01–5.63</td>
</tr>
<tr>
<td>g.66493737C&gt;T (1st intron in the MSTN gene) Allele model</td>
<td>P-value</td>
<td>1.40E-04</td>
<td>0.002</td>
<td>1.07E-06</td>
<td>3.22E-07</td>
<td>4.45E-05</td>
<td>8.12E-11</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>2.21</td>
<td>1.90</td>
<td>2.05</td>
<td>2.91</td>
<td>2.33</td>
<td>2.60</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.47–3.34</td>
<td>1.26–2.91</td>
<td>1.54–2.74</td>
<td>1.92–4.41</td>
<td>1.55–3.48</td>
<td>1.96–3.48</td>
</tr>
<tr>
<td>Dominant model</td>
<td>P-value</td>
<td>6.00E-05</td>
<td>0.008</td>
<td>2.32E-06</td>
<td>7.48E-07</td>
<td>1.99E-04</td>
<td>8.09E-10</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>3.67</td>
<td>2.34</td>
<td>2.92</td>
<td>5.68</td>
<td>3.84</td>
<td>4.69</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.97–7.93</td>
<td>1.28–4.56</td>
<td>1.86–4.65</td>
<td>2.84–13.23</td>
<td>1.64–9.31</td>
<td>2.81–8.20</td>
</tr>
<tr>
<td>recessive model</td>
<td>P-value</td>
<td>0.0441</td>
<td>0.0129</td>
<td>0.002</td>
<td>0.001</td>
<td>0.002</td>
<td>7.65E-06</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>2.13</td>
<td>2.71</td>
<td>2.38</td>
<td>3.19</td>
<td>3.05</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.00–5.24</td>
<td>1.15–6.49</td>
<td>1.33–4.35</td>
<td>1.51–7.52</td>
<td>1.59–7.59</td>
<td>1.96–5.42</td>
</tr>
<tr>
<td>g.66539967A&gt;G Allele model</td>
<td>P-value</td>
<td>0.001</td>
<td>0.002</td>
<td>6.91E-06</td>
<td>3.84E-07</td>
<td>1.66E-04</td>
<td>3.98E-10</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>1.98</td>
<td>1.89</td>
<td>1.93</td>
<td>2.94</td>
<td>2.22</td>
<td>2.56</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.26–2.97</td>
<td>1.31–2.84</td>
<td>1.47–2.58</td>
<td>1.92–4.49</td>
<td>1.50–3.38</td>
<td>1.94–3.44</td>
</tr>
<tr>
<td>Dominant model</td>
<td>P-value</td>
<td>0.004</td>
<td>0.011</td>
<td>1.22E-04</td>
<td>2.92E-04</td>
<td>0.001</td>
<td>1.12E-06</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>3.02</td>
<td>2.60</td>
<td>2.80</td>
<td>4.18</td>
<td>4.46</td>
<td>4.27</td>
</tr>
<tr>
<td></td>
<td>95% CI</td>
<td>1.37–6.60</td>
<td>1.25–6.06</td>
<td>1.68–4.88</td>
<td>1.84–12.05</td>
<td>1.96–12.37</td>
<td>2.48–8.33</td>
</tr>
</tbody>
</table>
vs. Both Poor (MP01 + MP02 + FP01 + FP02) (see Table 3). All populations except for the Both Good population ($P = 0.031$) were under HWE ($P > 0.05$). All comparisons between case–control populations showed statistical significance ($P < 0.05$) of the three models (i.e., allele, dominant and recessive models), indicating that the four SNPs are strongly associated with racing performance. Among the four SNPs, g.65809482T>C showed the strongest association with racing performance in Japanese Thoroughbreds.

Haplotype analysis among SNPs

The LD among the three SNPs g.65809482T>C, g.65868604G>T, and g.66493737C>T in the MSTN gene was remarkably high ($r^2 > 0.8$) (Table S2). Therefore, haplotype analyses using the three SNPs were performed in the case–control populations mentioned earlier, and the permutation $P$-values (10 000 re-samplings) showed statistical significance ($P < 1E-05$) for all the population pairs except for MG01 vs. MP01 ($P = 6.0E-4$) and MG02 vs. MP02 ($P = 3.6E-3$) (Table 3). The haplotypes T-G-C and C-T-T were the most common, while the T-G-C haplotype was most strongly associated with LE. The region defined by these SNPs encoded eight protein-coding genes, one pseudo gene, and two retrotransposons (Fig. S3). MSTN is a particularly strong candidate for this effect (Grobet et al. 1997; Mosher et al. 2007) in the genes, but other genes in the region are also worthy of further investigation. Among the eight protein-coding genes, three genes, namely, solute carrier family 40 member 1 (iron-regulated transporter 1), asparagine synthetase domain-containing protein 1, and ORM1-like protein 1 (Adoplin-1), are expressed in horse skeletal muscles based on CAGE (Cap Analysis of Gene Expression) analysis (Tozaki T., Hirota K. and Hasegawa T. in preparation). These results also suggest that genotyping for these three SNPs would be more effective than genotyping a single SNP at characterizing horses for this hereditary characteristic.

Comparison of the genotypes

The dominant models showed greater statistical significance than their recessive model counterparts (see Table 3), suggesting a dominant effect of the desirable alleles for racing performance. Hill et al. (2010) described the existence of the BRD among elite Thoroughbreds (experienced winners) by the differences in the genotype at the SNP locus g.66493737C>T and indicated that C/C horses are superior for short-distance races or sprints. Although our study design did not address BRD performance for Japanese Thoroughbreds, the following observations are pertinent. In accordance with the JRA racing programs, novice (2-year-old) horses are required to win one race in their first year to continue racing. The JRA races for novice and/or zero-winning horses are mainly short-distance races, for which the C allele described by Hill et al. (2010) is favourable. Therefore, winning and surviving Thoroughbreds in the current JRA racing programs may tend to have desirable alleles, such as the C allele at g.66493737C>T, because of the selection of those individual ranking up at the early stages of their athlete life. This would lead to the result that the desirable allele was correlated with good performance in the Japanese Thoroughbred population.

Conclusion

In this study, statistical significance ($P$-value) was shown on the genomic regions neighbouring the MSTN gene, suggesting that a gene in this region on ECA18 or the MSTN gene itself was one of the major genes affecting the racing performance of Thoroughbred racehorses. Because of the recent availability of over 50 000 SNPs on the commercial SNP-chip, it will be possible to perform a further detailed GWAS to identify the genes responsible for the racing performance in Thoroughbreds. Although this tool has great potential, we believe that the GWAS approach described here, using microsatellites for initial screening, remains
effective. We also propose the two-step selection procedure as a method to select research groups to avoid population stratification. These approaches would be useful in detecting such susceptibility regions in many livestock animals.

Acknowledgements

We thank Professor Dr. E. Bailey in University of Kentucky for helpful suggestions and advice which contributed to the improvement of the English in this manuscript. We thank the Ritto and Miho Training centres of the JRA for providing samples from their horses as study materials. This research was approved by the Equine Department of JRA and supported with a grant-in-aid (2008–2010).

Conflicts of interest

The authors have declared no conflicts of interest.

References


Supporting information

Additional supporting information may be found in the online version of this article.

Figure S1 Two-step selection for constructing subpopulations.
Figure S2 The differences in the total allele frequency distribution at a candidate locus by using pooled populations.
Figure S3 Genomic positions of the four candidate SNPs on ECA18.

Table S1 Primer sequences for asymmetric PCR by using high-resolution melting and unlabelled probe methods using LightScanner.

Table S2 SNPs tested for the racing performance susceptibility at the candidate regions on ECA1 and ECA18.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Refinement of quantitative trait loci on equine chromosome 10 for radiological signs of navicular disease in Hanoverian warmblood horses

M. S. Lopes*,†, U. Diesterbeck*, A. da Câmara Machado† and O. Distl*

*Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Bünteweg 17 p, 30559 Hannover, Germany.
†Biotechnology Centre of Azores, Department of Agriculture, University of Azores, Terra-Chã, 9701-851 Angra do Heroísmo, Portugal

Summary

Navicular disease is characterized by a progressive degenerative alteration of the equine podotrochlea. In this study, we refined a previously identified quantitative trait locus (QTL) on horse chromosome 10 for the abnormal development of canales sesamoidales (DCS) of the navicular bone in Hanoverian warmblood horses. Genotyping was done in 192 Hanoverian warmblood horses from 17 paternal half-sib groups. The whole marker set comprised 45 markers including seven newly developed microsatellites and 13 single nucleotide polymorphisms (SNPs) within positional candidate genes. Chromosome-wide significant QTL were confirmed and refined for DCS on horse chromosome (ECA) 10 at 0.16–2.70 Mb and at 14.45–36.37 Mb. Nine microsatellites and three SNP markers reached the highest multipoint Zmeans and LOD scores at 19.34–20.38 Mb and at 23.17–30.73 Mb with genome-wide error probabilities of \( P < 0.05 \). In addition, a significant association of a SNP within \( VSTM1 \) and a significant haplotype-trait association within \( IRF3 \) could be shown. These results support a possible role of the candidate genes \( VSTM1 \) and \( IRF3 \) within the QTL on ECA10 for DCS. This study is a further step towards the identification of the genes responsible for navicular disease in Hanoverian warmblood horses.

Keywords horse, horse chromosome 10, navicular disease, quantitative trait locus.

Introduction

Horse’s health is a fundamental prerequisite for its performance and durability in all sectors of the horse industry. Diseases of the locomotor system particularly interfere with the horse’s usability. Podotrochlosis, navicular disease or navicular syndrome is one of the main causes of chronic and often therapy-resistant forelimb lameness because of progressive degenerative alterations of the equine podotrochlea. The podotrochlea comprises the navicular bone, the bursa podotrochlearis and the insertion part of the deep digital flexor tendon. Horses affected by navicular disease show pathological changes that primarily affect the navicular bone, the navicular bursa and/or the distal end of the deep flexor tendon. Radiography of the navicular bone is the primary method to evaluate number, location, size and form of the nutrient foramina (canales sesamoidales) along the distal border of the navicular bone as well as the contour and structure of the navicular bone (Brunken 1986; MacGregor 1986; Dik & van den Broek 1995). Pathological signs include an increased number of canales sesamoidales, the appearance of branched canales sesamoidales, very deep or lollypop-shaped (bulbed ends) canales sesamoidales, a reduced radiographic density of the navicular bone (cyst-like lesion) and spurs at the margins of navicular bone (insertion desmopathy). The inheritance of pathological changes of the navicular bone has been shown by significant differences in prevalences between paternal progeny groups (Bos et al. 1986; Dik & van den Broek 1995; Stock & Distl 2006). Heritability estimates varied from 0.10 to 0.31 in threshold models (Diesterbeck et al. 2007). Furthermore, a recent study demonstrated quantitative trait loci (QTL) associated with radiological alterations in the navicular bone of Hanoverian warmblood horses (Diesterbeck et al. 2007). The phenotypic traits for navicular disease were based on the evaluation scheme of Brunken (1986). Diagnostic criteria were size, shape and distribution of canales...
Fine mapping of QTL for navicular disease

Materials and methods

Pedigree structure and phenotypic traits

A total of 192 Hanoverian horses were genotyped. These animals comprised 132 progeny and 12 grandchildren from 17 Hanoverian stallions, 36 dams of the progeny and 12 stallions out of these 17 Hanoverian sires. Diagnosis of alterations of the navicular bone was based on the radiologic examination of the dorsoproximo-palmarodistal projection (upright pedal route according to Oxspring 1935) of both front limbs. Radiologic appearance of the navicular bones was classified according to the evaluation scheme of Brunken (1986). Animals were classified as non-affected if (i) no canales sesamoidales distales; (ii) few (1–4) short indentations or conical canales sesamoidales distales; or (iii) few (1–4) elongated, narrow and straight canales sesamoidales distales were observed. Animals were classified as affected if (i) several (≥5) short indentations or conical canales sesamoidales distales; (ii) several (≥5) elongated, narrow and straight canales sesamoidales distales; (iii) few (1–4) lollypop-shaped and/or branched canales sesamoidales distales; (iv) several (≥5) lollypop-shaped and/or branched canales sesamoidales distales; (v) extensive osteolysis at margo distalis, no discernible canales sesamoidales distales; or (vi) canales sesamoidales proximales were visible.

Genotyping

For the refinement of the QTL on ECA10, the number of markers used was increased from 21 to 45. Four previously published microsatellites, seven newly developed microsatellites (ABGe351-ABGe357) and 13 SNPs (Table S1) were added to the previously used marker set. For the newly developed microsatellites, sequences were built with variations of di-, tri- and tetra-repeat motifs, with a minimum length of 15 repeats and a maximum length of 30 repeats, and were aligned with the horse genome assembly EquCab2.0 using the SSAHA2 package. Alignment results that obtained a maximum score per length (100% identity) were selected for primer design.

For SNP development positional and functional candidate genes for DCS in the identified QTL region on ECA10 were chosen using whole genome shotgun (WGS) sequences or equine ESTs which yielded significant BLAST hits to the syntenic region of ECA10 on human chromosome (HSA) 19 (http://www.ncbi.nlm.nih.gov/BLAST/) and the UCSC Horse genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway?hgsid=120144566&clade=mammal&org=Horse &db=0). Primers were designed using Primer3 after masking repetitive elements with RepeatMasker. Positions for all microsatellites and SNPs used were identified using the Basic Local Alignment Search Tool (BLAST) version 2.2.17) of NCBI.

Genotyping of microsatellites was carried out in a total volume of 12 μl containing 10 ng DNA, 8 μM dNTPs, 5% DMSO, 0.5 U Taq DNA polymerase (Qbiogene) in reaction buffer containing 15 μM MgCl₂. Primer concentration ranged from 2 to 10 pmol. All forward primers were fluorescently labelled at the 5’ end with IRD700 or IRD800. Reactions were performed in PTC 100™ or PTC 200™ thermocyclers (MJ Research) with 4-min denaturation at 94 °C, followed by 36 cycles of 30 s denaturation at 94 °C, 1 min annealing at the optimum temperature (Tₐ), 30 s extension at 72 °C. A final cooling step at 4 °C for 10 min

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 36–40
completed the PCR reaction. PCR products were size fractionated by gel electrophoresis on 6% polyacrylamide denaturing gels (Rotiphorese Gel40; Carl Roth) using an automated sequencer (LI-COR 4200/S-2 and 4300), and fragment lengths were determined using IRD700- and IRD800-fluorescence-labelled DNA ladders. Prior to electrophoresis, PCR products were diluted 1:10 with formamide loading buffer.

Screening for SNPs was performed by comparative sequencing of genomic DNA from eight unrelated stallions which sired eight of the 17 paternal progeny groups. PCRs for SNP genotyping were performed in a total volume of 30 μl containing 10 ng of genomic DNA as template, 30 pmol of each primer, 5 mM of each dNTP and 0.6 U Taq polymerase (Qbiogene) in reaction buffer. Amplification comprised an initial denaturation step at 94 °C for 4 min followed by 35 cycles of 94 °C for 30 s, 1 min annealing at the optimum temperature (T<sub>a</sub>), extension at 72 °C for 1 min and a final cooling step at 4 °C for 10 min. Amplified fragments were sequenced in both directions using the DYEnamic ET Terminator Cycle sequencing kit (GE Healthcare), and run on a MegaBACE 1000 automated sequencer (GE Healthcare). Sequence data was analysed using the Sequencher 4.7 program (GeneCodes). Genotyping of the identified SNPs was performed by sequencing of PCR products as described.

Data analysis

Mendelian inheritance and correctness of marker transmission in the pedigrees genotyped was confirmed using the Pedstats software (Wigginton & Abecasis 2005). Multipoint non-parametric linkage analysis (NPL) for 32 microsatellites and 13 SNPs was performed using the Merlin software (multipoint engine for rapid likelihood inference, version 1.1.2) (Abecasis et al. 2002). To estimate linkage between radiological findings and markers through the proportion of alleles shared by descendent (IBD) for affected animals, Zmean and LOD score were used as test statistics. The error probabilities (p) for linkage were also estimated according to Kong & Cox (1997). A multipoint chromosome-wide cosegregation of a marker allele among affected family members with the phenotypic expression of DCS in the navicular bones was assumed for P-values < 0.05. Genomewide probabilities were obtained by applying a Bonferroni correction: \( P_{\text{genome-wide}} = 1 - (1 - P_{\text{chromosome-wide}})^{1/r} \), where \( r \) = length of the respective equine chromosome in Mb (84 Mb according to the horse genome assembly EquCab2.0) divided by the total equine genome length (2680 Mb).

Statistical parameters of all microsatellite markers used, namely observed heterozygosity (HET) and polymorphism information content (PIC), Hardy–Weinberg equilibrium and genotypic and allelic associations, and the trend of the alleles with DCS as well as genotype-trait associations using \( \chi^2 \)-tests were calculated using the ALLELE, CASECONTROL and HAPLOTYPE procedures of SAS/Genetics (Statistical Analysis System, Version 9.2; SAS Institute).

Results

The increase of the marker density to a total of 45 markers confirmed the presence of the two QTL for DCS on ECA10 for Hanoverian warmblood horses. The average marker distance within identified QTL regions was 0.64 Mb (ABGe351-HMS023) and 1.29 Mb (ASB006-NVHEQ007). The average PIC of the marker set was 48.45% with a minimum of 8.66% and a maximum of 83.43%, while the mean observed HET was 54.28% ranging between 9.52% and 85.64%.

The non-parametric multipoint linkage analysis showed chromosome-wide significant Zmeans and LOD scores for DCS at 0.16–2.71 Mb and 14.45–36.37 Mb. The maximum (minimum) achievable Zmeans and LOD scores for deformed canales sesamoidales were 8.58 (–2.81) and 3.71 (–0.57) and thus high enough to reach genome-wide significance levels. Highest Zmean and LOD score were obtained at ABGe355 (Fig. 1; Table S2). After Bonferroni correction markers located within the region 19.34–20.38 Mb (SGCV30-LEX008) and markers in the QTL region at 23.17–30.73 Mb (PRPF31_Exon3:g.23166845G>A-ABGe355) reached the 5% threshold of the genome-wide error probability.

Association tests revealed a significant genotypic association (\( \chi^2 = 7.2; \, P = 0.03 \)) with DCS at 23.06 Mb (VSTM1_Intron2:g.23064508T>C). The genotype T/T was not affected by DCS, whereas the genotypes C/T and C/C
showed prevalences of 46.9% and 53.1%. For the SNP g.19440743G>A within IRF3 at 19.44 Mb, a significant allelic association ($\chi^2 = 5.0; P = 0.03$; odds ratio for DCS affection related with the allele A: 2.5) with DCS was evident. A significant haplotype-trait association ($\chi^2 = 8.49; P = 0.037$) was evident for the SNPs within IRF3 at 19.44 Mb (IRF3_Exon 7:g.19440743G>A. IRF3_Exon 7:g.19440761G>T). The haplotype G-G was significantly more frequent in horses free from DCS ($\chi^2 = 8.77; P = 0.003$) and the haplotypes A-G and G-T were significantly more prevalent in horses affected by DCS (A-G: $\chi^2 = 5.50; P = 0.02$; G-T: $\chi^2 = 3.58; P = 0.05$). These results corroborate the location of the QTL region at 14.45–36.37 Mb.

**Discussion**

The aim of this work was to refine QTL associated with the development of DCS of the navicular bone in Hanoverian warmblood horses on ECA10. In the whole genome scan previously performed for navicular disease in Hanoverian warmblood horses, relative marker positions were taken from horse linkage maps developed by Penedo et al. (2005) and Swinburne et al. (2006). The release of the horse genome assembly EquCab2 has made it possible to locate these microsatellites on the horse genome sequence. The QTL previously identified for DCS of the navicular bone in Hanoverian warmblood horses on ECA10 at 9.99 Mb and at 14.45–41.43 Mb could be confirmed, and through the increase of markers within and adjacent to these regions, refined. There were no indications for further QTL for DCS on ECA10 because Zmeans and LOD scores were far away from significance despite the high information content of the markers outside the QTL. Even if several markers were tested to close the gap of approximately 5 Mb between 2.71 and 7.55 Mb in our marker set, we could not find informative microsatellites in this region.

SNPs within the two candidate genes IRF3 and VSTM1 were in linkage disequilibrium with DCS, and this association across families indicates that these two candidates are worthwhile for further detailed analysis. Thus, this study presents an important step towards the identification of QTL-related candidate genes associated with DCS of the navicular bone in Hanoverian warmblood horses and the identification of further polymorphisms in these candidate genes should clarify their role for this syndrome. In addition, further SNPs have to be tested for association with DCS, particularly to identify candidate genes within the second QTL at the proximal end of ECA10. For this purpose, the SNP tables from the Broad Institute (http://www.broad.mit.edu/ftp/distribution/horse_snp_release/v2/) can be employed, although it is not assured that the SNPs reported are informative in the data used. Further potential candidate genes within the QTL may also include collagen, type XII, alpha1 (COL12A1), myeloid-associated differentiation marker (MYADM), osteoclast associated immunoglobulin-like receptor (OSCAR), interleukin 11 (IL11) and collagen, type X, alpha1 (COL10A1). Furthermore, it can be helpful to use the Equine Articular Cartilage cDNA Library (http://www.ncbi.nlm.nih.gov/sites/entrez) to identify genes which are expressed in the cartilage. There are ten ESTs located within the QTL region at 0.16–2.70 Mb and 160 ESTs located in the QTL region at 14.45–36.37 Mb on ECA10. This study is a further step towards unravelling the mechanisms that regulate the development of deformed canales sesamoidales of the navicular bone in Hanoverian warmblood horses.

**Acknowledgement**

M.S. Lopes was supported by DRCT M3.1.1/I/017A/2005.

**Conflicts of interest**

The authors have declared no potential conflicts.

**References**


**Supporting information**

Additional supporting information may be found in the online version of this article.

---

**Table S1** Primer sequences, product size in basepairs (bp), annealing temperature (Ta) and the corresponding location in the human gene, polymorphism information content (PIC) and observed heterozygosity (HET) for 13 intragenic single nucleotide polymorphisms (SNPs) in Hanoverian Warmblood horses.

**Table S2** Multipoint non-parametric chromosome-wide test statistics (Zmean and LOD scores) and their chromosome-wide significant error probabilities ($P_Z$, $P_L$) for deformed canales sesamoidales (DCS) of the navicular bone in Hanoverian warmblood horses.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Molecular heterogeneity of XY sex reversal in horses

T. Raudsepp*, K. Durkin*†, T. L. Lear‡, P. J. Das*, F. Avila*, P. Kachroo* and B. P. Chowdhary*

*Department of Veterinary Integrative Biosciences, College of Veterinary Medicine, Texas A&M University, College Station, TX-77843-4458, USA. †Unit of Animal Genomics, GIGA-R, Department of Animal Sciences, Faculty of Veterinary Medicine, University of Liège, Liège, Belgium. ‡M.H. Gluck Equine Research Center, Veterinary Science Department, University of Kentucky, Lexington, KY, USA

Summary
Male-to-female 64,XY sex reversal is a frequently reported chromosome abnormality in horses. Despite this, the molecular causes of the condition are as yet poorly understood. This is partially because only limited molecular information is available for the horse Y chromosome (ECAY). Here, we used the recently developed ECAY map and carried out the first comprehensive study of the Y chromosome in XY mares (n = 18). The integrity of the ECAY in XY females was studied by FISH and PCR using markers evenly distributed along the euchromatic region. The results showed that the XY sex reversal condition in horses has two molecularly distinct forms: (i) a Y-linked form that is characterized by Y chromosome deletions and (ii) a non-Y-linked form where the Y chromosome of affected females is molecularly the same as in normal males. Further analysis of the Y-linked form (13 cases) showed that the condition is molecularly heterogeneous: the smallest deletions spanned about 21 kb, while the largest involved the entire euchromatic region. Regardless of the size, all deletions included the SRY gene. We show that the deletions were likely caused by inter-chromatid recombination events between repeated sequences in ECAY. Further, we hypothesize that the occurrence of SRY-negative XY females in some species (horse, human) but not in others (pig, dog) is because of differences in the organization of the Y chromosome. Finally, in contrast to the Y-linked SRY-negative form of equine XY sex reversal, the molecular causes of SRY-positive XY mares (5 cases) remain as yet undefined.

Keywords deletions, horse, male-to female XY sex reversal, SRY, Y chromosome.

Introduction
Male-to-female sex reversal in horses is a developmental disorder in which phenotypic females or female-like animals are genetically males with 64,XY karyotype. The condition has been found in several breeds (Kent et al. 1986, 1988b; Makinen et al. 1999; Bugno et al. 2003; Switonski et al. 2005), and after X chromosome monosomy, it is the most common sex chromosome abnormality in horses (Lear & Bailey 2008; Villagomez et al. 2009). The male-to-female sex reversal has been described in humans (known also as Swyer syndrome) (Michala et al. 2008), cattle, river buffalo, and sheep (Ferrer et al. 2009; Villagomez et al. 2009).

Phenotypes of the XY sex reversal horses can vary in a broad range, from a very feminine to a greatly masculinized mare (Kent et al. 1988b). The majority of ‘feminine’-type animals have normal female external and internal genitalia, and no somatic or behavioural abnormalities (Power 1986; Bowling et al. 1987). Primary cause of infertility of these mares is typically because of ovarian dysgenesis and underdeveloped uterus (Power 1986; Pailloux et al. 1995; Makinen et al. 1999). The overall phenotype of the ‘feminine’-type XY mares is very similar to those of 64, XO females (Bowling et al. 1987).

Another less frequently observed phenotype shows various degrees of masculinization and virilization (Kent et al. 1986; Long 1988; Lear & Bailey 2008; Villagomez et al. 2009). The ‘masculine’-type XY mares usually have abnormally developed genital tract, and the gonads can range from ovotestes to testicular feminization (Kent et al. 1986; Bowling et al. 1987; Kent et al. 1988a; Howden 2004; Villagomez et al. 2009), leading to male pseudohermaphrodites. All XY female horses studied so far have been found to be infertile, except one that gave birth to a normal 64,XX filly (Sharp et al. 1980).
Most equine male-to-female sex reversal cases are sporadic, or have no supporting pedigree data. However, in some instances, familiar patterns of inheritance, particularly in certain Arabian sire lines, have been observed (Kieffer 1976; Kent et al. 1986; Bowling et al. 1987; Kent et al. 1988b).

Despite the relatively frequent occurrence of XY mares among animals with chromosome abnormalities, the molecular causes of the condition are poorly understood. One of the reasons is the very limited knowledge about the molecular organization of the horse Y chromosome. The only molecular test, initiated in 1995 (Pailhoux et al. 1995), is PCR with the Y-linked male sex-determining gene SRY. According to this, two molecularly distinct types of XY females have been identified. The most prevalent are mares with no SRY gene (Pailhoux et al. 1995; Abe et al. 1999; Makinen et al. 1999; Bugno et al. 2003). Another type, represented by a single case, is an SRY-positive mare (Switonski et al. 2005).

In this study, we carried out the first comprehensive molecular analysis of the Y chromosome in XY sex-reversed mares. This was performed by sequencing the bacterial artificial chromosome (BAC) clones containing the SRY-region (Raudsepp et al. 2004b; Raudsepp & Chowdhary 2008b; Paria 2009) and developing markers that precisely defined the deletion in sex-reversed individuals.

### Materials and methods

#### Animals

The study involved 18 female horses from various breeds. The animals were referred in 2001–2009 to the Texas A&M Molecular Cytogenetic Laboratory because of infertility and/or abnormal genitalia (Table 1). The laboratory received peripheral blood samples from each individual in sodium heparin and EDTA Vacutainers (VACUTAINER™, Becton Dickinson).

#### Cell cultures and chromosome analysis

Cell cultures and chromosome preparations were made using standard cytogenetic methods as previously described (Raudsepp & Chowdhary 2008a). For karyotyping, the slides were stained with Giemsa, the sex chromosomes were identified by C-banding (Arrighi & Hsu 1971), and at least 20 cells were analysed for each technique.

#### Fluorescence in situ hybridization, FISH

The approximate size of Y chromosome deletions in XY females were determined by FISH using selected BAC clones.

### Table 1 Summary of the phenotypes and genotypes of the female horses used in this study.

<table>
<thead>
<tr>
<th>No</th>
<th>Breed</th>
<th>Phenotype</th>
<th>Karyotype</th>
<th>PCR with SRY</th>
<th>STS content analysis</th>
<th>FISH analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Thoroughbred</td>
<td>Abnormal cervix, small ovaries (7–12 mm)</td>
<td>64,XYdel</td>
<td>Neg</td>
<td>Deleted Contig I and</td>
<td>Con I neg, the rest of Y pos</td>
</tr>
<tr>
<td>2</td>
<td>Thoroughbred</td>
<td>Small uterus and ovaries</td>
<td>64,XYdel</td>
<td>Neg</td>
<td>Deleted all MSY and PAR</td>
<td>All negative</td>
</tr>
<tr>
<td>3</td>
<td>Arabian</td>
<td>Male-like behaviour</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>4</td>
<td>Standardbred</td>
<td>Small uterus and ovaries, not cycling</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>5</td>
<td>Thoroughbred</td>
<td>Small uterus and ovaries, not cycling, aggressive in behaviour</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>6</td>
<td>Quarter Horse</td>
<td>Small uterus and ovaries, not cycling, aggressive in behaviour</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>7</td>
<td>Thoroughbred</td>
<td>Small stature, gonadal dysgenesis, suspected XO</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>8</td>
<td>Thoroughbred</td>
<td>Small ovaries (20 mm), no follicular activity</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>9</td>
<td>Arabian</td>
<td>N/a</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>10</td>
<td>Thoroughbred</td>
<td>Large stature, small ovaries (10 mm)</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>11</td>
<td>Thoroughbred</td>
<td>Immature reproductive tract, small ovaries</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>12</td>
<td>Thoroughbred</td>
<td>Normal stature, small ovaries (10 mm)</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>13</td>
<td>Standardbred</td>
<td>No pregnancies</td>
<td>64,XY</td>
<td>Neg</td>
<td>Deleted ~21 kb around SRY</td>
<td>All positive</td>
</tr>
<tr>
<td>14</td>
<td>Appaloosa</td>
<td>Normal external genitalia, hypoplastic uterus, underdeveloped mammary</td>
<td>64,XY</td>
<td>Pos</td>
<td>No ECAY deletions</td>
<td>All positive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>glands. Shows oestrous behaviour to stallions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Mixed warmblood</td>
<td>Abnormal: uterus, vulva located ventrally between rear legs at the location of mammary glands.</td>
<td>64,XY</td>
<td>Pos</td>
<td>No ECAY deletions</td>
<td>All positive</td>
</tr>
<tr>
<td>16</td>
<td>Standardbred</td>
<td>Male pseudohermaphrodite</td>
<td>64,XY</td>
<td>Pos</td>
<td>No ECAY deletions</td>
<td>All positive</td>
</tr>
<tr>
<td>17</td>
<td>Standardbred</td>
<td>Male pseudohermaphrodite</td>
<td>64,XY</td>
<td>Pos</td>
<td>No ECAY deletions</td>
<td>All positive</td>
</tr>
<tr>
<td>18</td>
<td>Quarter Horse</td>
<td>Male pseudohermaphrodite</td>
<td>64,XY</td>
<td>Pos</td>
<td>No ECAY deletions</td>
<td>All positive</td>
</tr>
</tbody>
</table>
Table 2 Markers and BAC clones used for FISH analysis.

<table>
<thead>
<tr>
<th>ECAY contig</th>
<th>Marker symbol</th>
<th>BAC clone</th>
<th>BAC library</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAR</td>
<td>PRKXY</td>
<td>164O9</td>
<td>CHORI-241</td>
<td>(Raudsepp &amp; Chowdhary 2008a,b)</td>
</tr>
<tr>
<td>V</td>
<td>ZFY</td>
<td>110.3H12</td>
<td>TAMU</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td>IV</td>
<td>AMELY</td>
<td>180P20</td>
<td>CHORI-241</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td>III</td>
<td>NLGNY</td>
<td>112E12</td>
<td>CHORI-241</td>
<td>(Paria 2009)</td>
</tr>
<tr>
<td>II</td>
<td>YE1</td>
<td>118L7</td>
<td>TAMU</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td>YSB12</td>
<td></td>
<td>190M2</td>
<td>CHORI-241</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td>I</td>
<td>KDM5D</td>
<td>60D8</td>
<td>CHORI-241</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td>SRY</td>
<td></td>
<td>140M23</td>
<td>CHORI-241</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24I23</td>
<td>CHORI-241</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>49.3F11</td>
<td>TAMU</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>79.4H1</td>
<td>TAMU</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>107.3H9</td>
<td>TAMU</td>
<td>(Raudsepp et al. 2004a,b)</td>
</tr>
</tbody>
</table>

from the physical map of ECAY (Raudsepp et al. 2004b; Raudsepp & Chowdhary 2008b) (Table 2). BAC DNA was isolated using the Plasmid Midi Kit (Qiagen) according to the manufacturer’s instructions. The BAC DNA was labelled and hybridized to metaphase chromosomes of XY females and control males following our standard FISH protocol (Raudsepp & Chowdhary 2008a). The results were analysed with a Zeiss Axioplan2 fluorescent microscope equipped with Isis V5.2 (MetaSystems GmbH) software.

Sequencing

Initial Y chromosome sequence information around the SRY gene was obtained by internal sequencing of two SRY-containing BAC clones, viz., 140M23 and 24I23 (CHORI-241 BAC library: http://bacpac.chori.org/quine241.htm). The first-step sequencing primers were designed in SRY-coding sequence and used for sequencing in both 3’ and 5’ directions. The newly obtained sequences served as templates to design primers for the next step. Such ‘walking’ was repeated several times. Sequencing reactions were carried out in 10-μl volume with 1 μg of BAC DNA template and BigDye chemistry and resolved on an ABI-3730 capillary sequencer.

Next, the two SRY-containing BAC clones, viz., 140M23 and 24I23, were completely sequenced at Broad Institute of MIT and Harvard (Cambridge, MA, USA) using Sanger sequencing technology. The draft sequence GenBank Accessions are AC214740 and AC215855, respectively. Additionally, a third SRY-containing BAC clone, 107.3H9 from TAMU BAC library (http://hbz7.tamu.edu/homelinks/bac_est/bac.htm), was completely sequenced at Research and Testing Laboratory (Lubbock) following standard Roche 454-sequencing protocols (Roche). A total of 260 150 sequences with an average length of 421 bp were derived and initially assembled using NGen (DNastar). The contigs were aligned with SeqMan, resulting in a continuous 116 857-bp sequence (GenBank HM103387). Sequences of the three BACs and all available STSs in the SRY-region (Raudsepp et al. 2004b) were aligned using Sequencher 4.7 (Gene Codes) software and analysed using RepeatMasker (http://www.repeatmasker.org/), GENSCAN (http://genes.mit.edu/GENSCAN.html) and NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Finally, 1 185 bp of the single exon of SRY (Table 3), including the 684-bp open reading frame, was sequenced from all SRY-positive sex-reversal females using BigDye chemistry. The sequences were aligned with the reference sequence (GenBank AB004572) using Sequencher 4.7 (Gene Codes) software.

Development of STS markers and STS content analysis

Initial analysis of the Y chromosome in XY mares was carried out by PCR using primers for the horse SRY gene and selected STS markers from ECAY maps (Raudsepp et al. 2004b; Raudsepp & Chowdhary 2008b) (Table 3). Thereafter, additional STS primers were designed from masked (RepeatMasker: http://www.repeatmasker.org/) sequences of the three completely sequenced BAC clones. Androgen receptor (AR) from the X chromosome (Raudsepp et al. 2004a) was used as a positive control for all PCR amplifications. A summary of the PCR primers developed and used in this study is presented in Table 3. The same primers were used for sequencing the PCR products and/or for internal BAC sequencing in the SRY-region.

STS content analysis was carried out on the genomic DNA isolated from EDTA-stabilized peripheral blood of the 18 XY females (Table 1). DNA from ‘Bravo’, the donor for CHORI-241 BAC library, served as the male control. Similarly, DNA from ‘Twilight’, the donor for the genome sequence project (Wade et al., 2009), served as the female control. STS content analysis for each animal was carried out in 10-μl volume duplicate PCR with 0.25 units Taq polymerase (JumpStart RedTaq, Sigma Aldrich), 50 mM KCl, 10 mM Tris–HCl (pH 8.4), 0.2 mM dATP, dCTP, dGTP and dTTP, 0.3 μM of each primer, and 50 ng of genomic DNA as template. The PCR products were resolved in 2% agarose gels containing ethidium bromide.

Results

Cytogenetic analysis

Cytogenetic analysis of Giemsa-stained metaphase spreads revealed that all 18 female horses (Table 1) had normal diploid chromosome number (2n = 64) with the XY sex chromosome complement, which is typical of males. This was confirmed by C-banding, which clearly showed the presence of one X chromosome and one Y chromosome in all cells analysed (Fig. 1a-d). Animals 1 and 2 had extremely small Y chromosomes (Fig. 1c, d), compared to other XY females and male controls. In contrast, the
Table 3 Summary data for horse Y chromosome genes and STS markers.

<table>
<thead>
<tr>
<th>Marker</th>
<th>ECAY region</th>
<th>Primer Sequences 5′–3″</th>
<th>PCR product size, bp</th>
<th>Reference</th>
<th>GenBank Accession No</th>
</tr>
</thead>
<tbody>
<tr>
<td>107-STS1</td>
<td>Con I: SRY-region</td>
<td>F: TGGGCAGTTAGAAAAATTCAAG R: CCTGCAACACTGTCCTGTAATTG</td>
<td>121</td>
<td>This study</td>
<td>HM103387</td>
</tr>
<tr>
<td>107-STS2</td>
<td>Con I: SRY-region</td>
<td>F: TTTTGTTGGTCATATTGGG R: ATCTACATGCTGGGGGCAAAC</td>
<td>381</td>
<td>This study</td>
<td>HM103387</td>
</tr>
<tr>
<td>107-STS3</td>
<td>Con I: SRY-region</td>
<td>F: TCACGTGAGAAACATTCTGG R: CAAGCAAAACAGATTGGAGCA</td>
<td>196</td>
<td>This study</td>
<td>HM103387</td>
</tr>
<tr>
<td>107-STS4</td>
<td>Con I: SRY-region</td>
<td>F: CATGGAGGAGAAGCCTTTTG R: TCAGCCTGAGTTTCCCAAT</td>
<td>425</td>
<td>This study</td>
<td>HM103387</td>
</tr>
<tr>
<td>110.3H12-M13</td>
<td>Con V</td>
<td>F: GGCCCAAGATATGCAAGGA R: GATGTTGTTGTTGCTCTGT</td>
<td>182</td>
<td>Raudsepp et al. 2004a,b</td>
<td>BV140834</td>
</tr>
<tr>
<td>81F8-SP6</td>
<td>Con III</td>
<td>F: GCAAAGGCTCTGAGAGAGGA R: CCATGCTTCTTGACCAGACA</td>
<td>182</td>
<td>Paria 2009</td>
<td>CU011557</td>
</tr>
<tr>
<td>118L7-T7</td>
<td>Con II</td>
<td>F: ATCTGCTCCTTTTGGTTTT R: CCCCCAGATTTACTGCCTTTG</td>
<td>225</td>
<td>Paria 2009</td>
<td>CU025611</td>
</tr>
<tr>
<td>126G2-T7</td>
<td>Con II</td>
<td>F: GCAACTTGCACTGATTGTCC R: ATTTGTGGAGGGCAGGT</td>
<td>200</td>
<td>This study</td>
<td>n/a</td>
</tr>
<tr>
<td>131N23-SP6</td>
<td>Con V</td>
<td>F: GCTTGTCCTAATTTGCTCTTG R: CTTTGGAGACGGGTAAATTG</td>
<td>152</td>
<td>Raudsepp &amp; Chowdhary 2008a,b</td>
<td>CU034701</td>
</tr>
<tr>
<td>168I4-T7</td>
<td>Con II</td>
<td>F: CCATTTGCACTCCCTGATCCA R: AAGGTTAGTAGGAGAGGGAAGG</td>
<td>101</td>
<td>Raudsepp et al. 2004a,b</td>
<td>BV140781</td>
</tr>
<tr>
<td>24I23-T7</td>
<td>Con I: SRY-region</td>
<td>F: ATCGTCTCCTGACCCTCCTTG R: GGTGACCCCTGCTTCTCGT</td>
<td>101</td>
<td>Raudsepp et al. 2004a,b</td>
<td>BV140808, CT942288, AC215855</td>
</tr>
<tr>
<td>26821-SP6</td>
<td>Con I</td>
<td>F: CAAGGAAGCCAGGAAGAGTG R: GTCTGCCCTGGCAGGATGAT</td>
<td>160</td>
<td>Paria 2009</td>
<td>CT951493</td>
</tr>
<tr>
<td>3∗STS4</td>
<td>Con I: SRY-region</td>
<td>F: ATGTGCCAGCTCTCCTTTTGTCTT R: ATGATGTTGTCGGCCATCTGA</td>
<td>690</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>3∗STS5</td>
<td>Con I: SRY-region</td>
<td>F: AAAACAGAGGTTGGCCCCCTTGA R: GCCGCTGAGAAGAAATGACCA</td>
<td>602</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>3∗STS6</td>
<td>Con I: SRY-region</td>
<td>F: ATTTTTCCCTCTGATCTCTT</td>
<td>838</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>3∗STS7</td>
<td>Con I: SRY-region</td>
<td>F: TCCTGCTTTTCTTATGACTC R: CCTGAAGAGCTGTTCTAAATGGA</td>
<td>797</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>3∗STS8</td>
<td>Con I: SRY-region</td>
<td>F: CAGGAGCTTTTGAAAGATGG R: TTGTGACCTTGCTGAATGTT</td>
<td>596</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>3∗STS9</td>
<td>Con I: SRY-region</td>
<td>F: TGAACCAACAGGGCCATTITT R: GACACGTGTAATAATGGGGGAAAGA</td>
<td>218</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>3∗STS10</td>
<td>Con I: SRY-region</td>
<td>F: TTCTCATACTGGGCCCATAGACTACA R: ACCTGCAAAACTCCCTCGTG</td>
<td>492</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>3∗STS12</td>
<td>Con I: SRY-region</td>
<td>F: GTGATTATGGGGCTTGTCGACT</td>
<td>399</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>3∗STS13</td>
<td>Con I: SRY-region</td>
<td>F: TGGAGCTCAGAGGGCAGGTA R: GCCTGCTAACTTGGAAGAAAACA</td>
<td>509</td>
<td>This study</td>
<td>AC214740</td>
</tr>
<tr>
<td>419P11-T7</td>
<td>PAR</td>
<td>F: ACAGGGGACGTCCTCACAAGA R: CCCAGAAAAGGCCGTAAG</td>
<td>151</td>
<td>Raudsepp &amp; Chowdhary 2008a,b</td>
<td>CU267996</td>
</tr>
<tr>
<td>49.3F11-M13</td>
<td>Con I: SRY-region</td>
<td>F: TGAAATCACAAGCATACACAAATG R: GGTCACTGAGAACCATCTGCCTG</td>
<td>172</td>
<td>Raudsepp et al. 2004a,b</td>
<td>BV140820</td>
</tr>
<tr>
<td>5.2A8-M13</td>
<td>Con IV</td>
<td>F: TCAGAATGGGACCTGCTCAAAR R: GGCCCGTGGCTTTTCTTTATT</td>
<td>274</td>
<td>This study</td>
<td>n/a</td>
</tr>
<tr>
<td>5∗STS1</td>
<td>Con I: SRY-region</td>
<td>F: ACTCCTCCCTCCAACACGT R: ATTTTAAGGAGGCGCAGAC</td>
<td>150</td>
<td>This study</td>
<td>AC214740, AC215855</td>
</tr>
</tbody>
</table>
Table 3 (continued).

<table>
<thead>
<tr>
<th>Marker</th>
<th>ECAY region</th>
<th>Primer Sequences 5’-3’</th>
<th>PCR product size, bp</th>
<th>Reference</th>
<th>GenBank Accession No</th>
</tr>
</thead>
</table>
| 5’STS2   | Con I: SRY-region | F: GCTCTGCCCCCTCCTAAAAAT  
R: AACGATGATGTCCACCATCATGAAT | 226 | This study | AC214740,  
AC215855 |
| 5’STS3   | Con I: SRY-region | F: TAGTTTAGCTCGGCGCGCAT 
R: TGGCATCTGACCATGCTACT | 347 | This study | AC214740,  
AC215855 |
| 5’STS4   | Con I: SRY-region | F: CAAACAACCACTTTAAACAA 
R: CATACCCTATTTTATTGGAATCGTA | 177 | This study | AC214740,  
AC215855 |
| 5’STS5   | Con I: SRY-region | F: TGTACATGTTGGGAAGTGG  
R: ATTAGAAGGCGCTTCCAAGA | 399 | This study | AC214740,  
AC215855 |
| 5’STS6   | Con I: SRY-region | F: TCACCAAGTGGCATCTGGA  
R: TCCCCAACATGTACACCTCA | 174 | This study | AC214740,  
AC215855 |
| 5’STS7   | Con I: SRY-region | F: AATGACGACATTTTAAAAGTACAAGTT 
R: TTTTCTCATTGGGGAAGTGG | 153 | This study | AC214740,  
AC215855 |
| 5’STS8   | Con I: SRY-region | F: TTCTAGGTCATCATTGGTGCTC  
R: TGGCTATCTTCAAAAGACATTAAGG | 848 | This study | AC214740,  
AC215855 |
| 5’STS9   | Con I: SRY-region | F: GCCATTGTTCCTCTTGTGTC  
R: TGGCTATCTCCAAAAGACATTAAG | 645 | This study | AC214740,  
AC215855 |
| 5’STS10  | Con I: SRY-region | F: TAGCTCTTGGCCTTCTCTT  
R: TTTTCTCATTGGGGAAGTGG | 188 | This study | AC214740,  
AC215855 |
| 5’STS12  | Con I: SRY-region | F: GCCCTTCATTCCTTTTCCTT  
R: AAAAGGCAGCTTCCAGGAT | 754 | This study | AC214740,  
AC215855 |
| 5’STS13  | Con I: SRY-region | F: GTTTTCCCAGTGGTGGTGG  
R: GGTCAAATTGGGTATGGTCAA | 755 | This study | AC214740,  
AC215855 |
| 5’STS14  | Con I: SRY-region | F: CACTAAACTGCAAAGACATTTTTCAT 
R: CAGGACCGTGGTGACAGTAA | 192 | This study | AC214740,  
AC215855 |
| 5’STS16  | Con I: SRY-region | F: TTGATGTTGATTCTGGAAAGAAG  
R: TGGGATCTGACTTCAGAAA | 552 | This study | AC214740,  
AC215855 |
| 5’STS17  | Con I: SRY-region | F: AGCACCGCAGATGTGAAAGA  
R: GGAGGCAAGAGATCCCTTG | 679 | This study | AC214740,  
AC215855 |
| 66M24-SP6| Con III | F: TTGGGTCTCCTCATTCAAGTGA  
R: GGATTGTGATGTCATCTCG | 169 | This study | CU001288 |
| 79.4H1-M13| Con I: SRY-region | F: AGTCCACACCAACCCACAGTGA  
R: TGAAGGGAAGTGGGGATTTTC | 248 | Raudsepp et al.  
2004a,b | BV140821 |
| 90B11-SP6| Con IV | F: TTGCTTTGGATTCTGGAAAGAAG  
R: TGGGATCTGACTTCAGAAA | 150 | This study | BV140791,  
CU015692 |
| AMELY    | Con IV | F: CCAACCCCAACACCACCGCACAAGGGGAAT  
R: AGCATAGGGGGGGAAGGCTGCA | 160 | Hasegawa et al.  
2000 | AB032194 |
| KDM5D    | Con I | F: AACACCGACCAATTTTTT  
R: GCAAAATTTCTGGGAAATCCA | 400 | Agulnik et al.  
1997; Lindgren et al.  
2001 | EU687564,  
U52364 |
| NLGN4Y   | Con III | F: GGGGATCCATCTTTTGTGGT  
R: GCTCAACAGACGCCTCTGCAC | 156 | Paria 2009 | EU687560 |
| PRKXY    | PAR | F: CCGTTTCTGACTGTTGAGGAGG  
R: TCTTGGGCAACCCGAAGTGGTGA | 229 | Raudsepp & Chowdhary  
2008a,b | ETO52965 |
| RBMY     | Con I: SRY-region | F: TCTGCCCTCTCTCTCTGAGCAT  
R: ACTCAAGGACCGCAATGAT | 180 | Paria 2009 | EU687561 |
| SH3B14   | Con I | F: GTGACCTCCACAGGAGCTGAG  
R: TCTGGCTATGCTCTGGTGA | 486 | Wallner et al.  
2004 | BV005745 |
| SRY1     | Con I | F: TGCCATTTACTCTGCTGCTT  
R: ATGGCAATTTTCTCGGTTCC | 200 | Paria 2009 | EU687565 |
Y chromosome in Animal 15 was larger than usual with a prominent and extended heterochromatic region (Fig. 1b). The presence of the Y chromosome was further confirmed by FISH using microdissected horse Y chromosome (Raudsepp & Chowdhary 1999) as a probe (Fig. 2a–b). Hybridization results confirmed the cytogenetic observations that the Y chromosomes in Animals 1 and 2 were abnormally small (Fig. 2b). Moreover, FISH with 8 BAC clones evenly distributed along the ECAY contig map (Table 2, Fig. 3a) showed that the two rudimentary Y chromosomes are molecularly different. The Y chromosome in Animal 1 gave FISH signals with all contig II–V and PAR BACs (Fig. 2c, d) and was negative only for contig I clones containing SRY and KDM5D (alias SMCY). In contrast, no FISH signals with any of the probes were observed in Animal 2, indicating that the rudimentary Y chromosome is comprised of little or no euchromatin (Fig. 3b). In the remaining 16 animals, FISH signals were observed with all MSY and PAR BACs, including the five overlapping clones in the SRY-region (Table 2, Fig. 2e, f; Fig. 3c). Taken together, the

**Table 3 (continued).**

<table>
<thead>
<tr>
<th>Marker</th>
<th>ECAY region</th>
<th>Primer Sequences 5’–3’</th>
<th>PCR product size, bp</th>
<th>Reference</th>
<th>GenBank Accession No</th>
</tr>
</thead>
</table>
| SRY2   | Con I       | F: CTTAAGCTTCTGCTATGTCCAGAAGTATCC  
R: GGGGAGGGAGGAATGTAACATAA | 1,185 | This study | EU687565, AC214740, AC215855 |
| Y3B12  | Con II      | F: GGGAGGCACTTGGAAAGTACA  
R: GGTGAGGAATGCACTGGAG | 400 | Wallner et al. 2004 | G72338 |
| YE1    | Con II      | F: CTTCACTCCGGACCAAGA  
R: GTGTCGTCGCTGGGCTTAAC | 199 | Wallner et al. 2004 | BV005727 |
| YM2    | Con I       | F: TGGCTAGATGTGTTTGGTTCTTAC  
R: TTTGCAAGCCAGCTACTACCTCCTT | 119 | Wallner et al. 2004 | BV005725 |
| ZFY    | Con V       | F: TGCACATTCTCCTTAATCT  
R: GCACATTTAAAGGAAACCTT | 342 | Lindgren et al. 2001 | AF178847, AF133198 |

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 41–52

**Figure 1** C-banded metaphase spreads (a–b) and Giemsa-stained karyotypes (c–d) of Animal 6 (a), Animal 15 (b), Animal 1 (c), and Animal 2 (d).
cytogenetic and FISH analyses detected massive Y chromosome deletions over 10–15 Mb size (Raudsepp et al. 2004b) in two animals, whereas the gross organization of the Y chromosomes in all other XY mares appeared to be normal.

Sequence analysis of the Y chromosome

Analysis by PCR with SRY primers divided the 18 XY females into two molecularly distinct groups: SRY negative (Animals 1–13, Table 1) and SRY positive (Animals...
14–18). All XY mares were positive for the AR control gene. Next, we used a panel of 20 MSY and PAR markers (Fig. 3a) and studied by PCR their presence or absence in individual animals. As expected, all tested markers were negative in Animal 2, confirming the FISH results that the small Y chromosome is exclusively heterochromatic. In Animal 1, only markers from contigs II–V and the PAR were present. No amplification by PCR was observed with the 4 markers in contig I: YM2, KDM5D, SH3B14, and 26B21-SP6 (Fig. 3a). Thus, the PCR and FISH results for Animal 1 were in complete agreement. Similarly, PCR analysis confirmed and refined FISH results for the remaining animals. All 20 ECAY markers were positive in Animals 3–18, showing that regardless of the presence or absence of SRY, the overall integrity of the Y chromosome was retained.

Next, we investigated whether the deletion of SRY is the only molecular signature of the Y chromosome in the 11 SRY-negative females (Animals 3–13). We hypothesized that if only the single exon of SRY (Hasegawa et al. 1999) is deleted, it should be possible to design a simple PCR test to distinguish between normal males and XY SRY-negative females. We reasoned that if primers are designed to flank the deletion, distinctly different-sized PCR products would be amplified from normal males and XY females. Such tests will show the size of deleted segments in sex-reversed females and will be a useful alternative to the current presence- or absence-based PCR analysis with SRY primers.

To develop this test, we carried out a detailed STS content analysis with a panel of 34 markers in a 146-kb region around SRY that was developed from three completely sequenced BAC clones (Fig. 3d). The analysis revealed that the deletion is not restricted to the SRY exon and includes at least 21 500 bp around it (Fig. 3d). We discovered that the deletion is flanked by approximately 28 kb of directed repeats (R1 and R2 arrows in Fig. 3d) that are present both in normal males and XY females. As the duplicated regions share a high degree of sequence similarity (approximately 96–100%), it was not possible to determine by PCR whether one or both segments are present in sex-reversed animals. As shown in Fig. 3d, the unique MSY-specific sequences (blue shades), which are suitable for primer design, are located tens of kilobase pairs away from the deletion. Thus, it was not possible to design primers for an alternative PCR test to distinguish between Y chromosomes in normal males and SRY-negative females.

Sequence analysis of the 146 432-bp genomic segment around the deletion (Fig. 3d) identified three MSY genes – SRY, RBMY, and ATP6V0CY. The single-copy SRY is part of the deletion, a multicopy gene RBMY (Paria 2009) is located in both directionally duplicated segments on either side of the deletion (Fig. 3d), and ATP6V0CY is a newly found Y-linked gene in horses.

Finally, FISH and STS content analyses did not detect any molecular differences between the Y chromosomes of SRY-positive XY females and normal males. Likewise, sequences of the single exon of SRY in all 5 SRY-positive females were 100% identical with each other and with the male reference sequence (data not shown). Therefore, the detailed analysis of the Y chromosome in XY females showed that the molecular causes of SRY-positive male-to-female sex reversal in horses are not Y-linked and remain as yet undefined.

**Discussion**

To our knowledge, this study represents the first comprehensive molecular characterization of the Y chromosome in XY sex-reversed female horses. Using recently available molecular tools for MSY and the PAR (Raudsepp et al. 2004b; Raudsepp & Chowdhary 2008b; Paria 2009), we considerably refined the underlying molecular aetiology of the SRY-negative (Pailhoux et al. 1995; Abe et al. 1999; Makinen et al. 1999, 2001; Bugno et al. 2003) form of sex reversal in horses.

**Molecular heterogeneity of ECAY in SRY-negative XY females**

A combination of cytogenetic, FISH, STS content, and sequence analysis revealed that the SRY-negative form of sex reversal is a molecularly heterogenous condition. Deletions that lead to the loss of SRY can be massive, including megabase pairs of DNA (Fig. 3b), or limited to a region around SRY, removing tens of kilobase pairs of DNA (Fig. 3d). Notably, none of the deletions is restricted to the SRY sequence only. Furthermore, as illustrated by Animals 1 and 2 (Table 1), the massive deletions differ from each other both in size and content (Fig. 3b). The loss of entire ECAY euchromatin in Animal 2 makes this case genetically very similar to X monosomy. This is in agreement with the animal’s phenotype (Table 1) and supports earlier observations that mares with XY and XO sex chromosome complements are frequently indistinguishable by overall appearance (small stature), behaviour (abnormal oestrus behaviour), and gonadal phenotype (underdeveloped ovaries and uterus) (Chandley et al. 1975; Bowling et al. 1987; Long 1988). However, as none of the earlier cytogenetic studies report about finding an abnormally small Y chromosome, the molecular cause of the XO-like phenotype in XY mares is probably not because of the massive loss of Y euchromatin. Female gonadal dysgenesis might rather be because of the absence of the second X chromosome.

Systematic mapping of the horse Y chromosome and the development of Y-linked markers started only recently (Raudsepp et al. 2004b). This is why all earlier molecular studies of XY mares have been limited to a few Y-linked genes, viz., SRY (Pailhoux et al. 1995; Abe et al. 1999; Makinen et al. 1999, 2001; Bugno et al. 2003), AMELY, STS-Y and ZFY (Makinen et al. 2001; Bugno et al. 2003). These tests confirmed the absence of SRY, and the presence of the three other markers, but had no knowledge about the
relative order or map location of these genes. Therefore, the present study is the first in which the Y chromosomes of XY mares were analysed using detailed map information (Raudsepp et al. 2004b; Raudsepp & Chowdhary 2008b) and partial sequence data for ECAY. This allowed fine demarcation of ECAY deletions (Fig. 3) and showed that all SRY-negative mares, except Animals 1 and 2, share the same large 21.5-kb deletion that includes SRY (Fig. 3d) but no other known functional MSY genes. The results are in agreement with the common idea that the female-like phenotype or the lack of male phenotype in SRY-negative mares is primarily because of the loss of SRY (Pailhoux et al. 1995; Makinen et al. 1999). However, the cause of female gonadal dysgenesis, another characteristic feature of both the XY and XO mares (Table 1), is not so clear. In females with X monosomy, haploinsufficiency for certain PAR genes has been considered as the cause of gonadal dysgenesis (Ellison et al. 1997; Blaschke & Rappold 2001). In this study, using SRY and a panel of 20 Y-linked markers (Fig. 3a), we showed that, except for the SRY-region, the rest of the Y chromosome is undisturbed. Thus, in contrast to XO females, the SRY-negative XY mares have two normal copies of the PAR. Therefore, given that the molecular causes of gonadal dysgenesis in XY and XO females are the same, they are more likely associated with incomplete dosage of some non-PAR Xp genes that escape X inactivation.

Possible molecular mechanisms of Y chromosome deletions

Previously, it has been proposed that SRY-negative sex-reversal syndrome in horses is because of abnormal X-Y recombination in male meiosis, resulting in the loss of SRY from the Y to the X chromosome (Abe et al. 1999; Makinen et al. 2001; Bugno et al. 2003). The assumption was based on observations in humans where such uneven crossovers are common and give rise to 46,XY SRY-negative females and 46,XX SRY-positive males (Rosser et al. 2009). However, it must be noted that the molecular organization of the human and horse Y chromosomes are very different. In humans, the SRY is located only 35 kb away from the pseudoautosomal boundary (PAB) (Skaletsky et al. 2003), which makes uneven X-Y meiotic exchange feasible. In contrast, the horse SRY is located in the very proximal part of MSY (Raudsepp et al. 2004b), far from the PAB (Fig. 3a), and its involvement in abnormal meiotic exchange is unlikely (Raudsepp & Chowdhary 2008b).

A more plausible explanation of Y deletions lies in the intrinsic nature of Y chromosome sequences. A typical feature of the haploid and non-recombining Y chromosome is the presence of massive and numerous reversed (palindromic) and directional repeats (Skaletsky et al. 2003; Lange et al. 2009; Hughes et al. 2010). Sequence similarity between palindrome arms or repeated units might be as high as 99%–100%, facilitating inter- and intra-chromatid gene conversion and recombination (Lange et al. 2009). As such genetic exchanges help to maintain Y chromosome structural and functional integrity, the majority of important spermatogenesis genes are located in palindrome arms or directional repeats (Skaletsky et al. 2003; Lange et al. 2009).

However, this mechanism also has a downside: inter-chromatid recombination can frequently lead to duplications and deletions which, in humans, are associated with spermatogenic failure, sex reversal, and Turner syndrome (Lange et al. 2009). In this study, we show that in the horse Y chromosome, the SRY-region is surrounded by two almost 100% identical directional repeats (Figs 3d & 4a). Not coincidentally, both repeated segments contain one copy of a known spermatogenesis-related gene, RBMY (Elliott 2004). It is possible that the deletion of the SRY-region, as seen in SRY-negative XY mares, is the result of an inter-chromatid recombination between the two repeats (Fig. 4b). Theoretically, this inter-chromatid exchange should remove from the Y chromosome of XY females not only SRY but also one of the directed repeats (Fig. 4c). However, this can be proven only by direct Y chromosome sequencing, and not by PCR. If our assumption is correct, the SRY-region is expected to become duplicated on another chromatid (Fig. 4c) and should be present in the Y chromosomes of paternal male full- or half-sibs of the XY females. Possibly, duplication of SRY and triplication of repeats have no significant phenotypic effects, and the carriers remain undetected. Further, we infer that the molecular mechanism of the two massive deletions in Animals 1 and 2 might be quite similar to that presented in Fig. 4, with the only difference that recombination events have taken place between more distantly located repeats resulting in larger deletions.

Taken together, the key factors causing SRY-negative sex reversal are Y-linked, although some X-linked genes might also contribute to the phenotype. Thus, the primary mutation, sporadic or genetically predisposed, is derived paternally. Indeed, as shown in earlier studies, there are certain sire lines that produce more XY females than expected by chance (Kieffer 1976; Kent et al. 1986; Bowling et al. 1987; Kent et al. 1988b). However, the 13 SRY-negative XY mares analysed in this study were from different breeds (Table 1) and families and were therefore not informative regarding the pedigree analysis.

The SRY-positive male-to-female sex reversal

The SRY-positive females are phenotypically distinct from the SRY-negative mares, as shown by a single previous report (Switoski et al. 2005) and this study. These mares display features that are not typically observed in SRY-negative animals, such as stallion-like behaviour, large body size, abnormal external genitalia, elevated blood testosterone levels, and male rather than female gonadal dysgenesis (Table 1). Based on this, it is possible that many more SRY-positive XY females have been described earlier.
but because of the lack of molecular tools, they were not properly identified (Kieffer 1976; Kent et al. 1986, 1988a,b; Crabbe et al. 1992; Howden 2004). Therefore, the ‘masculine’-type equine XY sex reversal might be more prevalent than the single previous report (Switonski et al. 2005) and the 5 cases described in this study.

Detailed molecular analysis of the Y chromosome showed that the ‘feminine’ (SRY-negative) and ‘masculine’ (SRY-positive) XY sex reversals differed genetically more than just the absence or presence of SRY. While SRY-negative animals showed various ECAY deletions, the Y chromosome of SRY-positive mares was the same as in normal males; even the unusually large Y chromosome in Animal 15 was molecularly normal by PCR, FISH, and sequence analysis, and the abnormal size could be attributed to substantially enlarged heterochromatic region (Fig. 1b). We infer that the SRY-positive condition is likely not Y-linked. However, human studies suggest that a Y-linked ‘growth’ gene might be responsible for the larger than usual body size (adult height) (Ogata & Matsuo 1992; McDonough 2003) of XY females. Further, it has been proposed that testicular feminization and male pseudohermaphroditism in SRY-positive mares might be caused by a mutation in the X-linked androgen receptor (AR) gene (Crabbe et al. 1992; Howden 2004; Switonski et al. 2005). To date, there is no experimental evidence to support this. It is also possible that, similarly to humans, the SRY-positive sex reversal is a genetically heterogeneous disorder (Sarafoglou & Ostrer 2000). In humans, some cases are Y-linked, showing different missense and frameshift mutations in the SRY-coding region (Salehi et al. 2006; Shahid et al. 2008; Marchina et al. 2009). However, almost 80% of 46,XY women have no mutations in SRY or other Y-linked genes, indicating the involvement of X-linked or autosomal factors. The few different autosomal (Barburo et al. 2009; Biason-Lauber et al. 2009; Schimmer & White 2010) and X-linked (Sarafoglou & Ostrer 2000) mutations described in XY women do not give consistent answers, and the underlying genetic causes of SRY-positive sex reversal in humans and horses alike remain poorly understood.

Occurrence of XY sex reversal in horses and other species

Male-to-female XY sex reversal, especially the SRY-negative form, is a relatively frequent chromosomal abnormality in horses. Incidence of XY mares in our cytogenetic analysis practice is as high as 26% among all chromosomally abnormal animals studied during the period between 2001 and 2009. This is in line with earlier data showing that 12%–30% of cytogenetic abnormalities in horses count for XY females (Bowling et al. 1987; Power 1990). The condition has also been described in humans (Sarafoglou & Ostrer 2000; Michala et al. 2008), where about 10%–20% XY women are SRY negative, and 80–90% have normal SRY. Both SRY-positive and SRY-negative XY females have been found in cattle (Ferrer et al. 2009; Villagomez et al. 2009) and mouse (Arnold & Chen 2009). In contrast, surprisingly few and exclusively SRY-positive ‘masculine’-type XY females have been reported for other species, such as river buffalo (Di Meo et al. 2008), dog (Nowacka-Woszuk et al. 2007; Whyte et al. 2009), and sheep (Ferrer et al. 2009). Notably, no XY sows have been found in pigs (Villagomez et al. 2009). This disparity might be because of uneven cytogenetic sampling of populations in different species. However, given the well-organized cytogenetic screening system of domestic species, especially pigs, in France and many other European countries (Ducos et al. 2008), this is unlikely. Instead, we hypothesize that the prevalence of Y-linked XY SRY-negative condition is related to the diversity of the organization of the Y chromosome in different species (Skaletsky et al. 2003; Raudsepp et al. 2004b; Murphy et al. 2006; Hughes et al. 2010). For example, SRY is a single-copy gene in human (Skaletsky et al. 2003), mouse (Landrigan & Tucker 1997), and horse (Paria 2009). This implies that any Y chromosome rearrangement that causes the loss of SRY will result in an SRY-negative condition. In contrast, species like cat (Pearks Wilkerson et al. 2008), rabbit (Gerdales & Ferrand 2006), rat (Turner et al. 2007), and several other rodents (Landrigan & Tucker 1997) have multiple copies of SRY. Consequently, deletion of one SRY copy leaves other copies intact to carry out their function. Therefore, one plausible reason why species like pigs, dogs or sheep have no SRY-negative XY females might be the presence of multicity SRY. It could also be that SRY location in relation to inverted and directional repeats is different across species, thus facilitating SRY deletions in some species but not in others. It is anticipated that the ongoing mapping and sequencing projects for the
mammalian Y chromosomes will soon provide better explanations for these ideas. Likewise, the availability of genome-wide analysis tools, such as SNP-chips and tiling arrays, are expected to improve the discovery of genes and rearrangements underlying the SRY-positive form of sex reversal in horses and other species.

Acknowledgements

This work was supported by USDA grant 2006-04801 and CVMBS LINK Endowment for Equine Genomics. The University of Kentucky Agricultural Experiment Station number for this paper is 10-14-077.

Conflict of interest

The authors have declared no potential conflicts.

References


Paria N., 2009. Discovery of Candidate Genes for Stallion Fertility, Ph.D. Thesis, College of Veterinary Medicine, Texas A&M University, College Station. p. 223.
Genetic diversity in the Maremmano horse and its relationship with other European horse breeds

M. Felicetti*,1, M. S. Lopes†, A. Verini-Supplizi*, A. da Câmara Machado†, M. Silvestrelli*, D. Mendonça† and O. Distl‡

*Department of Pathology, Diagnostic and Veterinary Clinic, University of Perugia, Via San Costanzo 4, 06126 Perugia, Italy.
†Biotechnology Centre of Azores, Department of Agriculture, University of Azores, Terra-Câmara, 9701-851 Angra do Heroísmo, Portugal.
‡Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Bünneweg 17p, 30559 Hannover, Germany

Summary

The Maremmano is an Italian warmblood horse breed from central Italy. We characterized the genetic diversity and the degree of admixture in Maremmano in comparison to 14 other European horse breeds using 30 microsatellites. Between-breed diversity explained about 9 per cent of the total genetic diversity. Cluster analysis, genetic distances and genetic differentiation coefficients showed a close relationship of Maremmano with Hanoverian and Lusitano in accordance with breed history.

Keywords clustering analysis, Equus caballus, Maremmano.

The Maremmano is an Italian warmblood horse mostly bred in the provinces of Grosseto and Viterbo (Central Italy). It is believed that the origin of this horse breed goes back to local Etruscan horse populations which were crossbred with modern breeds in the last centuries. Pedigree analysis of the Maremmano horses revealed four male lines contributing 11.1% to the genetic diversity of all Maremmano stallions (Silvestrelli 1991). In 1980, Maremmano breeders established a stud book for Maremmano.

The objective of this study was to investigate the genetic structure and the degree of admixture of Maremmano together with the Italian nucleus of Lipizzan and Lusitano as well as twelve further horse breeds previously characterized by Aberle et al. (2004). This data set consisted of Hanoverian, Arabian, Exmoor, Icelandic, Przewalski, Sorraia and all German coldblood horse breeds. We used the same marker set of 30 autosomal microsatellites and the same PCR conditions and reference samples as described elsewhere (Aberle et al. 2004).

A total of 146 animals that were not closely related were genotyped in this study: Maremmano (n = 50), the Italian nucleus of Lipizzan (n = 49) and Lusitano (n = 47) from Portugal. DNA was extracted from EDTA-blood using standard methods.

Genetic variability of the breeds genotyped in this study was determined by the mean number of alleles, observed and expected heterozygosities, breed-specific alleles and the molecular variance (AMOVA) using GENALEX 6 (Peakall & Smouse 2006). Excess and deficiency of heterozygotes, which are deviations from Hardy-Weinberg equilibrium, were estimated using GENEPOP (Raymond & Rousset 1995). Molecular genetic relationships among populations were estimated using Wright’s FST and Nei’s standard genetic distance (GST; Nei 1973) by bootstrapping 1000 replicates using MICROSOAT (Minch 1997). Phenograms based on Nei’s GST and genetic distances among all 549 animals were drawn using the unweighted pair group method with arithmetic mean algorithm (UPGMA) by PHYLIP (Felsenstein 1989) and displayed by TREEVIEW (Page 1996). The Bayesian clustering procedure of STRUCTURE was employed to investigate the genetic structure and the degree of admixture of the 15 horse breeds (Pritchard et al. 2000). A 20 000 initial burn-in was used, followed by 100 000 MCMC iterations as recommended by Falush et al. (2007) with 10 independent replicates each. All runs used an admixture model with correlated frequencies and the parameter of individual admixture alpha.

The mean number of alleles was 4.7 for the Lipizzan, 6.7 for the Lusitano and 7.3 for the Maremmano. For Maremmano and Lipizzan breeds the total observed heterozygosity was higher than the expected, whereas for Lusitano breed

Address for correspondence

Prof. Dr O. Distl, Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Bünneweg 17p, 30559 Hannover, Germany.
E-mail ottmar.distl@tiho-hannover.de

1These authors have contributed equally to this work.

Accepted for publication 20 September 2010
19 out of the 30 loci showed observed heterozygosity values lower than the expected ones (Table S1).

The AMOVA indicated that for Maremmano, Lipizzan and Lusitano 9% of the total genetic variability is attributed to significant differences between the horse breeds, whereas 91% of the observed variation was from within the breeds. The AMOVA performed for the five riding horse breeds (Maremmano, Lipizzan, Lusitano, Arabian, Hanoverian), considered separately from the others, showed 11% of the genetic variability due to breed differences. When the German coldblood breeds were included, the variation among breeds increased only slightly to 12%; with all 15 breeds the total genetic variability rose to 14.6%.

Maremmano showed the lowest genetic differentiation with the Hanoverian (5.5%); in addition, the Lipizzan and Lusitano, when matched with the Maremmano, showed the lowest $F_{ST}$ values (11.4% and 6%, respectively). Among the three breeds genotyped here, the Maremmano showed the least differentiation when compared to the coldblood breeds (Table S2). Similar results were obtained when the genetic differentiation based on Nei’s $G_{ST}$ among breed pairs was used. The only exception was for the Maremmano, which showed lowest genetic differentiation (14.7%) with the Lusitano.

The phylogenetic tree based on Nei’s $G_{ST}$ (Fig. 1) and the dendrogram based on the proportion of shared alleles (Fig. S1) displayed three main clusters representing riding horses, the Exmoor, Przewalski and Sorraia group and the German draught horses. The Icelandic horses did not cluster with any of the other breeds.

Clustering using STRUCTURE separated for $K = 3$ horses into riding horse breeds, ancient and isolated breeds, and German draught horses. For $K = 9$, Maremmano clustered together with Lusitano, but Hanoverian and Arabian were separate clusters. When $K = 14$, Maremmano and Lusitano also clustered in their own pre-defined populations (Fig. S2; Table S3).

From the three breeds genotyped in this study, both the Maremmano and the Lusitano showed a high level of genetic variability and similar to that observed for the same breeds in other studies (Luı́s et al. 2007; Zuccaro et al. 2008). The high level of genetic differentiation observed for the Maremmano may partly reflect contributions from several breeds, although a significant proportion of the Maremmano is descending from a reduced number of male lines. The Italian nucleus of Lipizzan horses showed low levels of variation similar to those observed by Achmann et al. (2004). The small number of founders resulting in a small effective population size and the traditional pure-breeding system within a nucleus without any crossbreeding may explain the reduced variability within the Lipizzan. In agreement with the results of Luı́s et al. (2007) in the Lusitano, levels of observed heterozygosity were lower than their expected counterpart.

Pairwise $F_{ST}$ values among Maremmano, Lipizzan, Lusitano and Hanoverian were in a similar range as those among closely related coldblood breeds (Aberle et al. 2004; Druml et al. 2007). Inclusion of Hanoverian and German coldblood increased the proportion of variance among breeds only slightly, indicating a close relationship of these breeds with Maremmano, Lipizzan and Lusitano in contrast to Arabian, pony and primitive horse breeds. Two genetic differentiation measures and the two model-based clustering approaches also revealed a genetic proximity of Maremmano with the Hanoverian and the Lusitano. Two male lines founded by thoroughbred stallions (Aiace and Ingres) and a Trakehner stallion might have created the relationship between Maremmano and Hanoverian, as thoroughbred and Trakehner stallions have been intensely used in the Hanoverian warmblood (Hamann & Distl 2008). Regarding the close genetic proximity between Maremmano and Lusitano, it is believed that Iberian horses, the ancestors of the Lusitano and the Andalusian, were in the Stato dei Reali Presidi di Spagna (1557–1800), located between Stato Pontificio and Granducato di Toscana, and therefore might have influenced the founder lines of the Maremmano breed.
In conclusion, the Maremmano retained a high genetic diversity and the results reported here can be used to prevent genetic erosion of the Maremmano breed.

Acknowledgements

Thanks to Dr Luca Buttazzoni for technical support and Mr Gianluca Alunni for expert technical assistance. This work was supported by SELMOL-MIPAF (Silvestrelli) and by FTC and FEDER within the projects POCTI/CVT/41890/2001 and POCl2010 Ref. GG/GGP/ME611. CBA-UAc was supported by FCT and DRCT and Lopes by DRCT M3.1.1/I/017A/2005.

Conflict of interest

The authors have declared no potential conflicts.

References


Supporting information

Additional supporting information may be found in the online version of this article.

Figure S1 UPGMA dendrogram constructed from allele-sharing distances among 549 animals from 15 different horse breeds.

Figure S2 Graphical presentation of the population structure analyses for a sample of 549 horses from 15 different horse breeds obtained by STRUCTURE.

Table S1 Number of alleles, observed and expected heterozygosity for Maremmano, Lusitano and Italian Lipizzan horses (146 horses) based on 30 microsatellite loci.

Table S2 Nei’s standard genetic distance and Wright’s distance among 15 horse breeds.

Table S3 Estimated memberships to inferred clusters obtained by STRUCTURE and individual assignments of 549 horses according to their own predefined breed or to another breed as obtained using GeneClass2.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance

E. W. Hill*, J. Gu*, B. A. McGivney* and D. E. MacHugh*†
*Animal Genomics Laboratory, UCD School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland. †UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland

Summary

Athletic performance is influenced by a complex interplay among the environment and a suite of genes, which contributes to system-wide structure and function. In a panel of elite and non-elite Thoroughbred horses (n = 148), we genotyped 68 SNPs in 17 putative exercise-relevant genes chosen from a genome scan for selection. We performed a series of case–control and quantitative association tests for relationships with racecourse performance. Thirteen SNPs in nine genes were significantly (P < 0.05) associated with a performance phenotype. We selected five SNPs in four genes (ACSS1, ACN9, COX4I1, PDK4) for validation in an independent sample set of elite and non-elite Thoroughbreds (n = 130). Two SNPs in the PDK4 gene were validated (P < 0.01) for associations with elite racing performance. When all samples were considered together (n = 278), the PDK4_38973231 SNP was strongly associated (P < 0.0005) with elite racing performance. Individuals with the A:A and A:G genotypes had a 16.2–16.6 lb advantage over G:G individuals in terms of handicap rating. Re-sequencing of the PDK4 gene and further genotyping will be required to identify the causative variant that is likely influencing exercise-induced variation in expression of the gene. Notwithstanding, this information may be employed as a marker for the selection of racehorses with the genetic potential for superior racing ability.

Keywords Thoroughbred, exercise, SNP, association test, gene.

Introduction

Intense selection for athletic phenotypes in the Thoroughbred horse (Equus caballus) during the past 300 years has resulted in structural and functional system-wide adaptations that have significantly enhanced the physiological characteristics that enable elite athletic performance (Constantinopol et al. 1989; Jones et al. 1989; Evans et al. 1993). The athleticism of the Thoroughbred is attributed to a range of extreme physiological characteristics including a large muscle mass to body weight ratio, high skeletal muscle mitochondrial density and oxidative enzyme activity and considerable intramuscular stores of energy substrates (Hinchcliff et al. 2008).

While the phenotypic adaptations to elite athleticism in Thoroughbred horses are well described, the understanding of the molecular contributions to such exquisitely adapted exercise-related phenotypes is still in its infancy (Harrison & Turrion-Gomez 2006; Eivers et al. 2009; Gu et al. 2009; McGivney et al. 2009; Hill et al. 2010). Genetic contributions to athletic performance phenotypes in humans are well documented and more than 220 genes have been described (Bray et al. 2009). Although it is likely that Thoroughbred racing performance is also influenced by a large number of genes, only one performance-associated sequence variant in an exercise-relevant gene has previously been reported for the horse (Hill et al. 2010). The recent and strong selection for exercise-related traits has left signatures in the genome of the Thoroughbred. In a population genetics-based genome scan, positively selected loci have been identified in the extreme tail-ends of the distributions for statistics (FST and Ewens–Watterson test) that identify departures from patterns of genetic variation expected under neutral genetic drift (Gu et al. 2009). Such outlier approaches have led to an understanding of the selective forces that have shaped the recent evolution of human populations (Akey 2009; Oleksyk et al. 2010; Pritchard...
et al. 2010). Within the positively selected genomic regions, enrichments for genes involved in insulin signalling, fat substrate utilization and muscle strength have been identified. Genes in these functional categories likely play key roles in contributing to the lean, muscular, athletic phenotype that is typical for Thoroughbreds.

Genomic regions that have been targets for selection represent the most likely regions to contain structural genetic variation contributing to functional and phenotypic variance in exercise-relevant traits. Therefore, to identify genes that represent the most likely targets for selection, we investigated whether genes within these regions may contain sequence variants that contribute to the genetic variation in racetrack performance in the Thoroughbred population (Gaffney & Cunningham 1988).

We interrogated the EquCab2.0 SNP database for Thoroughbred SNPs located within the genomic sequence of 20 putative exercise-relevant genes located within the top-ranked outlier genetic regions previously described (Gu et al. 2009). A panel of 68 SNPs in 17 genes was selected for genotyping. To investigate associations between the sequence variants and racing phenotypes, we genotyped a group of Thoroughbred horses \( n = 148 \) and performed a series of population-based case–control genetic association investigations by separating the samples on the basis of retrospective racecourse performance. In addition, we performed quantitative trait association tests using best race distance and handicap rating (Racing Post Rating, RPR) as phenotypes.

**Materials and methods**

**Genomic DNA samples**

More than 1400 registered Thoroughbred horse tissue samples (hair or fresh blood) were collected from stud farms, racing yards and sales establishments in Ireland and New Zealand between 1997 and 2009. Genomic DNA was extracted from either fresh whole blood or hair samples using a modified version of a standard phenol/chloroform extraction method (Sambrook & Russell 2001).

Horses were categorized based on retrospective racecourse performance records as ‘Elite’ Thoroughbreds (TBE) or ‘Other’ Thoroughbreds (TBO). To minimize confounding effects of racing over obstacles, only horses with performance records in Flat races were considered for inclusion in the study cohorts. Elite Thoroughbreds were Flat racehorses that had won at least one Group race (Group 1, Group 2 or Group 3) or a Listed race – the highest standard and most valuable elite Flat races are known as Group races and Listed races are the next in status. Other Thoroughbreds had competed in at least one race, but had never won a race and had handicap (Racing Post Rating, RPR) ratings <80. Race records were derived from three sources. Europe race records were derived from the Racing Post on-line database (http://www.racingpost.com); Australasia and South East Asia race records were derived from Arion Pedigrees (http://www.arion.co.nz); and North America race records were derived from the Pedigree Online Thoroughbred database (http://www.pedigreequery.com). In all cases, pedigree information was used to control for genetic background by attempting to exclude samples sharing relatives. Over-representation of popular sires within the pedigrees was avoided where possible.

**Study sample set**

A panel of Thoroughbred samples \( n = 148 \) was selected from the repository and separated into two distinct performance cohorts; elite (TBE, \( n = 86 \); mean RPR = 115) and other (TBO, \( n = 62 \); mean RPR = 59) (Table 1). In each cohort, there was no sharing of sires or dams. The elite performer group contained horses (of which 84 were Group race winners) that competed in a total of 1170 races and won 425, including 215 Group races, of which 91 were at Group 1 level. The other (non-elite) cohort competed in 537 races and won 15, none of which was a Group race.

**Validation sample set**

An additional \( n = 130 \) Thoroughbred samples were genotyped for six SNPs. The samples were subdivided into elite \( (n = 97; \text{mean RPR} = 113) \) and other \( (n = 33; \text{mean RPR} = 49) \) cohorts. There was some sharing of sires within and among the validation set cohorts (Table 1).

<table>
<thead>
<tr>
<th>Table 1 Thoroughbred sample sets.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
</tr>
<tr>
<td>Study set</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Validation set</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>All</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

The original sample set contained unrelated individuals that were categorized based on retrospective racing performance as elite (TBE) or other (TBO). The validation sample set contained some sharing of sires within each performance cohort.

RPR, Racing Post Rating handicap rating.
<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene description</th>
<th>Gene ontology: Biological Process</th>
<th>KEGG pathway</th>
<th>Chr</th>
<th>Distance from marker (Mb)</th>
<th>Dh/sd</th>
<th>P</th>
<th>P_C</th>
<th>P &lt; 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACN9</td>
<td>ACN9 homolog (S. cerevisiae)</td>
<td>GO:0005996~monosaccharide metabolic process</td>
<td>hsa00010:Glycolysis/ Gluconeogenesis</td>
<td>4</td>
<td>1.66</td>
<td>−6.12</td>
<td>&lt;0.001</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>ACSS1</td>
<td>Acyl-CoA synthetase short-chain family member 1</td>
<td>GO:0006732~coenzyme metabolic process</td>
<td>hsa05110:Vibrio cholerae infection</td>
<td>22</td>
<td>0.55</td>
<td>NS</td>
<td>NS</td>
<td>0.42</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>ACTA1</td>
<td>Actin, alpha 1, skeletal muscle</td>
<td>GO:0006936~muscle contraction</td>
<td>hsa04510:Focal adhesion</td>
<td>1</td>
<td>2.83</td>
<td>NS</td>
<td>NS</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>ACTN2</td>
<td>Actinin, alpha 2</td>
<td>GO:0006936~muscle contraction</td>
<td>hsa04510:Focal adhesion</td>
<td>1</td>
<td>3.67</td>
<td>NS</td>
<td>NS</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>ADHFE1</td>
<td>Alcohol dehydrogenase iron-containing protein 1</td>
<td>GO:0005114 oxidation reduction</td>
<td>hsa00010:Glycolysis/ Gluconeogenesis, hsa00071: fatty acid metabolism, hsa04614:Renin-angiotensin system</td>
<td>9</td>
<td>0.11</td>
<td>−9.44</td>
<td>&lt;0.001</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>AGT</td>
<td>Angiotensinogen (serpin peptidase inhibitor, clade A, member B)</td>
<td>GO:0001944~vasculature development</td>
<td>hsa0100:Metabolic pathways</td>
<td>1</td>
<td>4.05</td>
<td>NS</td>
<td>NS</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>COX4I1</td>
<td>Cytochrome c oxidase subunit IV isoform 1</td>
<td>GO:0004129 cytochrome c oxidase activity</td>
<td>hsa0100:Metabolic pathways</td>
<td>3</td>
<td>3.53</td>
<td>−4.72</td>
<td>0.002</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>CYP51A1</td>
<td>Cytochrome P450, family 51, subfamily A, polypeptide 1</td>
<td>GO:0005114 oxidation reduction</td>
<td>hsa0100:Metabolic pathways</td>
<td>4</td>
<td>2.71</td>
<td>−6.12</td>
<td>&lt;0.001</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>GGPS1</td>
<td>Geranylgeranyl diphosphate synthase 1</td>
<td>GO:0004255~cellular lipid metabolic process</td>
<td>hsa0100:Metabolic pathways</td>
<td>1</td>
<td>4.77</td>
<td>NS</td>
<td>NS</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>GSN</td>
<td>Gelsolin (amyloidosis, Finnish type)</td>
<td>GO:0007015~regulation of actin cytoskeleton</td>
<td>hsa04810:Regulation of actin cytoskeleton</td>
<td>25</td>
<td>0.69</td>
<td>−6.78</td>
<td>&lt;0.001</td>
<td>0.35</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>MTFR1</td>
<td>Mitochondrial fission regulator 1</td>
<td>GO:0005739 mitochondrial regulation of mitochondrial fission</td>
<td>hsa00071:Vibrio cholerae infection</td>
<td>9</td>
<td>0.52</td>
<td>−9.44</td>
<td>&lt;0.001</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>MUSK</td>
<td>Muscle, skeletal, receptor tyrosine kinase</td>
<td>GO:0007517~muscle organ development</td>
<td>hsa00071:Vibrio cholerae infection</td>
<td>25</td>
<td>9.86</td>
<td>−6.78</td>
<td>&lt;0.001</td>
<td>0.35</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>NDUF4B</td>
<td>NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 8, 19 kDa</td>
<td>GO:0006119~oxidative phosphorylation</td>
<td>hsa0190:Oxidative phosphorylation,</td>
<td>25</td>
<td>0.06</td>
<td>−6.78</td>
<td>&lt;0.001</td>
<td>0.35</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>PDK4</td>
<td>Pyruvate dehydrogenase kinase, isozyme 4</td>
<td>GO:0005996~monosaccharide metabolic process</td>
<td>hsa00010:Glycolysis/ Gluconeogenesis</td>
<td>4</td>
<td>0.38</td>
<td>−6.12</td>
<td>&lt;0.001</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>PON1</td>
<td>Paraoxonase 1</td>
<td>GO:0006644~phospholipid metabolic process</td>
<td>hsa00361:gamma-Hexachlorocyclohexane degradation</td>
<td>4</td>
<td>0.09</td>
<td>−6.12</td>
<td>&lt;0.001</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>PTGS1</td>
<td>Prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase)</td>
<td>GO:0006631~fatty acid metabolic process</td>
<td>hsa0090:Arachidonic acid metabolism</td>
<td>25</td>
<td>0.26</td>
<td>−6.78</td>
<td>&lt;0.001</td>
<td>0.35</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>SGCE</td>
<td>Sarcoglycan, epsilon</td>
<td>GO:0007517~muscle organ development</td>
<td>hsa04510:Focal adhesion</td>
<td>4</td>
<td>0.36</td>
<td>−6.12</td>
<td>&lt;0.001</td>
<td>0.45</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>TNC</td>
<td>Tenascin c</td>
<td>GO:0007165~signal transduction</td>
<td>hsa04510:Focal adhesion</td>
<td>25</td>
<td>5.96</td>
<td>−6.78</td>
<td>&lt;0.001</td>
<td>0.35</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>
Table 2 (Continued)

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene description</th>
<th>Gene ontology: Biological Process</th>
<th>KEGG pathway</th>
<th>Distance from marker (Mb)</th>
<th>Dh/sd1</th>
<th>P</th>
<th>Fst2</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOMM20</td>
<td>Translocase of outer mitochondrial membrane 20 homolog (yeast)</td>
<td>GO:0006886~intracellular protein transport</td>
<td>hsa000600:Sphingolipid metabolism</td>
<td>1</td>
<td>4.96</td>
<td>NS</td>
<td>NS</td>
<td>0.45</td>
</tr>
<tr>
<td>UGCC</td>
<td>UDP-glucose ceramide glucosyltransferase</td>
<td>GO:0006643~membrane lipid metabolic process</td>
<td>hsa000600 Sphingolipid metabolism</td>
<td>25</td>
<td>9.01</td>
<td>-6.78</td>
<td>&lt;0.001</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Twenty candidate genes were selected on the basis of gene ontology (GO biological process, GO cellular compartment or KEGG pathway) and their presence in one of the top ranked regions with a signature of selection in the Thoroughbred. The chromosome, distance from microsatellite marker and selection statistics are shown.

1Dh/sd = Ewens-Watterson test statistic for deviation from expected heterozygosity.

2Fst = among population differentiation.

Selection of candidate genes

A panel of 20 genes (Table 2) was selected from approximately 100 putative exercise-relevant genes among the 1202 predicted genes located within the genomic regions lying at the extremes of the distributions for departure from expected neutral variation identified in a population genetics–based genome scan (Gu et al. 2009). Candidate genes were selected based on gene ontology and were prioritized according to the statistical ranking of the genomic outlier. With the exception of COX4I1, all genes were selected from the top three ranked regions for each of the two statistics (Dh/sd and Fst). COX4I1 was included because of its relevance to exercise adaptation determined in previous gene expression experiments (Elvers et al. 2009) and its location within the region ranked fifth in the Dh/sd distribution.

Selection of SNPs from EquCab2.0 SNP database and SNP genotyping

The EquCab2.0 SNP database was interrogated for SNPs discovered in the Horse Genome Sequencing Project (Wade et al. 2009) that were located within the genomic sequence of each of the 20 candidate genes. No SNPs were identified in the regions containing the genomic sequence for the AGT and CYP51A1 genes, and these genes were therefore excluded. MUSK was excluded as it contained no Thoroughbred SNPs and was located >9 Mb from the microsatellite marker that displayed a signature of selection (Gu et al. 2009). Of the 17 genes that were included, 13 contained genomic sequence variants that had been identified in Thoroughbred. For four genes (UCGC, SGCE, TOMM20, ACSS1 and COX4I1), SNPs which were initially identified in other equine populations were included. In total, 66 SNPs were selected from the EquCab2.0 SNP database for genotyping (Table S1). SNP genotyping was carried out using iPLEX® technology (Jurinke et al. 2004) by Sequenom Inc. at their facilities in San Diego, USA.

Statistical analysis

SNP-phenotype associations were investigated for 57 SNPs in a series of case–control tests: TBE vs TBO; TBE > 8 f vs TBE > 8 f; TBE > 7 f vs TBE > 8 f; TBE > 7 f vs TBO; TBE > 8 f vs TBO; TBE > 7 f vs TBO; and TBE males vs TBO males. The sample sizes for each set were as follows: TBE, n = 86; TBO, n = 62; TBE > 8 f, n = 36; TBE > 8 f, n = 51; TBE > 7 f, n = 41; TBE males, n = 41; TBO males, n = 18.

All statistical analyses, including tests of association were performed using PLINK Version 1.05 (http://pngu.mgh.harvard.edu/purcell/plink) (Purcell et al. 2007). Quality control analyses included computation of sample allele frequencies and percentage of missing genotypes. A series of case–control association tests were performed for all loci. Statistical significance was assessed using the Cochran-Armitage test for trend and an unconditioned genotypic model. The linear regression model was used to evaluate quantitative trait association using best race distance (furlongs) and highest lifetime Racing Post Rating (RPR) as the phenotypes. Best race distance was defined as the distance (furlongs) of the highest grade Group race won by an individual. In cases where multiple races of the same grade were won, the distance of the race in which the most prize money was won was used.

Results

Genotyping assay performance

Fifty-eight SNPs had call rates of >90%. Four SNPs had call rates of <15% and were excluded (Table S2). Of the
successful genotyping assays, seven displayed a minor allele frequency (MAF) <0.02 and were excluded. An exact test for deviation from Hardy–Weinberg proportions was applied at each locus (Wigginton et al. 2005). Six SNPs in four genes [ACSS1 (3 SNPs), TNC (1 SNP), ACTN2 (1 SNP) and PTGS1 (1 SNP)] showed significant departures from Hardy–Weinberg equilibrium (HWE) (P < 0.05). However, we did not exclude SNPs that deviated significantly from Hardy–Weinberg proportions from subsequent analyses as the Thoroughbred population does not meet many of the requirements for HWE. In addition, the tests for association remain valid under departure from Hardy–Weinberg proportions, albeit with a potential loss in power if they reflect systematic genotyping errors. Therefore, 57 SNPs were included in the genetic association tests.

Case–control association tests

We performed a series of population-based case–control association tests to investigate SNP associations with retrospective racecourse performance phenotypes in Thoroughbred horses. Results for all case–control tests are available in Table S3.

Significant association with elite race winning performance

The three PDK4 SNPs (PDK4_38968139, PDK4_38969307 and PDK4_38973231) were significantly (P < 0.01) associated with elite race winning performance (TBE). The PDK4 SNPs were also significantly associated with performance in four of the other case–control association tests (TBE > 8 f vs TBO; TBE ≤ 8 f vs TBO; TBE ≤ 7 f vs TBO and TBE vs TBO males), and in each case the PDK4_38973231 SNP had the strongest association. The strongest association was with elite race performance (i.e. TBE vs TBO) (P = 0.0017; odds ratio = 2.20). A full set of results is available in Table S3.

Significant association with short distance elite race winning performance

When the elite cohort was separated into subgroups of individuals that had won their best race over short distances (TBE ≤ 8 f) and long distances (TBE > 8 f), the PON1_38697145 SNP was significantly associated with elite short distance racing (P = 0.0358, odds ratio = 3.47). Further subdivision of the short distance subgroup into individuals that won their best race over even shorter distances (TBE ≤ 7 f) revealed a significant association with the ACN9_40279726 SNP (P = 0.0448, odds ratio = 2.07). When the frequencies of alleles among the short distance elite race winning cohort (TBE ≤ 8 f and TBE ≤ 7 f) was compared with the non-winning cohort, the ADHFE1_18802749 (P = 0.0494, odds ratio = 4.20), GSN_25024464 (P = 0.0354, odds ratio = 3.18) and GSN_25028755 (P = 0.0454, odds ratio = 3.03) SNPs had significantly different allele frequency distributions.

Significant association with elite race winning performance among males

Male selection in the Thoroughbred population is particularly pronounced, with an approximately 1:45 ratio of breeding males to females (Indecon 2004). Therefore, we investigated SNP associations with performance in males only. In addition, two of the PDK4 SNPs (PDK4_38969307 and PDK4_38973231), the ACTN2_74842283 (P = 0.0437, odds ratio = 3.25), PTGS1_25991437 (P = 0.0197, odds ratio = 3.76), PTGS1_26007699 (P = 0.0051, odds ratio = 4.60) and COX4I1_32772871 (P = 0.0442, odds ratio = 2.23) SNPs were significantly associated with elite race winning performance among males.

Quantitative trait association tests

We performed a quantitative trait association test using the best race distance (furlongs) for each individual as the phenotype. The ACN9_40279726 (P = 0.0321) and PON1_38697145 (P = 0.0350) SNPs were significantly associated with best race distance. We also performed a quantitative trait association test using handicap rating (Racing Post Rating) for each individual as the phenotype (Table S4). Two of the PDK4 SNPs were significantly associated with handicap rating (PDK4_38969307, P = 0.0369; and PDK4_38973231, P = 0.0252).

Validation of associations

Five SNPs in four genes were selected for validation in an independent sample set (Table 3): ACN9_40279726; ACSS1_759076; COX4I1_32772871; PDK4_38969307; and PDK4_38973231. The two PDK4 SNPs were significantly associated with elite racing performance in the validation set (PDK4_38969307, P = 0.0255; PDK4_38973231, P = 0.0150). The PDK4_38973231 SNP consistently had the strongest association, and when all samples (n = 278) were considered, the significance of association became stronger (PDK4_38973231, P = 0.0004, odds ratio = 1.97 C.I. (95% = 1.35–2.87).

We then attempted to determine the most parsimonious genetic model for PDK4_38973231 by repeating the analysis with coding variables for additive, recessive and over-dominant models (Table 4). For PDK4_38973231, a dominant model in which the A:A and A:G genotypes were favourable provided the best explanation for the data (P = 0.0003), with the A:A and A:G genotypes more common among elite (70%) than non-elite (47%) racehorses.

Both PDK4 SNPs were also validated for association with RPR (Validation set: PDK4_38969307, P = 0.0369;
Four SNPs in three genes (ACSS1, COX4I1 and PDK4) were validated for association with elite racing performance (TBE vs TBO). $\chi^2$, P-value ($P$) and odds ratios (OR) are shown for the study sample set, the validation sample set and a cohort containing all genotyped samples.

This comparison was between elite horses that won their best race ≤7 f and elite horses that won their best race >8 f.

For the first time, we report here the association of a SNP with elite racing performance. Notwithstanding, this information may be employed as a marker for the selection of racehorses with the genetic potential for superior racing ability.

**Table 3** Case-control association test results.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene SNP</th>
<th>Location</th>
<th>Study</th>
<th>Validation</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>COX4I1</td>
<td>32 772 871</td>
<td>$\chi^2$</td>
<td>3.46</td>
<td>0.0628</td>
</tr>
<tr>
<td>4</td>
<td>PDK4</td>
<td>38 969 307</td>
<td>$\chi^2$</td>
<td>7.41</td>
<td>0.0065</td>
</tr>
<tr>
<td>4</td>
<td>PDK4</td>
<td>38 973 231</td>
<td>$\chi^2$</td>
<td>9.87</td>
<td>0.0017</td>
</tr>
<tr>
<td>22</td>
<td>ACSS1</td>
<td>759 076</td>
<td>$\chi^2$</td>
<td>3.84</td>
<td>0.0501</td>
</tr>
<tr>
<td>4</td>
<td>ACN9</td>
<td>40 279 726</td>
<td>$\chi^2$</td>
<td>6.21</td>
<td>0.0130</td>
</tr>
</tbody>
</table>

A dominant model in which the A:A and A:G genotypes were favourable provided the best explanation for the association of the PDK4_38973231 SNP.

**Table 4** Genetic model for the PDK4_38973231 A/G SNP.

<table>
<thead>
<tr>
<th>Model</th>
<th>TBE (n)</th>
<th>TBO (n)</th>
<th>$\chi^2$</th>
<th>DF</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype</td>
<td>35/92/55</td>
<td>10/35/50</td>
<td>13.74</td>
<td>2</td>
<td>0.0010</td>
</tr>
<tr>
<td>Trend</td>
<td>162/202</td>
<td>55/135</td>
<td>12.22</td>
<td>1</td>
<td>0.0005</td>
</tr>
<tr>
<td>Allelic</td>
<td>162/202</td>
<td>55/135</td>
<td>12.68</td>
<td>1</td>
<td>0.0004</td>
</tr>
<tr>
<td>Dominant</td>
<td>127/55</td>
<td>45/50</td>
<td>13.32</td>
<td>1</td>
<td>0.0003</td>
</tr>
<tr>
<td>Reccessive</td>
<td>35/147</td>
<td>10/85</td>
<td>3.476</td>
<td>1</td>
<td>0.0623</td>
</tr>
</tbody>
</table>

PDK4_38973231, $P = 0.0252$; All: PDK4_38969307, $P = 0.0017$; PDK4_38973231, $P = 0.0008$ (Table S4). At locus PDK4_38973231, A:A and A:G horses had on average a significantly higher handicap rating (A:A 98.3 ± 32.8; A:G 97.9 ± 29.0) than individuals with the G:G (81.7 ± 31.4) genotype (Table S5).

**Discussion**

In a series of case–control and quantitative association tests, we have identified SNPs associated with racing performance phenotypes located within genomic regions that have been targets of selection during the development of the Thoroughbred. It is likely that these SNPs are linked to functional variants in genes or regulatory elements that are associated with physiological adaptations that enable superior racing performance in Thoroughbreds.

For the first time, we report here the association of a SNP with elite race winning performance. When all individuals with a RPR record (n = 228) were considered, the A:A and A:G genotypes (PDK4_38973231) had on average a 16.2–16.6 lb handicap advantage over G:G horses. The expression of PDK4 is co-ordinated by the transcriptional co-activator PGC-1α (Wende et al. 2005), which has been identified as one of the critical control factors in the adaptation to exercise (Arany 2008). Specifically, PGC-1α is a key regulator of energy metabolism that regulates insulin sensitivity by controlling glucose transport, mediates exercise-induced angiogenesis (Chinsomboon et al. 2009) and co-ordinates mitochondrial biogenesis via its interaction with nuclear encoded mitochondrial protein genes (Scarpulla 2008). The regulation of glucose utilization is tightly controlled by the uptake of glucose by glucose transporters, the rate of glycolytic flux and the conversion of pyruvate to acetyl-CoA in mitochondria via the catalytic function of the pyruvate dehydrogenase complex (PDC). The critical rate-limiting step in the oxidation of glucose is the regulation of assembly of the PDC, which is controlled by pyruvate dehydrogenase kinase (PDK). PDK blocks the formation of the PDC, resulting in the beta-oxidation of fatty acids to acetyl-CoA as the substrate for oxidative phosphorylation. The oxidation of fatty acids is highly efficient in the generation of ATP and is controlled by the expression of PDK4 in skeletal muscle during and after exercise (Pilegaard & Neufer 2004).

Structural and functional genomics approaches represent powerful strategies to dissect key components of the molecular contribution to performance phenotypes and the biology of the equine athlete. We have previously identified a significant increase (+7.4-fold) in equine skeletal muscle PDK4 mRNA during recovery from exercise (Eivers et al. 2009). This observation is consistent with prolonged inhibition of the PDC to decrease glucose oxidation and increase mitochondrial fatty acid oxidation (Wende et al. 2005). Variation in gene expression is strongly influenced by structural genetic variation, and therefore the integration of functional data in the interpretation of structural variation is highly relevant. Although it is unlikely that the PDK4 SNP (PDK4_38973231) described in this study is directly influencing the control of gene expression, it is likely to be in linkage disequilibrium with a SNP that is affecting the expression of the PDK4 gene. Re-sequencing efforts and further genotyping will be required to determine the functional variant in the PDK4 gene associated with elite performance. Notwithstanding, this information may be employed as a marker for the selection of racehorses with the genetic potential for superior racing ability.
**Acknowledgements**

We thank the numerous contributors of Thoroughbred horse samples. This work was financed by a Science Foundation Ireland President of Ireland Young Researcher Award (04/Y11/B539) to EH.

**Conflicts of interest**

EH, JG and DM are named in patent applications. The remaining authors have declared no conflicts of interest.

**References**


**Supporting information**

Additional supporting information may be found in the online version of this article.

**Table S1** SNP details: Locus ID/Local SNPID, gene name, chromosome location and SNP assay.

**Table S2** Genotyping assay results: Assay ID, genotyping call rate, total number of individuals genotyped and allele frequencies.
Table S3 Case–control association test results: $\chi^2$, P-value ($P$) and odds ratios (OR) for the study sample set for a series of case–control association tests.

Table S4 Quantitative association test results.

Table S5 Quantitative trait means for PDK4_38973231: Means for Racing Post Ratings for each of the three PDK4_38973231 SNP genotypes.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer reviewed and may be re-organized for online delivery, but are not copy edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Genetics of swayback in American Saddlebred horses

D. Cook, P. C. Gallagher and E. Bailey

Department of Veterinary Science, MH Gluck Equine Research Center, University of Kentucky, Lexington, KY 40546-0099, USA

Summary

Extreme lordosis, also called swayback, lowback or softback, can occur as a congenital trait or as a degenerative trait associated with ageing. In this study, the hereditary aspect of congenital swayback was investigated using whole genome association studies of 20 affected and 20 unaffected American Saddlebred (ASB) Horses for 48 165 single-nucleotide polymorphisms (SNPs). A statistically significant association was identified on ECA20 (corrected $P = 0.017$) for SNP BIEC2-532523. Of the 20 affected horses, 17 were homozygous for this SNP when compared to seven homozygotes among the unaffected horses, suggesting a major gene with a recessive mode of inheritance. The result was confirmed by testing an additional 13 affected horses and 166 unaffected horses using 35 SNPs in this region of ECA20 (corrected $P = 0.036$). Combined results for 33 affected horses and 287 non-affected horses allowed identification of a region of homozygosity defined by four SNPs in the region. Based on the haplotype defined by these SNPs, 80% of the 33 affected horses were homozygous, 21% heterozygous and 9% did not possess the haplotype. Among the non-affected horses, 15% were homozygous, 47% heterozygous and 38% did not possess the haplotype. The differences between the two groups were highly significant ($P < 0.00001$). The region defined by this haplotype includes 53 known and predicted genes. Exons from three candidate genes, TRERF1, RUNX2 and CNPY3 were sequenced without finding distinguishing SNPs. The mutation responsible for swayback may lie in other genes or in regulatory regions outside exons. This information can be used by breeders to reduce the occurrence of swayback among their livestock. This condition may serve as a model for investigation of congenital skeletal deformities in other species.

Keywords equine, lordosis, single-nucleotide polymorphisms, whole genome scanning.

Introduction

Lordosis, specifically the dorsal concave curvature of the spine, is normal and healthy in most mammals. However, extreme lordosis is associated with pathology in horses (Rooney & Pickett 1967; Rooney 1969). This condition is also known in horses as swayback, lowback or softback. Rooney & Robertson (1996) noted that ‘variable degrees of lordosis seem to be common in certain lines of American Saddlebred horses’. Figure 1 shows the characteristic conformation for a horse with lordosis (Fig. 1a) and a normal horse (Fig. 1b). A familial aspect was suspected but not established in previous work. A study of swayback among Saddlebred horses led Gallagher et al. (2003) to devise a method to measure the extent of lordosis and to characterize the variation found among horses in this breed. Based on this study, a threshold for considering horses to be swayback was defined, and 5% of 294 horses were considered affected. Studies of families suggested, but did not prove, a recessive mode of inheritance.

In our experience, breeders are of mixed opinions regarding this trait. Swayback horses have not been routinely identified as experiencing pain, and some horses with this trait have performed well. Most breeders regard the condition as a conformation defect and avoid breeding stock with this condition. Identifying the genetic determinant(s) for this condition would provide breeders with the opportunity to better understand the genetics of the trait and use that information in their selection programs.
Traditionally, family studies have been most useful to map hereditary traits in horses (for example, Trommershausen-Smith 1978; Bailey et al. 1997; Locke et al. 2001). However, while swayback horses are not uncommon, breeders avoid matings they believe likely to produce affected horses. Consequently, family studies are difficult. However, with the advent of the horse genome sequence and the availability of dense arrays of single-nucleotide polymorphisms (SNPs) for whole genome association (WGA) studies, population studies can be used to investigate the genetics of traits in horses (Wade et al. 2009). The purpose of this study was to determine whether a hereditary component contributes to the swayback trait in American Saddlebred (ASB) horses and, if so, to identify the location of genes for this trait.

Materials and methods

Phenotypic assessment of lordosis

Phenotypic assessment of lordosis was based on the measurement of back contour (MBC) (Gallagher et al. 2003). Two points were chosen on the horse’s back, one at the top of the withers (the highest point of the dorsal spinous process in the region of thoracic vertebrae T2-T3) and one on the point of the rump (the highest point on top of the horse’s hips). The shortest distance between those points was measured in centimetres and designated ‘A’. Next, the distance along the contour of the back was measured in centimetres and designated ‘B’. MBC was designated as the difference between ‘A’ and ‘B’. Figure 1c illustrates the points of measurement on a swayback horse.

Horses measured for MBC

Measurements and tissue (hair or blood) samples for ASB horses used in this study came from a sample set of 749 ASB horses collected from private and commercial farms in and around Kentucky. The average age was 7.1 years and ranged from 1 month to 29 years old. ASB horses used in the Illumina and Sequenom assays were selected from among this group of samples.

Horses for illumina assay

For the genome scan using the Illumina Equine SNP50 Chip, 40 ASB horses were selected, based on their lordosis phenotype. Twenty horses were selected based on having an MBC >8.0 cm, and 20 were selected based on having MBC <5.0 cm. The control group included 19 half-siblings for horses in the affected group to reduce the chance of population substructure producing spurious associations.

Horses for sequenom assay

A total of 426 horses were selected for screening with 35 SNPs located in the genomic region suggested by the WGA study. Three horses were tested in duplicate to control for the quality of the testing. Twilight, the Thoroughbred horse used for the whole genome sequencing, was also tested for quality control of known SNPs located in the genomic region suggested by the WGA study. Among the ASB horses, 33 (including the 20 original cases) had values for MBC of 7.0 or greater, 287 had MBC <7.0 cm.

Figure 1 Image of a swayback 3-year-old horse (a), image of a 3-year-old horse with a normal back (b) and an image showing the method of determining the measurement of back curvature (MBC). MBC was determined as the difference between the shortest distance (blue) and the distance along the contour of the back (red) from points A and B.
and 106 were parents siblings or offspring of swayback horses. The relatives were tested to assist with haplotype determination.

DNA extraction

DNA was extracted from blood or hair follicles for testing. DNA from blood samples was extracted using Puregene whole blood extraction kit (Gentra Systems Inc.) according to its published protocol. Hair samples were processed using 20–30 hair bulbs according to the modified protocol using a Gentra DNA purification kit. The hair bulbs were placed in 200 μl Gentra Cell Lysis solution containing 0.01 mg proteinase K (Sigma-Aldrich) and incubated at 5 °C overnight. DNA was then purified by following remaining steps of the protocol.

Illumina equine SNP50 genotyping

Initial SNP genotyping of 40 samples in the case/control group was performed utilizing the Illumina Equine 50 SNP chip for a WGA study. DNA was provided to the core facility at the Mayo Clinic in Rochester, MN for genotyping.

Sequenom SNP genotyping

Ten SNPs from the WGA study and 39 additional SNPs from the EquCab2.0 SNP database (http://www.broadinstitute.org/ftp/distribution/horse_snp_release/v2/) were selected for testing using the MassArray iPLEX Gold assay on the Sequenom platform. Genotyping was performed at Proactive Genomics, LLC. The 49 SNPs selected were identified as being between positions 41 530 793 and 44 585 118 and are listed in Table 2. Following testing, 43 of 49 the SNPs provided quality genotyping data. Of these, 35 SNPs with minor allele frequency >0.01 were used for final association and haplotype analysis. Furthermore, 96% of the submitted samples had fewer than 0.05 genotypes missing and were thus included in analysis.

Analyses

Genotyping data analysis was performed with PLINK v1.06 (Purcell et al. 2007). Association analysis by case/control chi-square was performed on the Illumina data. To minimize error because of the multiplicity of SNPs tested, an MPERM analysis with 10 000 permutations was performed and a statistic, EMP2 (referred to here as corrected P), verified possible associations. Once association was identified, different sized haplotypes were tested by chi-square analysis to identify the size and location of the highest associated haplotype in the region. Selected haplotypes were phased for all horses in this study to clarify familial patterns of transmission. Chi-square association analysis was also performed on the Sequenom data to verify association of the SNPs included from Illumina assay and to identify any new associations. The 35 SNP haplotypes were then phased to verify familial patterns and to identify possible recombination locations. Haplotype analysis was also performed in Haploview (Barrett et al. 2005) for validation of PLINK analysis. Haplotypes were identified using the HAP and PHASE options of PLINK. Haplotype frequencies were determined by direct counting.

Sequencing

To investigate candidate genes, exon sequences of two affected horses were compared to those of two unaffected horses. The two affected horses were selected as unrelated horses with MBC >7.0 and homozygous for the swayback-associated ECA20 haplotype. Unaffected horses were selected as unrelated horses with MBC <7.0 and not possessing the ECA20 haplotype associated with swayback. When SNPs were identified and confirmed on these four horses, additional control and case horses were tested to determine whether there was an association with the swayback trait. PCR template for sequencing was amplified in 20 μl PCRs using 1× PCR buffer with 2.0 mM MgCl₂.

Table 1 Results from Illumina Equine SNP50 assay of DNA from 20 swayback and 20 normal back Saddlebred horses. Chromosome location (CHR), base position on chromosome (BP), SNP identifier (SNP-ID), 2 × 2 chi-square value (CHISQ), P-value (P) and P-value from Monte Carlo correction for number of comparison (corrected P) are shown.

<table>
<thead>
<tr>
<th>CHR</th>
<th>BP</th>
<th>SNP-ID</th>
<th>CHISQ</th>
<th>P</th>
<th>corrected P</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>41530793</td>
<td>BIEC2-532523</td>
<td>20.28</td>
<td>6.69E-06</td>
<td>0.01699</td>
</tr>
<tr>
<td>X</td>
<td>51065036</td>
<td>BIEC2-1124071</td>
<td>15.66</td>
<td>7.60E-05</td>
<td>0.2554</td>
</tr>
<tr>
<td>20</td>
<td>42062440</td>
<td>BIEC2-532578</td>
<td>15.53</td>
<td>8.12E-05</td>
<td>0.2904</td>
</tr>
<tr>
<td>7</td>
<td>7444796</td>
<td>BIEC2-978005</td>
<td>13.73</td>
<td>0.000211</td>
<td>0.6762</td>
</tr>
<tr>
<td>X</td>
<td>51061227</td>
<td>BIEC2-1124068</td>
<td>13.47</td>
<td>0.000243</td>
<td>0.7091</td>
</tr>
<tr>
<td>7</td>
<td>18839866</td>
<td>BIEC2-9840003</td>
<td>13.33</td>
<td>0.000261</td>
<td>0.7326</td>
</tr>
<tr>
<td>28</td>
<td>35119684</td>
<td>BIEC2-739844</td>
<td>13.33</td>
<td>0.000261</td>
<td>0.7326</td>
</tr>
<tr>
<td>20</td>
<td>41576546</td>
<td>BIEC2-532530</td>
<td>13.07</td>
<td>0.000301</td>
<td>0.7931</td>
</tr>
<tr>
<td>20</td>
<td>41604741</td>
<td>BIEC2-532534</td>
<td>13.07</td>
<td>0.000301</td>
<td>0.7931</td>
</tr>
<tr>
<td>20</td>
<td>41625809</td>
<td>BIEC2-532535</td>
<td>13.07</td>
<td>0.000301</td>
<td>0.7931</td>
</tr>
</tbody>
</table>
200 μM of each dNTP, 1 μL genomic DNA from hair lysate, 0.2 U FastStart Taq DNA polymerase (Perkin Elmer, Wal- tham, Mass.) and 50 nM of each primer. Template product was quantified on a 1% agarose gel, then amplified with BigDye Terminator v1.1 cycle sequencing kit according to manufacturer’s instructions (Applied Biosystems), cleaned using Centri-Sep columns (Princeton Separations Inc.), and run on an ABI 310 genetic analyzer (Applied Biosystems). Primers were designed in Primer 3 (Steve and Skaletsky 1998) using 8 intronic sequences and seven exonic sequences of TRERF1 (transcriptional regulating factor for CYP11A1), 15 exonic sequences for RUNX2 (transcription
d factor associate with osteoblast differentiation) and six exonic sequences for CNPY3 (regulates cell surface expression of Toll Receptor 4) (See Table S1).

**Results**

**Distribution of MBC**

The distribution of MBC values among the 749 horses in this study was similar to that found by Gallagher et al. (2003). MBC for the 749 ASB horses ranged from 0 to 17 cm. Most MBC values appeared to fall within a normal
population distribution with a mean of $3.6 \pm 1.9$ cm, median of 4 cm and mode of 4 cm. An MBC of 7 cm or greater was selected to classify horses as affected based on this value being approximately two standard deviations from the mean.

Whole genome scan with illumina equine SNP50 beadchip

The Illumina Equine SNP50 chip (Illumina) was effective in typing DNA from the initial 40 horses. The 40 individuals had an average call rate of 0.96. After filtering for minimum minor allele frequency and genotyping, 48 165 SNPs were retained for data analysis.

Table 1 shows the top ten statistically significant results from a $2 \times 2$ chi-square analysis comparing the distribution of SNPs in the DNA of affected and non-affected horses. The most significant association was found on ECA20 for SNP BIEC2-532523 ($P = 6.69 \times 10^{-6}$). Of the ten highest P-values, five occurred for SNPs on ECA20 in the 531-kb region between positions 42 062 440 and 41 530 973. The multiplicity of comparisons can result in the spurious discovery of high chi-square values; therefore, to control for multiple comparisons, PLINK was used to conduct a Monte Carlo simulation with 10 000 permutations to calculate a corrected P-value (EMP2). Only the association with SNP BIEC2-532523 on ECA20 remained significant (corrected $P = 0.017$).

Among the affected horses with MBC 7 cm or greater, 17 of 20 were homozygous for the T allele of the BIEC2-532523 T/C SNP. Among the 20 controls, only seven were homozygous for this allele. This SNP fell within a larger region of homozygosity spanning approximately 3 Mb (base position 41 604 741–44 512 270), which was identified with the homozygosity function in PLINK by scanning sliding windows of 35 SNPs, moving one SNP at a time and allowing one heterozygote per window to be considered homozygous (data not shown).

Sequenom

To verify the statistical associations found with the Illumina assay, SNPs from the candidate region on ECA20 were tested, including ten SNPs from the Illumina Equine SNP50 chip and 25 additional SNPs from the EquCab2.0 SNP database. Association chi-squared analyses were performed separately for the additional 13 affected and 181 unaffected horses not previously included in the WGA assay. The association analyses for just these new samples are shown in Table 2 (CHISQ1 and P1). Based on these 13 affected horses, the association with BIEC2-532523 remained statistically significant with a $P$-value of 0.036. Because this was a comparison with new samples dictated by the original Illumina assay, no statistical correction is necessary to correct for multiplicity of testing, as performed for the previous experiment. Of the 35 SNPs, seven showed statistically significant associations ($P < 0.05$). When all 33 affected horses and 287 controls were compared, 21 of the 35 selected SNPs showed statistical significance in their distributions between the two groups based on their relative frequencies.

Haplotype analysis

Haplotypes from the region defined in the Sequenom assay (41 530 793 and 44 585 118) were compared among affected and unaffected horses. A minimum haplotype that included the maximum number of affected horses and the lowest number of non-affected horses was identified using the four SNPs BIEC2-532523, 532534, 532578 and 532658 and spanned 1 073 074 bp. Intervening SNPs did not affect haplotype assignment. These four SNPs allowed identification of 13 haplotypes. The haplotypes and their frequencies among the affected and non-affected horses are shown in Table 3.

Table 3 Haplotypes defined by BIEC-532523, 532534, 532578 and 532658, and haplotype frequencies among swayback and non-affected ASB horses.

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Affected ($N = 33$)</th>
<th>Non-affected ($N = 287$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGTG</td>
<td>0.80</td>
<td>0.388</td>
</tr>
<tr>
<td>CACT</td>
<td>0.03</td>
<td>0.232</td>
</tr>
<tr>
<td>CACG</td>
<td>0.06</td>
<td>0.167</td>
</tr>
<tr>
<td>TACT</td>
<td>0.04</td>
<td>0.164</td>
</tr>
<tr>
<td>TGCG</td>
<td>0.05</td>
<td>0.019</td>
</tr>
<tr>
<td>TGTT</td>
<td>0.00</td>
<td>0.012</td>
</tr>
<tr>
<td>CATG</td>
<td>0.02</td>
<td>0.003</td>
</tr>
<tr>
<td>TGCT</td>
<td>0.00</td>
<td>0.003</td>
</tr>
<tr>
<td>CGCT</td>
<td>0.00</td>
<td>0.003</td>
</tr>
<tr>
<td>CATT</td>
<td>0.00</td>
<td>0.002</td>
</tr>
<tr>
<td>TACG</td>
<td>0.00</td>
<td>0.002</td>
</tr>
<tr>
<td>TATG</td>
<td>0.00</td>
<td>0.002</td>
</tr>
<tr>
<td>CTGT</td>
<td>0.00</td>
<td>0.002</td>
</tr>
</tbody>
</table>

ASB, American Saddlebred.
comparison of the combined data set to the subset of horses with swayback was highly significant \( (P < 0.000001) \).

### Candidate genes

The SNPs defined an overall region of homozygosity for swayback horses ranging from 41.5 to 44.5 Mb on ECA20. The annotated horse genome at ENSEMBL genome browser (Hubbard et al. 2009) showed that this region contains 53 known and predicted genes. Three genes were selected as possible candidates based on predicted or known function in other species; sequence comparisons for TRERF1 (15 exons), RUNX2 (seven exons, eight introns) and CNPY3 (six exons) between normal and swayback horses did not identify SNPs associated with swayback. Data on SNPs found and their occurrence among the case and control horses are shown in Table S2.

### Discussion

The distribution of MBC measurements in this study confirmed the results from the earlier study by describing a normal distribution of the MBC phenotype, with 5% falling two standard deviations above the mean (Gallagher et al. 2003). In that previous study, the mean was 4.05 cM, while the mean found in this study was smaller, 3.6 cM. Differences in age of horses may account for this difference, because the previous study showed a positive correlation for age and MBC; the mean age in the first study was 7.8 years, and the mean age in this study was 7.1 years.

The WGA study demonstrated the presence of a recessive gene responsible for the swayback trait in horses. This was suggested in the initial WGA study with 40 horses and the Illumina SNP chip (corrected \( P = 0.017 \)). The association was confirmed in a subsequent study with a second set of affected horses using SNPs from this targeted area \( (P = 0.036) \). Using data from all horses, a haplotype showing the strongest association with swayback was identified based on four SNPs (TGTG) described in Table 3. This haplotype spanned 1 073 074 bases and the region harboured 53 known and predicted genes.

The TGTG haplotype was the most common and suggests that the haplotype occurred in the breed before the mutation causing swayback. If a gene present in this haplotype was mutated to cause swayback, then only knowing the specific mutation would allow us to distinguish between these haplotypes. Of course, mutations which subsequently occurred within the swayback-causing haplotype might allow us to use tests for other markers to identify haplotypes completely associated with swayback, although not all horses with this gene for swayback.

The high frequency of this haplotype and this phenotype among ASB horses might be the consequence of selection by breeders. If a single copy of the gene produced a desirable phenotypic effect, such as improved gait, selection for that trait may negate selection against swayback and result in a net increase in the frequency of the gene in the breed. Comparisons of gene frequencies for the TGTG haplotype among horses with different performance phenotypes are needed to answer this question.

While the high homozygosity of this haplotype among the swayback horses demonstrated the presence of a recessive gene for the trait, not all swayback horses were homozygous for the region. We found five different haplotypes among the swayback horses, and 30% were not homozygous for the haplotype associated with the recessive swayback condition. The swayback phenotype may have multiple possible causes, of which the hereditary recessive condition is only one. While 70% of the swayback horses were homozygous for this haplotype, we observed seven (21%) swayback horses that were heterozygous and three (9%) that did not have the haplotype at all (Table 4). There may be multiple causes of swayback among Saddlebred horses, and the recessive gene suggested by this study may be only one of them. Other genes, accidents affecting skeletal integrity or even management practices may cause swayback in the absence of the recessive gene implicated by this study. Nevertheless, considering the high prevalence of this haplotype among affected horses, this hereditary recessive condition is probably the most common cause of swayback among Saddlebred horses.

As noted above, the region contains 53 known or predicted genes for the horse. We selected three genes for DNA sequencing in the hope of identifying the causative mutation. One gene in this region, RUNX2, had been implicated in skeletal defects based on information from OMIM. RUNX2 has been found to be a scaffold for factors involved in skeletal gene expression (Stein et al. 2004), and it plays a role in osteoblast and chondrocyte differentiation and migration (Fujita et al. 2004). From assessment of other likely gene functions, the candidate gene TRERF1 was identified based on its function as a transcription factor. CNPY3, a trinucleotide repeat-containing gene, was also considered, because repeat expansion has been shown to

---

**Table 4 Haplotype by swayback/non-affected status**.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Number of horses</th>
<th>TGTG homozygotes</th>
<th>TGTG heterozygotes</th>
<th>No TGTG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swayback</td>
<td>33</td>
<td>23 (70%)</td>
<td>7 (21%)</td>
<td>3 (9%)</td>
</tr>
<tr>
<td>Non-affected</td>
<td>287</td>
<td>44 (15%)</td>
<td>135 (47%)</td>
<td>108 (38%)</td>
</tr>
<tr>
<td>Combined</td>
<td>320</td>
<td>67 (21%)</td>
<td>142 (44%)</td>
<td>111 (35%)</td>
</tr>
</tbody>
</table>

*Chi-square for swayback vs. combined = 47.08, \( P < 0.00001 \).
play a role in various diseases. Such a repeat expansion within SCA1 on human 6p is responsible for spinocerebellar ataxia type 1 (Kameya et al. 1994). However, exon sequencing of the three candidate genes did not identify SNPs or other genetic markers associated with the trait.

Rooney & Robertson (1996) distinguished between senile lordosis and congenital lordosis in horses. Senile lordosis was a consequence of ageing. Congenital lordosis occurred as a consequence of hypoplasia of the articular facets of thoracic vertebrae and followed birth as a result of weight bearing (Rooney & Pickett 1967). We believe that the condition we have been investigating in ASB horses is the congenital form, because most of the affected horses were under the age of 10. However, this should be confirmed by sequential measurement of MBC in horses of different ages, concentrating especially on young horses, to determine the progression of lordosis.

Discovery of the mutation responsible for swayback is a goal that remains ahead of us. In connection with this project, exons of several candidate genes were sequenced. However, the cause of the trait could be because of aspects of gene expression which are not encoded in exons. As we learn more about the genome of animals, we realize that even the introns and the DNA between genes can play a role in gene regulation. The region of interest might be reduced by further studies using additional genetic markers from this region, including more SNPs, microsatellites or other genetic polymorphisms. Another approach to understanding this condition may be to investigate differences in gene expression between affected and unaffected horses. Discovery of a gene which shows differential expression would help to focus this work. However, the choice of tissues and the age at which horses are tested may be factors which confound such an approach.

More insight into the variation responsible for early-onset extreme lordosis in horses may be beneficial for studies of human juvenile kyphosis and juvenile idiopathic scoliosis (IS). As with the horse, these two congenital conditions exhibit an early age of onset. Familial IS only accounts for 10% of all cases in humans, while 90% appear to be sporadic with unknown or environmental aetiological factors (Cheng et al. 2007). Through familial linkage analysis, candidate regions for IS susceptibility have been identified on human chromosomes 6p, 10q and 18q (Wise et al. 2000). In a more recent study, regions on 6, 9, 16 and 17 were identified through genome-wide screening (Miller et al. 2005). It is of particular interest to note that the segment on HSA6 implicated in that study is syntenic with the region on ECA20 that was found in this study to be associated with swayback.

**Acknowledgements**

The authors are grateful to the Morris Animal Foundation and to the American Saddlebred Horse Association for funding the project. Samples were generously provided by private horse breeders. This work was in connection with the doctoral research of DC. The work is also in connection with a project of the University of Kentucky Agricultural Experiment Station as paper number 10-00-000.

**Conflicts of interest**

The authors have declared no potential conflicts.

**References**


Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Primer pairs used to sequence intronic and exonic regions of candidate genes for swayback condition in American Saddlebred horses.

Table S2 Results from sequencing candidate genes RUNX2, TRERF1 and CNPY2 in swayback and normal horses to investigate association of this gene with swayback.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Comparative human–horse sequence analysis of the CYP3A subfamily gene cluster

A. Schmitz*, S. Demmel*, L. M. Peters*, T. Leeb†, M. Mevissen* and B. Haase†
*Division Veterinary Pharmacology and Toxicology, Vetsuisse Faculty, University of Bern, 3012 Bern, Switzerland. †Institute of Genetics, Vetsuisse Faculty, University of Bern, 3012 Bern, Switzerland

Summary
Cytochrome P450 enzymes (CYP450s) represent a superfamily of haem–thiolate proteins. CYP450s are most abundant in the liver, a major site of drug metabolism, and play key roles in the metabolism of a variety of substrates, including drugs and environmental contaminants. Interaction of two or more different drugs with the same enzyme can account for adverse effects and failure of therapy. Human CYP3A4 metabolizes about 50% of all known drugs, but little is known about the orthologous CYP450s in horses. We report here the genomic organization of the equine CYP3A gene cluster as well as a comparative analysis with the human CYP3A gene cluster. The equine CYP450 genes of the 3A family are located on ECA 13 between 6.97–7.53 Mb, in a region syntenic to HSA 7 99.05–99.35 Mb. Seven potential, closely linked equine CYP3A genes were found, in contrast to only four genes in the human genome. RNA was isolated from an equine liver sample, and the approximately 1.5-kb coding sequence of six CYP3A genes could be amplified by RT-PCR. Sequencing of the RT-PCR products revealed numerous hitherto unknown single nucleotide polymorphisms (SNPs) in these six CYP3A genes, and one 6-bp deletion compared to the reference sequence (EquCab2.0). The presence of the variants was confirmed in a sample of genomic DNA from the same horse. In conclusion, orthologous genes for the CYP3A family exist in horses, but their number differs from those of the human CYP3A gene family. CYP450 genes of the same family show high homology within and between mammalian species, but can be highly polymorphic.

Keywords cytochrome P450 3A subfamily, genomic organization, horse, human, polymorphism.

Introduction
Cytochrome P450 enzymes (CYP450) constitute a large superfamily of membrane-bound haem–thiolate containing monoxygenases, so called because they catalyse the incorporation of one oxygen atom from molecular oxygen into a substrate. CYP450 can be found in virtually all living organisms, from bacteria, in which cytochromes are soluble enzymes, to mammalian species, where cytochromes are bound to the smooth endoplasmic reticulum (Fink-Gremmels 2008). Members of the CYP450 family diverged from each other as early as 2 billion years ago, resulting in <40% similarity in amino acid sequences within the same CYP450 family. The CYP450 subfamilies emerged more than 400 million years ago, and within a subfamily the amino acid similarities are usually >55% (Nelson et al. 1996). To date, in humans, 57 functional genes and 58 pseudogenes have been sequenced (http://drnelson.uthsc.edu/cytochromeP450.html). CYP450s have diverse biological functions that include biosynthesis and regulation of cellular effectors such as synthesis of steroids and other endogenous compounds. Within the complex group of CYP450s, the enzyme families 1–3 are primarily involved in the biotransformation of drugs and toxins and are therefore referred to as drug-metabolizing enzymes. The pharmaceutical industry has developed an increasing interest in the structure and function of CYP450s, as they may also represent an important locus of drug–drug interactions.
Determination of drug-metabolizing CYP450s has become a routine step in the drug development process and is required by the U.S. Food and Drug Administration (FDA) (Guengerich 2003).

The human CYP3A subfamily includes CYP3A4, CYP3A5, CYP3A7 and CYP3A43 encoded by a gene cluster on human chromosome 7. This is one of the most versatile biotransformation systems and facilitates the metabolism of about half of the known drugs. CYP3A4 is abundant in both the intestinal epithelium and the liver, and probably represents the most important of all drug-metabolizing enzymes. It accounts for nearly 50% of the CYP450 enzymes in humans (Wilkinson 2005). A high inter-individual variability in expression levels of CYP3A genes has been found, as well as an abundance of polymorphisms within the subfamily. This variability is further enhanced by induction and inhibition of the CYP3A enzymes by certain drugs and dietary constituents (Qiu et al. 2008).

CYP3A paralogs have been identified among others in rat, mouse, rabbit, dog, minipig, cow and monkey. In rats there are seven genes, in mouse eight genes, and there are only three isoforms in dogs and a single isoform in the rabbit (http://drnelson.uthsc.edu/cytochromeP450.html).

A direct extrapolation between experimental animal species and humans is not always possible, for reasons such as differences in enzyme activity, abundance, specificity and regulation (Nelson 1999; Spatzenegger et al. 2007). For example, CYP2C enzymes share important functional similarities in laboratory animal species, such as testosterone-6β hydroxylation, which are not present in CYP2C of humans. It was demonstrated that small sequence changes can produce large effects on activity and function: e.g. the canine CYP3A26 displays a much lower testosterone-6β hydroxylase activity than the canine CYP3A12, although they share 96% amino acid identity (Fraser et al. 1997).

Horses have been subject to more and more refined drug treatments during recent years, with treatment regimes being based on extrapolation from human medicine or clinical experience. Identification of CYP450s responsible for the metabolism of a drug is important with respect to clinical drug–drug interactions, genetic polymorphism and toxicity. Knowledge of CYP450s involved in drug metabolism and of individual variations in the metabolism of certain drugs is still missing for the horse. The first horse CYP450s were recombinantly expressed in insect cells only very recently (Maio Knych & Stanley 2008; Maio Knych et al. 2009). As the horse genome sequence was recently completed, much faster progress in CYP450 research is now possible.

The objective of this study was to obtain the genomic structure and DNA sequence of the equine CYP3A subfamily, including existing polymorphisms, to facilitate further functional investigations of the equine CYP3A subfamily.

Materials and methods

DNA sequences

The human sequence investigated covered the region 99.2–99.5 Mb of the build 37.1 HSA 7 sequence. The corresponding equine sequence contained the region 6.97–7.53 Mb of the build 2.0 on ECA 13.

Sequence analysis


Dot plot and percent identity plot (pip) analyses were performed with the programs PipMaker and MultiPipMaker (Schwartz et al. 2000). For detailed comparisons, local and global pairwise alignments were calculated with the program LASTZ (Huang & Miller 1991).

DNA and RNA extraction

Genomic DNA was isolated from liver tissue of a ten-year-old female Arabian horse using the Nucleon BACC 2 genomic DNA extraction kit (GE Healthcare). Additionally, total RNA was isolated from a liver sample of the same horse using TRizol™ reagent (Invitrogen, Karlsruhe, Germany).

RT-PCR and determination of the CYP3A cDNA sequences

Aliquots of 1 µg total RNA were reverse transcribed to cDNA using 20 pmol (T)24V primer and SuperScript™ III reverse transcriptase (Invitrogen). On the basis of a multiple alignment, primers were designed to distinguish between the different paralogs of the equine CYP3A subfamily (Table S1). One microlitre of the cDNA was used as template in a polymerase chain reaction. The reaction was performed in a total volume of 20 µl using TopTaq DNA polymerase (Qiagen), and products were inspected for yield and purity on agarose gels. Direct sequencing of the PCR products was performed after shrimp alkaline phosphatase (Roche) and exonuclease I treatment (N.E.B.). PCR products were sequenced on an ABI 3730 capillary sequencer (Applied Biosystems) using the PCR primers and additional internal sequencing primers. The sequences of the primers are listed in Tables S2 and S3 of the supporting information.
Sequencing data were assembled with Sequencher 4.9 (GeneCodes).

PCR and mutation analysis


PCR and direct sequencing were performed as described previously. PCR primers were used as sequencing primers, with two exceptions, in which nested primers were designed for sequencing. Sequences are listed in Tables S4 and S5.

For the prediction of the effect of an amino acid exchange on protein function, we used the programs POLYPHEN (http://genetics.bwh.harvard.edu/pph/) and pMUT (http://mmb2.pcb.ub.es:8080/PMut/+).

Results

We used the repeat masked human 27205 bp CYP3A4 sequence as the query in a BLASTN analysis to identify homologous sequences in the horse genome. On the basis of these initial BLASTN results, we selected a 560- kb interval on ECA 13 that included the entire equine CYP3A gene cluster.

Pairwise dot plot analyses of the human CYP3A cluster on HSA 7 and the equine 560- kb region revealed that the horse CYP3A gene cluster contains seven potential genes and one potential pseudogene, in contrast to the human region, which includes only four genes and two pseudogenes (Fig. 1). The number of exons (13) and the lengths of the internal exons were consistent between human CYP3A4 and all seven equine genes.

A search in the CYP450 nomenclature database indicated that names had already been assigned to the equine genes (http://drnelson.uthsc.edu/Nomenclature.html), although not in a continuous order (see Fig. 1). Out of all seven paralogs, the CYP3A89 mRNA and protein displayed the highest similarity with the human CYP3A4. All genes except CYP3A129 were situated on the plus strand. The pseudogene CYP3A128P, situated between CYP3A97 and CYP3A129, aligns with exons 8–13 of the human CYP3A4 gene, but exons 8 and 9 lack one base pair compared to exons 8 and 9 of the human CYP3A4, and exon 12 showed a poor alignment below 60%. We did not undertake to sequence this pseudogene on genomic DNA.

MultiPip pairwise sequence comparison of the CYP3A genes from human and horse (Fig. 2) showed that the exons were highly conserved between the two species. Many non-coding regions also exhibited a high degree of sequence conservation. The GC content in all CYP3A genes of the horse was ≤40%, and LINEs constituted the main repeat type found in the CYP3A genes.

The 1512- bp coding sequence of six of seven potential genes could be amplified and completely sequenced from a cDNA template. Sequences were submitted to the EMBL database and assigned the following accession numbers: FN669292, FN669293, FN669294, FN669295, FN669296 and FN669297.

Sequencing revealed a total of 19 single nucleotide polymorphisms (SNPs) and one 6-bp deletion in the coding sequences of all six genes with respect to the reference sequence (Table 1). The reference sequence used in this

Figure 1 Dot plot analysis of the equine CYP3A gene cluster with the corresponding sequence of the human CYP3A gene cluster. Sequence sections were chosen that contain the entire CYP3A gene cluster without flanking genes. Except for the equine CYP3A89, no other CYP450 has been annotated in build 2.0 of the equine genome sequence at the NCBI Map Viewer, but hypothetical genes (LOCs) have been annotated based on similarities with other CYP450 genes. Names for the CYP450s were extracted from the cytochrome P450 homepage (http://drnelson.uthsc.edu/cytochromeP450.html), where suggested protein sequences have been published and named.

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 72–79
work was constructed by joining the exons extracted from the genomic DNA sequence (EquCab 2.0).

Seven of the SNPs that we detected had already been entered in the SNP database from the Broad Institute at MIT and Harvard (http://www.broadinstitute.org/ftp/distribution/horse_snp_release/v2). We did not confirm the established SNPs on the genomic DNA of this horse as we did for the formerly unknown SNPs (Table 1).

CYP3A93 exhibited the highest variability, with 7 SNPs in the coding sequence, four of which were located in exon 1.

The majority of SNPs were found to be either synonymous or to code for conserved amino acid exchanges which are not expected to influence the protein function. This was calculated by using the programs Polyphen and pMUT. The only non-conserved amino acid exchange we found was the exchange of tryptophan with serine in exon 3 of CYP3A93. The replacement of the large side chain of tryptophan 58 with the much smaller side chain of a serine (p.Trp58Ser) was predicted to be ‘possibly damaging’ (Polyphen) or ‘pathological’ (pMUT) to the protein. Amplification of CYP3A95 rendered, apart from the expected product of 1512 bp, a second product of about 900 bp.

Sequencing of the gel-extracted 900-bp product provided three sequences that covered 695 bp in total. The three sequences between 200 and 300 bp comprised bases of exons 1, 2, exons 4, 5, 6 and exons 11, 12. Thus, our data suggest the existence of an alternatively spliced CYP3A95 mRNA isoform. We did not pursue the effort of obtaining a full sequence of the hypothetic splice variant.

In CYP3A96, a 6-bp deletion (p.Thr499_Val500del) was detected in exon 13, 8 bp before the stop codon, rendering a protein with only 501 amino acids instead of 503. All SNPs and the 6-bp deletion were confirmed on genomic DNA of the same horse. It is interesting to note that even if the SNPs investigated on the genomic level were rated to be heterozygous, in most cases the mutated variant was clearly predominating on the cDNA level, indicating potential allelic imbalances in the mRNA expression. We were unable to amplify the mRNA of CYP3A129 from a liver cDNA template, which may be because of low or absent expression of this gene in liver.

Alignment of the coding sequences and deduced protein sequences with the homologous sequences of the human CYP3A subfamily (Table 2) revealed high identity between...
human and horse CYP3A subfamilies (79–86% for the coding sequence and 68–81% for the amino acid sequence), but also very high identity between the members of the equine CYP3A subfamily (88–91.7% for the coding sequence and 83.9–88.5% on the protein level).

**Discussion**

This study describes the detailed genomic organization of the equine CYP3A gene cluster for the first time. Comparison and genomic characterization of the human CYP3A gene cluster with the complete horse CYP3A gene cluster became possible through the recent completion of the horse genome sequence.

Comparative analysis of the human chromosome region containing the CYP3A gene cluster and the corresponding equine region illustrated that the horse genome contains six highly similar CYP3A genes and two potential pseudogenes, in contrast to the human genome, which has only four known functional CYP3A genes and two pseudogenes.

Our findings of six functional CYP3A genes in the horse genome versus only four in the human genome are in accordance with reports about CYP3A genes in other mammalian species such as dog, cow, mouse and rat. In an extensive research project, CYP3A loci from 16 different species were compared, and conclusions were drawn for the evolution of the CYP3A genomic loci over a period of 450 million years (Qiu et al. 2008). Among the species investigated in this paper, the horse belongs to the ones with the highest amount of CYP3A genes, a level met only by mouse and opossum. Most CYP3 genes build gene clusters within a genomic CYP3 locus that develops through independent gene duplications in distantly related species. A dominant role of tandem duplications in the development of the CYP3 clusters has been suggested (Thomas 2007; Qiu et al. 2008). Frequent gene duplications and losses have also been attributed to other so-called unstable CYP450 gene families, which all share a tendency for positive selection in

<table>
<thead>
<tr>
<th>CYP450 Gene</th>
<th>Polymorphism (cDNA)</th>
<th>Polymorphism (genomic DNA)</th>
<th>Position within Gene</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP3A93</td>
<td>c.17G&gt;C</td>
<td>g.6996581G&gt;C</td>
<td>Exon 1</td>
<td>p.Ser6Thr</td>
</tr>
<tr>
<td></td>
<td>c.26C&gt;T</td>
<td>g.6996590C&gt;T</td>
<td>Exon 1</td>
<td>p.Thr9Met</td>
</tr>
<tr>
<td></td>
<td>c.64A&gt;G</td>
<td>g.6996628A&gt;G</td>
<td>Exon 1</td>
<td>p.Ile122Val</td>
</tr>
<tr>
<td></td>
<td>c.66T&gt;C</td>
<td>g.6996630T&gt;C</td>
<td>Exon 1</td>
<td>Silent</td>
</tr>
<tr>
<td></td>
<td>c.87G&gt;A</td>
<td>g.7000744G&gt;A</td>
<td>Exon 2</td>
<td>Silent</td>
</tr>
<tr>
<td></td>
<td>c.173G&gt;C</td>
<td>g.7002936G&gt;C</td>
<td>Exon 3</td>
<td>p.Trp58Ser</td>
</tr>
<tr>
<td></td>
<td>c.748G&gt;A</td>
<td>g.7014254G&gt;A</td>
<td>Exon 8</td>
<td>p.Al350Thr</td>
</tr>
<tr>
<td>CYP3A89</td>
<td>c.1010T&gt;C</td>
<td>g.7079070T&gt;C</td>
<td>Exon 10</td>
<td>p.Val337Ala</td>
</tr>
<tr>
<td></td>
<td>c.1095G&gt;A</td>
<td>g.7081607G&gt;A</td>
<td>Exon 11</td>
<td>Silent</td>
</tr>
<tr>
<td>CYP3A94</td>
<td>c.17G&gt;A</td>
<td>g.7133787G&gt;A</td>
<td>Exon 1</td>
<td>p.Ser6Asn</td>
</tr>
<tr>
<td></td>
<td>c.558T&gt;C</td>
<td>g.7145876T&gt;C</td>
<td>Exon 7</td>
<td>Silent</td>
</tr>
<tr>
<td>CYP3A95</td>
<td>c.640C&gt;G</td>
<td>g.7229727C&gt;G</td>
<td>Exon 7</td>
<td>p.His214Asp</td>
</tr>
<tr>
<td></td>
<td>c.718G&gt;C</td>
<td>g.7234153G&gt;C</td>
<td>Exon 8</td>
<td>p.Val240Leu</td>
</tr>
<tr>
<td></td>
<td>c.1174A&gt;T</td>
<td>g.7241704A&gt;T</td>
<td>Exon 11</td>
<td>p.Thr392Ser</td>
</tr>
<tr>
<td></td>
<td>c.1254G&gt;A</td>
<td>g.7243321G&gt;A</td>
<td>Exon 12</td>
<td>Silent</td>
</tr>
<tr>
<td>CYP3A96</td>
<td>c.437T&gt;A</td>
<td>g.7316374T&gt;A</td>
<td>Exon 6</td>
<td>p.Phe146Tyr</td>
</tr>
<tr>
<td></td>
<td>c.1496_1501delCCGTGA</td>
<td>g.7334180del</td>
<td>Exon 13</td>
<td>p.Thr499_</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Val500del</td>
</tr>
<tr>
<td>CYP3A97</td>
<td>c.126T&gt;A</td>
<td>g.7380650T&gt;A</td>
<td>Exon 2</td>
<td>Silent</td>
</tr>
<tr>
<td></td>
<td>c.356C&gt;T</td>
<td>g.7387399C&gt;T</td>
<td>Exon 5</td>
<td>p.Thr119Ile</td>
</tr>
<tr>
<td></td>
<td>c.1498G&gt;A</td>
<td>g.7413137G&gt;A</td>
<td>Exon 13</td>
<td>p.Val500Met</td>
</tr>
</tbody>
</table>

Table 1: CYP450 Gene Polymorphisms.

<table>
<thead>
<tr>
<th>CYP3A4</th>
<th>CYP3A5</th>
<th>CYP3A7</th>
<th>CYP3A43–1</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Identity at the nucleotide level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP3A93</td>
<td>85.4</td>
<td>83.7</td>
<td>83.8</td>
</tr>
<tr>
<td>CYP3A89</td>
<td>86.4</td>
<td>84.7</td>
<td>84.4</td>
</tr>
<tr>
<td>CYP3A94</td>
<td>85.3</td>
<td>83.2</td>
<td>83.6</td>
</tr>
<tr>
<td>CYP3A95</td>
<td>83.7</td>
<td>82.3</td>
<td>81.9</td>
</tr>
<tr>
<td>CYP3A96</td>
<td>84.3</td>
<td>83</td>
<td>82.5</td>
</tr>
<tr>
<td>CYP3A97</td>
<td>83.3</td>
<td>81.9</td>
<td>81.7</td>
</tr>
</tbody>
</table>

% Identity at the amino acid level

| CYP3A93  | 79.5    | 78.9    | 76.5      | 71.2      |
| CYP3A89  | 81.3    | 80.3    | 76.3      | 73.2      |
| CYP3A94  | 80.1    | 77.9    | 76.9      | 71.2      |
| CYP3A95  | 79.1    | 77.3    | 75.3      | 72        |
| CYP3A96  | 79.1    | 77.5    | 76.5      | 70.6      |
| CYP3A97  | 75.3    | 75.5    | 73.4      | 68.7      |

Published by Broad Institute.

Table 2: Percent identities at the cDNA level between the human CYP3A cluster and the equine variants found.

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 72–79
amino acid sequence, especially in the substrate binding regions of the enzyme (Thomas 2007; Qiu et al. 2008; Chen et al. 2009). It was hypothesized that phylogenetically stable genes have core functions in development and physiology and metabolize mainly endogenous substrates, whereas unstable genes have accessory metabolic functions associated with unstable environmental interactions such as toxin and pathogen exposure (Thomas 2007). The high number of CYP3A genes in the horse could possibly reflect a tendency of this herbivore to metabolize a wide variety of different substrates that might be contained in the forage. It could be assumed that wild horses had to be able to digest a huge variety of plant species depending on many environmental factors in different geographical regions and climate conditions.

CYP450 genes are known to be rich in variations and there are a higher number of polymorphisms in CYP450 genes than for other investigated gene sets (Solus et al. 2004). More than 30 SNPs have been identified for the human CYP3A4 gene, out of which about 20 produce an amino acid exchange (http://www.cypalleles.ki.se/cyp3a4.htm). In general, variants in the coding regions of CYP3A4 occur at allele frequencies <5% (Lamba et al. 2002). Allelic frequency of the variants in horses and the amount of variation between different breeds still remains to be analysed using population studies.

Automatic prediction of the influence of certain variations on protein function, as performed in this study, showed that out of all 13 recorded non-synonymous SNPs in the horse CYP3A cluster, only one was predicted to affect protein function. This finding coincides well with results from studies in humans. Although a strong inter-individual variability in CYP3A-dependent clearance is reported in man, it is considered that different allelic variants only contribute to the variability to a small extent. Inter-individual differences in enzyme expression as a result of a variety of factors seem to be more relevant (Lamba et al. 2002).

It is interesting to note that 14 of the 19 SNPs we found in the horse CYP3A cluster are positioned in the first half of the coding sequence (codons 1–250), whereas only 5 SNPs were found in the second half. When comparing the distribution of SNPs within the coding sequence between human CYP3A4 and the horse CYP3A, we found that the majority of SNPs were also in the first half of the coding sequence in the human CYP3A4 (http://www.cypalleles.ki.se/cyp3a4.htm).

This might underline the typical structure of the CYP450 family, which is usually a small N-terminal predominantly beta-strand domain with the membrane-spanning part and a larger helical C-terminal domain containing the functionally important active site and the haem (Williams et al. 2004).

One characteristic feature of human CYP3A4 is the phenylalanine cluster forming a hydrophobic core above the active site (Williams et al. 2004). In the horse CYP3A cluster, the phenylalanine at positions 108, 213, 219 and 220 was found to be sometimes replaced by a leucine or isoleucine. In contrast, positions 241 and 304 were highly conserved, and all CYP3As had a phenylalanine at that position. Functional differences between the different CYP3A proteins from this cluster are not easy to predict from the sequence level. In the human CYP3A cluster, the two most widely expressed isoforms CYP3A4 and CYP3A5 had 84% identity at the amino acid level (Guengerich 1997) and share most substrate specificities, whereas the number of substrates can differ between the isoforms (Daly 2006). In contrast, the human 2C subfamily has 4 known isoforms, CYP2C9, CYP2C8, CYP2C18 and 2C19, which are more than 80% homologous to each other in the amino acid sequence but can differ markedly in function (Guengerich 1997). In extreme cases, a single amino acid exchange can change substrate specificity completely (Lindberg & Negishi 1989; Ramarao & Kemper 1995).

Our failure to amplify the coding sequence of CYP3A129 from the cDNA of horse liver can imply different situations: either the gene, although showing a complete set of exons, does not code for a functional protein and is therefore one of the numerous pseudogenes or splice variants that have been described in many other species, or as another consequence, CYP3A129 could be an enzyme not expressed predominantly in the liver and may instead be found in other tissues at the protein level. This has been reported for the human CYP3A43, which is expressed in prostate and testis rather than in liver. Sex-dependent expression of CYP450s has been predominantly described in rats and was hypothesized to be a consequence of several situations, including extensive inbreeding (Mugford & Kedderis 1998). It may be less likely that the horse shows a sex specific set of CYP450s like the strongly inbred rat. However, a male horse should also be tested for the expression of CYP3A129 to rule this out.

In humans, a correlation between CYP3A5 genotype and cancer susceptibility has been drawn. Certain combined CYP3A4/CYP3A5 haplotypes show differential susceptibility to prostate cancer (Rebeck et al. 1998; Paris et al. 1999). CYP3A5*1 homozygotes may have higher systolic blood pressure (Givens et al. 2003; Qiu et al. 2008). More research is necessary before such conclusions can be drawn for horses.

In conclusion, orthologous genes for the CYP3A family exist in horses, but their number differs from the human CYP3A gene family. CYP genes of the same family show high homology within and between species but can be highly polymorphic. These data provide a basis for functional analyses of the equine CYP3A family and perhaps some day in the future drug treatments of horses will be adjusted to their individual CYP genotypes to increase the efficacy of pharmacological interventions.
Acknowledgement

This study was supported by a grant from the Swiss National Science Foundation.

Conflicts of interest

The authors have declared no potential conflicts.

References


Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Primers placed in the UTR for amplification of coding sequence from the cDNA template.

Table S2 Internal sequencing primers used for several CYP450s.

Table S3 Internal sequencing primers.

Table S4 Primers for PCR on genomic DNA.


© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 72–79.
Table S5 Internal sequencing primers for genomic PCR products.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
A genome-wide scan for tying-up syndrome in Japanese Thoroughbreds

T. Tozaki*, K. Hirota*, S. Sugita†, N. Ishida†, T. Miyake‡, H. Oki† and T. Hasegawa†

*Department of Molecular Genetics, Laboratory of Racing Chemistry, Utsunomiya, Tochigi, Japan. †Equine Research Institute, Japan Racing Association, Utsunomiya, Tochigi, Japan. ‡Comparative Agricultural Sciences, Graduate school of Agriculture, Kyoto University, Sakyo, Kyoto, Japan.

Summary

Tying-up syndrome, also known as recurrent exertional rhabdomyolysis in Thoroughbreds, is a common muscle disorder for racehorses. In this study, we performed a multipoint linkage analysis using LOKI based on the Bayesian Markov chain Monte Carlo method using 5 half-sib families (51 affected and 277 nonaffected horses in total), and a genome-wide association study (GWAS) using microsatellites (144 affected and 144 nonaffected horses) to map candidate regions for tying-up syndrome in Japanese Thoroughbreds. The linkage analysis identified one strong L-score (82.45) between the loci UCDEQ411 and COR058 (24.9–27.9 Mb) on ECA12. The GWAS identified two suggestive genomic regions on ECA12 (24.9–27.8 Mb) and ECA20 (29.3–33.5 Mb). Based on both results, the genomic region between UCDEQ411 and TKY499 (24.9–27.8 Mb) on ECA12 was the most significant and was considered as a candidate region for tying-up syndrome in Japanese Thoroughbreds.

Keywords horse, microsatellite, recurrent exertional rhabdomyolysis, Thoroughbred, tying-up syndrome.

Introduction

Tying-up syndrome is a common intermittent condition that primarily affects the muscles in horses. It is characterized by clinical signs ranging from slight stiffness to immobility, signs of pain and reluctance to move, which may develop after mild-to-moderate exercises (McLean 1973; Rossdale et al. 1985; Harris 1989, 1991). It is also known as azoturia, exertional rhabdomyolysis, paralytic myoglobinuria, Monday morning disease and exertional myopathy, because the clinical signs of this syndrome vary. Tying-up syndrome occurs because of specific inherited abnormalities and can be classified into two distinct types: polysaccharide storage myopathy (PSSM) in Quarter horses and recurrent exertional rhabdomyolysis (RER) in Thoroughbreds. A mutation in the gene encoding skeletal muscle glycogen synthase (GYS1) was recently reported to be strongly associated with PSSM in Quarter horses and Belgian Draught horses (McCue et al. 2008a, b; Stanley et al. 2009). This GYS1 point mutation appears to be a gain-of-function mutation that results in the accumulation of a glycogen-like, less-branched polysaccharide in skeletal muscles. It is inherited as a dominant trait (McCue et al. 2008b).

For more than a century, tying-up syndrome, also known as RER, in Thoroughbreds has been recognized as a syndrome of muscle pain and cramping associated with exercises (McKenzie et al. 2003). This syndrome occurs in approximately 5.0% of all Thoroughbreds (MacLeay et al. 1999; McGowan et al. 2002) in the United Kingdom and United States of America and is hence responsible for substantial economic loss. A statistical genetic analysis of pedigrees indicated that the susceptibility of Thoroughbreds to tying-up syndrome is a heritable trait (Oki et al. 2005). Although the occurrence of this syndrome may be influenced by multiple factors such as sex, temperament and diet, an in vitro muscle contracture test involving various breeding trials suggested that tying-up syndrome in Thoroughbreds can be modelled as a genetic trait with an autosomal dominant mode of inheritance (Dranchak et al. 2005). Many studies based on proteomic approaches have been performed to identify the susceptibility genes for tying-up syndrome (Lentz et al. 2002; McKenzie et al. 2002,
2004). However, no critical susceptibility genes have yet been identified. A recent linkage analysis using the microsatellites around the RYR1, CACNA1S and ATP2A1 genes, which were expected to be susceptibility genes because of their role in myoplasmic calcium regulation in skeletal muscles, did not reveal associations with tying-up syndrome in Thoroughbreds (Dranchuk et al. 2006).

Many significant advances have recently been achieved in horse genome mapping, and linkage maps of horses have been created (Penedo et al. 2005; Swinburne et al. 2006). Horse–human comparative maps were published in 2007 and 2008 (Tozaki et al. 2007a; Raudsepp et al. 2008). With the complete horse genome sequence available to the equine genome community, we expect an acceleration in discoveries of genetic traits of interest to horse breeders (Wade et al. 2009). In our previous studies, we proposed a genome-wide association study (GWAS) that can be systematically performed by combining primary screening for genome-wide markers with pooled DNA samples and secondary screening for candidate markers implicated from the genome-wide scan in Thoroughbreds (Tozaki et al. 2005, 2007b).

In this study, we aimed at identifying candidate genomic regions susceptible for tying-up syndrome in Japanese Thoroughbred populations by performing a Markov chain Monte Carlo (MCMC) linkage analysis and a GWAS using microsatellites.

Materials and methods

Thoroughbreds and DNA purification

Thoroughbreds with and without tying-up syndrome were obtained from the Ritto and Miho Training centers of the Japan Racing Association (JRA). In terms of clinical infrastructure and expertise, these training centers of the JRA offer state-of-the-art horse clinic facilities for the diagnosis and treatment of various equine disorders. All the clinical information regarding the JRA’s racehorses is stored in a computerized database system (JARIS: Japan Racing Information System). Of the 3927 Thoroughbreds registered as JRA racehorses on October 2002, 223 Thoroughbreds with tying-up syndrome were selected for DNA isolation. On the basis of the information available on the relatives of the 223 horses, five paternal half-sib Thoroughbred families (designated A through E) were then constructed for linkage analysis using LOKI (Table 1). The half-sib families were unrelated for at least three generations of paternal pedigree. Furthermore, from these 223 horses, we selected 144 individuals that had no paternal, maternal or sibling relationships with each other for the GWAS. The control population comprised 144 normal horses without tying-up syndrome that had been randomly selected from both training centers; these horses had more racing experience and also had no paternal, maternal or sibling relationships with each other. No statistically significant differences were observed between both populations for the ratio of gender and/or for training centre.

All blood samples for DNA isolation were stored at −40 °C. Genomic DNA was extracted from the stored blood by using the MagExtractor System MFX-2000 (Toyobo, Osaka, Japan). The concentrations of the DNA samples were measured twice using a multi-wavelength (200–310 nm) absorption spectrometer (DU®7500 Spectrophotometer; Beckman Coulter, Brea, CA, USA) and diluted to a concentration of 25 ng/µl. The concentrations of the samples were measured again before they were combined for the case–control association studies. Finally, the DNA samples were combined with their respective pools. The concentration of all the pooled DNA samples was adjusted to 25 ng/µl.

Microsatellites

For the linkage analysis, 117 microsatellites on autosomal chromosomes were selected from the workshop linkage map (Penedo et al. 2005). For the GWAS performed using pooled DNA, 986 microsatellites on autosomal chromosomes were used as genetic markers for the first screening for the genome-wide scan. The markers for the GWAS were mainly selected from the workshop linkage map, with an average interval of 6.3 cM between markers using 766 microsatellites. The microsatellites mapped on the AHT linkage map (Swinburne et al. 2006) and a human–horse comparative map (Tozaki et al. 2007a) were also added to increase the marker density. The added marker improved marker intervals of over 10 cM in the workshop linkage map, and it was expected that all the marker intervals would become less than 10 cM in the first screening. In the second screening, the additional microsatellites on the chromosomes that showed statistical significance for tying-up syndrome in the first screening were used as markers for individual genotyping.

Microsatellite genotyping

For microsatellite genotyping, we used the following 3 primers: a sequence-specific forward primer conjugated with

### Table 1 Number of tying-up and non-tying-up offspring for each sire family.

<table>
<thead>
<tr>
<th>Family</th>
<th>Tying-up</th>
<th>Non-tying-up</th>
<th>Disease rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>72</td>
<td>21.7</td>
</tr>
<tr>
<td>B</td>
<td>11</td>
<td>86</td>
<td>11.3</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
<td>66</td>
<td>12.0</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
<td>23</td>
<td>23.3</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>30</td>
<td>11.8</td>
</tr>
<tr>
<td>Total</td>
<td>51¹</td>
<td>277</td>
<td>18.4</td>
</tr>
</tbody>
</table>

¹These were sampled from 233 affected horses.
a 5'-TGACCGGCAGCAAAATTG-3' tail at its 5' end, a sequence-specific reverse primer, and a FAM-labelled 5'-TGACCAGCAGCAAAATTG-3' primer (Applied Biosystems, Foster City, CA, USA). The underlined sequences were used for fluorescence detection. PCR was performed as described in Tozaki et al. (2001). The reaction products were analysed using the ABI 3130 genetic analyzer (Applied Biosystems), and the genotyping data were processed by the Genotyper software (Applied Biosystems) to identify alleles and to measure the peak height of the alleles for each microsatellite.

Linkage analysis using LOKI

We carried out multipoint linkage analysis assuming multiple putative quantitative trait loci (QTL) simultaneously in the oligogenic model. We used the LOKI 2.4 program (http://www.stat.washington.edu/thompson/Genepi/Loki.shtml), which is based on the Bayesian reversible jump MCMC method (Heath 1997), and can estimate the posterior distribution of a number of parameters of interest (e.g. the number of QTL contributing to the trait, their location and the genotype effects for each QTL). The analytical method basically followed the study of Shmulewitz et al. (2006). The LOKI results are reported as L-scores, which, as estimates of the Bayes factor (Kass & Raftery 1995), show the posterior ratio between the probability that a QTL signal is 'real' and that it is because of chance alone. To compute the L-scores, the prior probability \( p \) of finding a QTL linked to a 1-cM bin is 1/\( t \), where \( t \) is the total map length of the genome. For a given iteration, the prior probability that at least 1 QTL is linked to a 1-cM bin is 1 - (1 - 1/t)^n, where \( n \) is the number of QTL in the model at that iteration. The posterior probability \( q \) is 1 or 0, depending on whether or not a QTL is linked. The posterior \( p/q \) ratio for each linkage group is then averaged over iterations to give the L-score (for more information, see Neuman et al. (2002) or the LOKI usage manual). For all the other parameters, uniform priors were adopted. According to the Bayes factor calibration tables, an L-score of >10 is suggestive of linkage while an L-score of >20 provides evidence of linkage (Raftery 1996).

In this multipoint linkage analysis, phenotypic information regarding tying-up syndrome was recorded as binary data (0.0 or 1.0) but handled as continuous quantitative characters during the computation by the LOKI program. The pedigrees of all the Thoroughbreds were traced back for three generations. The total number of MCMC chain lengths was 30,000, with a burn-in of 10,000. The convergence of the estimates was confirmed by at least 2 runs with different starting seeds, especially for the QTL location with the largest L-score. The linkage analysis was separately performed for each half-sib family.

GWAS

The differences in the total allele content (TAC) in the microsatellites between different DNA pools were evaluated by calculating \( \Delta \text{TAC} \) (Collins et al. 2000; Tozaki et al. 2007b). The peak height of each allele in a pool was determined from an electrophoretogram, and the sum of all the peak heights in the pool was calculated. The height of each allele was then divided by the sum of all peak heights and multiplied by 100 to obtain the allele percentage. The two pools were then compared, and the absolute value of the difference in the percentage for the two pools was computed for each allele. \( \Delta \text{TAC} \) was then computed using the following formula:

\[
\Delta \text{TAC} = \sum \text{ABS}(X_1 - X_2)/2,
\]

where \( X_1 \) is the allele percentage in pool 1 and \( X_2 \) is the allele percentage in pool 2. In the first screening, a \( \Delta \text{TAC} \) value of >10 was regarded as positive (Collins et al. 2000; Tozaki et al. 2007b). This procedure was repeated for each of the markers.

A chi-square test of association using the genotypes of individual horses was performed to compare the allele frequency between the case and control populations for seven microsatellites that were chosen during the first screening. Additionally, 13 microsatellites located near the positive microsatellites in the first screening were included. The SNPAlyze ver. 7.0 Standard program (Dynacom, Yokohama, Japan) was used for computation.

In addition, the permutation test was performed to test for deviation in the allele frequencies of markers and haplotypes. Haplotype frequencies for two markers were estimated by the maximum-likelihood method with an expectation-maximization algorithm. Distribution of the test statistics was estimated by evaluating the statistics for a random sampling of 10,000 iterated permutations; the total numbers of both cases and controls were fixed to avoid false-positive results in multiple testing. We calculated the LD between pairs of markers by using the standard definition of \( D' \). The SNPAlyze ver. 7.0 Standard program was used for computation.

Results and discussion

Linkage analysis with LOKI

The MCMC linkage analysis performed using LOKI identified a significant L-score on ECA12 (82.45) only in family B (Table 2). The location for the best L-score was 26.5 cM on ECA12, which produced a sharp peak (Fig. 1) between the loci UCDEQ411 and COROS58. This suggests that the region between these markers (24.9–27.9 Mb) on ECA12 is a candidate region for tying-up syndrome in Japanese Thoroughbreds. In the other families, namely, families A, C, D and E, a suggestive L-score (≥10) for linkage is not observed.
at this position (Table 2). The expected number of QTL for each family is listed in Table 3. On average, one or two QTL are expected to segregate in these families. However, considering the low L-scores in the families A, C, D and E, the expected number of QTL in those families would not be suitable for estimation. Therefore, the expected number of QTL was calculated as 1.51 for family B, suggesting the presence of at least one QTL segregating in family B.

If only outbred sire half-sib family populations are used in the linkage analysis, then only the sire heterozygous for the target QTL will be useful for QTL detection. In families A, C, D and E, the QTL of the sires used could be homozygous on ECA12. In addition, considering the small sample size for each family (Table 1), the power for the detection of QTL would be low, especially for families D and E.

GWAS

In the first screening, pooled DNA samples from the case and control populations were used for the GWAS involving 986 microsatellites. From these, seven microsatellites with a ΔTAC of >10, namely, COR028 on ECA3; TKY331 on ECA5; TKY499, COR058 and UCDEQ411 on ECA12; UM011 on ECA20; and L12.2 on ECA29, were subjected to the second screening with individual genotyping of the 144 case and control samples. The microsatellites UM011 (P = 0.0055) on ECA20 and TKY499 (P = 0.0258) on ECA12 showed statistically significant associations (P < 0.05) for tying-up syndrome without Bonferroni correction; however, UM011 showed a statistically significant association only with Bonferroni correction (i.e. significant level with 0.05/7) in the second screening. Those suggestive genomic regions were then subjected to the third screening. The third screening was a permutation test to evaluate the differences in the haplotype frequencies of two neighbouring microsatellites (Table 4). Several microsatellites located near the positive markers on each chromosome, namely, AHT017 and UCDEQ497 on ECA12; and LEX064, TKY136, UMNe056, TKY694, HMS082, UMNe065, TKY1115, TKY821, TKY539, TKY547 and

<table>
<thead>
<tr>
<th>Family</th>
<th>L-score</th>
<th>Position (cM)</th>
<th>L-score</th>
<th>Position (cM)</th>
<th>L-score</th>
<th>Position (cM)</th>
<th>L-score</th>
<th>Position (cM)</th>
<th>L-score</th>
<th>Position (cM)</th>
<th>L-score</th>
<th>Position (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5.19</td>
<td>25.5</td>
<td>1.86</td>
<td>145.5</td>
<td>1.44</td>
<td>30.5</td>
<td>4.97</td>
<td>64.5</td>
<td>1.96</td>
<td>132.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1.63</td>
<td>21.5</td>
<td>1.85</td>
<td>78.5</td>
<td>1.55</td>
<td>4.5</td>
<td>1.42</td>
<td>55.5</td>
<td>1.82</td>
<td>44.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>8.15</td>
<td>28.5</td>
<td>1.22</td>
<td>64.5</td>
<td>1.80</td>
<td>14.5</td>
<td>3.08</td>
<td>84.5</td>
<td>1.73</td>
<td>100.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>7.50</td>
<td>38.5</td>
<td>2.13</td>
<td>20.5</td>
<td>1.74</td>
<td>6.5</td>
<td>6.40</td>
<td>12.5</td>
<td>2.06</td>
<td>95.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>2.23</td>
<td>28.5</td>
<td>0.91</td>
<td>88.5</td>
<td>6.52</td>
<td>95.5</td>
<td>1.84</td>
<td>0.5</td>
<td>1.73</td>
<td>44.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.50</td>
<td>70.5</td>
<td>3.95</td>
<td>97.5</td>
<td>2.23</td>
<td>89.5</td>
<td>1.91</td>
<td>3.5</td>
<td>1.54</td>
<td>82.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.19</td>
<td>28.5</td>
<td>1.35</td>
<td>40.5</td>
<td>1.49</td>
<td>23.5</td>
<td>2.87</td>
<td>40.5</td>
<td>1.58</td>
<td>12.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.78</td>
<td>50.5</td>
<td>0.82</td>
<td>21.5</td>
<td>1.60</td>
<td>88.5</td>
<td>1.78</td>
<td>85.5</td>
<td>1.77</td>
<td>106.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.72</td>
<td>10.5</td>
<td>1.10</td>
<td>11.5</td>
<td>1.93</td>
<td>55.5</td>
<td>1.20</td>
<td>7.5</td>
<td>2.36</td>
<td>8.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.92</td>
<td>0.5</td>
<td>0.65</td>
<td>23.5</td>
<td>1.37</td>
<td>23.5</td>
<td>5.70</td>
<td>23.5</td>
<td>1.81</td>
<td>4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.62</td>
<td>5.5</td>
<td>0.78</td>
<td>16.5</td>
<td>1.37</td>
<td>6.5</td>
<td>1.17</td>
<td>7.5</td>
<td>2.26</td>
<td>5.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.33</td>
<td>16.5</td>
<td>82.45</td>
<td>26.5</td>
<td>1.62</td>
<td>10.5</td>
<td>1.29</td>
<td>41.5</td>
<td>1.28</td>
<td>14.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.14</td>
<td>9.5</td>
<td>0.34</td>
<td>7.5</td>
<td>1.07</td>
<td>17.5</td>
<td>3.96</td>
<td>21.5</td>
<td>1.61</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.98</td>
<td>68.5</td>
<td>1.02</td>
<td>90.5</td>
<td>1.75</td>
<td>77.5</td>
<td>1.32</td>
<td>124.5</td>
<td>1.88</td>
<td>25.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.61</td>
<td>1.5</td>
<td>1.01</td>
<td>43.5</td>
<td>2.74</td>
<td>76.5</td>
<td>1.94</td>
<td>17.5</td>
<td>1.41</td>
<td>37.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.09</td>
<td>52.5</td>
<td>2.42</td>
<td>42.5</td>
<td>2.79</td>
<td>85.5</td>
<td>3.08</td>
<td>73.5</td>
<td>1.82</td>
<td>81.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.98</td>
<td>13.5</td>
<td>0.74</td>
<td>22.5</td>
<td>1.39</td>
<td>27.5</td>
<td>2.05</td>
<td>38.5</td>
<td>2.08</td>
<td>29.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.00</td>
<td>4.5</td>
<td>7.89</td>
<td>40.5</td>
<td>1.42</td>
<td>8.5</td>
<td>4.96</td>
<td>22.5</td>
<td>1.80</td>
<td>4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.68</td>
<td>1.5</td>
<td>1.23</td>
<td>15.5</td>
<td>1.38</td>
<td>6.5</td>
<td>1.63</td>
<td>2.5</td>
<td>1.90</td>
<td>17.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.13</td>
<td>2.5</td>
<td>0.99</td>
<td>5.5</td>
<td>1.69</td>
<td>49.5</td>
<td>1.14</td>
<td>68.5</td>
<td>1.21</td>
<td>51.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.74</td>
<td>6.5</td>
<td>0.96</td>
<td>66.5</td>
<td>3.81</td>
<td>63.5</td>
<td>2.24</td>
<td>30.5</td>
<td>1.88</td>
<td>16.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.33</td>
<td>44.5</td>
<td>1.17</td>
<td>60.5</td>
<td>1.34</td>
<td>4.5</td>
<td>1.96</td>
<td>63.5</td>
<td>1.52</td>
<td>51.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.94</td>
<td>0.5</td>
<td>0.26</td>
<td>0.5</td>
<td>1.00</td>
<td>0.5</td>
<td>0.58</td>
<td>0.5</td>
<td>1.78</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.12</td>
<td>2.5</td>
<td>0.57</td>
<td>10.5</td>
<td>1.61</td>
<td>7.5</td>
<td>0.62</td>
<td>34.5</td>
<td>0.89</td>
<td>19.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.98</td>
<td>21.5</td>
<td>3.23</td>
<td>30.5</td>
<td>1.24</td>
<td>27.5</td>
<td>1.09</td>
<td>29.5</td>
<td>2.06</td>
<td>29.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.01</td>
<td>8.5</td>
<td>1.65</td>
<td>2.5</td>
<td>1.15</td>
<td>1.5</td>
<td>1.52</td>
<td>8.5</td>
<td>1.09</td>
<td>11.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.94</td>
<td>14.5</td>
<td>1.24</td>
<td>6.5</td>
<td>1.22</td>
<td>2.5</td>
<td>2.97</td>
<td>4.5</td>
<td>1.44</td>
<td>11.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>14.5</td>
<td>0.01</td>
<td>6.5</td>
<td>0.01</td>
<td>2.5</td>
<td>0.01</td>
<td>4.5</td>
<td>0.01</td>
<td>11.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.57</td>
<td>19.5</td>
<td>1.06</td>
<td>15.5</td>
<td>1.31</td>
<td>48.5</td>
<td>1.69</td>
<td>7.5</td>
<td>1.79</td>
<td>37.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.27</td>
<td>2.5</td>
<td>0.59</td>
<td>11.5</td>
<td>1.67</td>
<td>19.5</td>
<td>2.02</td>
<td>22.5</td>
<td>1.74</td>
<td>8.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.87</td>
<td>10.5</td>
<td>1.45</td>
<td>2.5</td>
<td>1.32</td>
<td>38.5</td>
<td>1.38</td>
<td>38.5</td>
<td>0.88</td>
<td>20.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1 L-score for ECA12 in Family B constructed by MCMC linkage analysis.

Table 3 The expected number of QTL for tying-up syndrome with LOKI analyses.

<table>
<thead>
<tr>
<th>QTL Number</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family A</td>
<td>0.0000</td>
<td>0.6159</td>
<td>0.2757</td>
<td>0.0800</td>
<td>0.0223</td>
<td>0.0050</td>
<td>0.0010</td>
<td>1.5275</td>
</tr>
<tr>
<td>Family B</td>
<td>0.0000</td>
<td>0.6203</td>
<td>0.2785</td>
<td>0.0725</td>
<td>0.0204</td>
<td>0.0046</td>
<td>0.0031</td>
<td>1.5180</td>
</tr>
<tr>
<td>Family C</td>
<td>0.0125</td>
<td>0.3442</td>
<td>0.2958</td>
<td>0.1854</td>
<td>0.0839</td>
<td>0.0424</td>
<td>0.0200</td>
<td>2.1596</td>
</tr>
<tr>
<td>Family D</td>
<td>0.0000</td>
<td>0.5208</td>
<td>0.2926</td>
<td>0.1181</td>
<td>0.0465</td>
<td>0.0135</td>
<td>0.0032</td>
<td>1.7330</td>
</tr>
<tr>
<td>Family E</td>
<td>0.0067</td>
<td>0.3595</td>
<td>0.2937</td>
<td>0.1790</td>
<td>0.0962</td>
<td>0.0384</td>
<td>0.0167</td>
<td>2.1609</td>
</tr>
</tbody>
</table>

Table 4 Microsatellites tested on chromosome 12 and 20 for tying-up susceptibility.

<table>
<thead>
<tr>
<th>ECA</th>
<th>Markers</th>
<th>Mb</th>
<th>cM</th>
<th>cM</th>
<th>P-value</th>
<th>Tying-up</th>
<th>Non-tying-up</th>
<th>Marker pairs</th>
<th>P-value</th>
<th>LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>UCDEQ411</td>
<td>24.9</td>
<td>25.6</td>
<td>37.6</td>
<td>0.2051</td>
<td>8.54E-01</td>
<td>1.24E-01</td>
<td>UCDEQ411/TKY499</td>
<td>0.0051</td>
<td>0.25</td>
</tr>
<tr>
<td>12</td>
<td>TKY499</td>
<td>27.8</td>
<td>0.0258</td>
<td>9.82E-01</td>
<td>4.17E-01</td>
<td>0.0429</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>COR058</td>
<td>27.9</td>
<td>25.6</td>
<td>46.8</td>
<td>0.1470</td>
<td>7.30E-02</td>
<td>1.25E-02</td>
<td>COR058/AHT017</td>
<td>0.2689</td>
<td>0.25</td>
</tr>
<tr>
<td>12</td>
<td>AHT017</td>
<td>66.4</td>
<td>0.5778</td>
<td>9.49E-01</td>
<td>2.73E-01</td>
<td>0.3864</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>UCDEQ497</td>
<td>32.5</td>
<td>58.0</td>
<td>70.2</td>
<td>0.7302</td>
<td>4.51E-01</td>
<td>3.42E-01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>LEX064</td>
<td>15.3</td>
<td>49.9</td>
<td>0.2860</td>
<td>6.85E-01</td>
<td>8.26E-01</td>
<td>LEX064/TKY136</td>
<td>0.7958</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TKY136</td>
<td>20.4</td>
<td>0.2835</td>
<td>9.13E-01</td>
<td>5.97E-01</td>
<td>0.0393</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>UMNNo56</td>
<td>29.3</td>
<td>23.4</td>
<td>0.0200</td>
<td>6.81E-01</td>
<td>7.00E-01</td>
<td>UMNNo56/TKY694</td>
<td>0.0817</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TKY694</td>
<td>31.2</td>
<td>0.0109</td>
<td>5.14E-01</td>
<td>3.43E-01</td>
<td>0.0611</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>HMS082</td>
<td>26.1</td>
<td>0.0190</td>
<td>5.12E-01</td>
<td>3.20E-01</td>
<td>0.0169</td>
<td>0.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>UMNNo65</td>
<td>31.5</td>
<td>0.0095</td>
<td>3.36E-01</td>
<td>3.22E-01</td>
<td>0.1712</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>UMN011</td>
<td>33.5</td>
<td>27.6</td>
<td>0.0055</td>
<td>6.61E-01</td>
<td>2.36E-02</td>
<td>UMN011/TKY1115</td>
<td>0.0152</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TKY1115</td>
<td>34.9</td>
<td>0.0457</td>
<td>8.56E-01</td>
<td>8.85E-02</td>
<td>0.1612</td>
<td>0.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TKY821</td>
<td>39.7</td>
<td>0.5675</td>
<td>6.3E-01</td>
<td>3.39E-01</td>
<td>0.9583</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TKY539</td>
<td>40.3</td>
<td>0.7375</td>
<td>6.13E-01</td>
<td>3.63E-01</td>
<td>0.2774</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TKY547</td>
<td>40.8</td>
<td>0.5920</td>
<td>9.97E-01</td>
<td>7.61E-02</td>
<td>0.7054</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TKY1048</td>
<td>48.9</td>
<td>0.4340</td>
<td>8.75E-01</td>
<td>2.24E-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 The genomic position of chromosome 12 and 20 (EquCab2.0) is given in Mbp.
2 The positions of markers along equine chromosome 12 and 20, according to the AHT linkage map, are given in cM.
3 The positions of markers along equine chromosome 12 and 20, according to the international equine gene mapping workshop report, are given in cM.
4 P-values based on permutation test.
TKY1048 on ECA20, were also genotyped (Table 4). Among these microsatellites, the microsatellites located in the region between 29.3 and 34.9 Mb on ECA20 also showed a statistically significant association ($P < 0.05$) for tying-up syndrome. For the microsatellite pairs, significant permutation $P$ values were also observed on ECA12 and ECA20 (Table 4), i.e. UCDEQ411/TKY499 ($P = 0.0051$) on ECA12 and HMS082/UMNe065 ($P = 0.0169$) and UM011/TKY1115 ($P = 0.0152$) on ECA20. The results for UCDEQ411/TKY499 ($P = 0.0051$) on ECA12 corresponded well with the results for linkage analysis, with a high L-score (Fig. 1). Although the linkage analysis did not detect any QTL on ECA20, the results of GWAS showed statistically significant associations between markers on ECA12 and ECA20, suggesting that ECA12 and ECA20 are candidate susceptibility regions for tying-up syndrome in Japanese Thoroughbreds.

The power of whole-genome LD screening depends on the number of markers analysed, the number of samples in the case and control populations, the extent of LD in the horse genome and the marker density required for effective association. The number of case and control samples used in this study ($N = 144$) may be rather small to detect the susceptibility loci for tying-up syndrome. Furthermore, we could not clarify the mode of inheritance of the candidate QTL, probably because of the use of the less informative binary phenotype. Further investigations into the Thoroughbreds diagnosed with tying-up syndrome in other racing organizations and countries are expected.

It is known that the linkage analysis is not influenced by population stratification. Therefore, linkage analysis using LOKI is useful for the detection of QTL for complex traits in outbred populations. The combination of linkage analysis with LOKI and our GWAS approach with microsatellites (Tozaki et al. 2007b) can accelerate the detection of QTL or candidate regions for complex traits.

A previous study reported that tying-up syndrome occurred in 7.7% of the JRA’s Thoroughbreds and that the prevalence of this syndrome in fillies, colts and geldings was 11.0%, 5.7%, and 5.8%, respectively (Oki et al. 2005). The gender-based difference in prevalence may support the theory that this syndrome is a sex-linked disorder. Further investigation involving the sex chromosomes will thus be worthwhile. Oki et al. (2005) also clarified that tying-up syndrome is heritable and estimated the heritability to be approximately 0.42 by using the threshold model. It is thought that the ECA12 and ECA20 candidate regions identified in this study may partly explain the heritability of tying-up syndrome.

## Conclusion

In this study, we performed MCMC linkage analysis and GWAS and identified two candidate regions on horse chromosomes ECA12 and ECA20 for tying-up syndrome in Japanese Thoroughbreds. The results obtained by both methods provided strong evidence for the presence of QTL on ECA12.

Although the recent advances in the use of single-nucleotide polymorphisms (SNPs) in genomic studies are impressive, we still believe that our GWAS approach using microsatellites will be quite useful in cases where sufficient genomic information or tools (such as SNP chips) are not available. The GWAS described here was performed using pooled DNAs and is a cost-effective method that can also be used for many livestock animals.

## Acknowledgements

We thank the Ritto and Miho Training centers of the JRA for providing samples from their horses as study materials. We also thank the Equine Department of JRA for approving and supporting the study with a grant-in-aid (2005–2007).

## Conflicts of interest

The authors have not declared any potential conflicts.

## References


Fine mapping of a quantitative trait locus for osteochondrosis on horse chromosome 2

C. Dierks, K. Komm, V. Lampe and O. Distl
Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, 30559 Hannover, Germany

Summary
In this study, we refine a quantitative trait locus for equine osteochondrosis (OC) on horse chromosome (ECA) 2 to a genome-wide significant interval at 20.08–30.94 Mb. The marker set contained 27 newly developed microsatellites equidistantly distributed over ECA2 and 44 nucleotide polymorphisms, located in 16 positional candidate genes for OC. Genotyping was performed in 211 Hanoverian horses from 14 paternal half-sib groups. A NCDN-associated SNP and haplotype were significantly associated with OC in fetlock and/or hock joints. This study is a further step towards the identification of genes responsible for OC in horses.

Keywords association, horse, neurochondrun, osteochondrosis.

Osteochondrosis (OC) is an inherited orthopaedic disorder frequently detected radiographically in young horses (Stock et al. 2005; Wittwer et al. 2006). Heritabilities for OC have been estimated in different horse breeds in a range of $h^2 = 0.10–0.37$ (Schober 2003; Stock et al. 2005; Wittwer et al. 2007; Van Grevenhof et al. 2009). A focal disturbance of chondrocyte differentiation and maturation during the enchondral ossification, most commonly seen in fetlock, hock and stifle joints, is causing OC in horses. Osteochondrotic lesions include subchondral bone cysts, fissures, cartilage flaps and osteochondral fragments. Osteochondral fragments are present in case of an osteochondrosis dissecans (OCD). The clinical manifestation of OC may lead to premature retirement of the animal as a result of chronic or recurrent lameness (Stock & Distl 2006).

A chromosome-wide significant quantitative trait locus (QTL) was identified for fetlock and hock OC in Hanoverian warmblood horses at 22.0–49.0 cM (this corresponds to 17.80–32.50 Mb on ECA2). These chromosome-wide significance levels were empirically determined using 1000 replicates (Dierks et al. 2007). The objective of this study was to confirm and refine the position of this QTL using a dense marker set based on newly developed microsatellites and single nucleotide polymorphisms (SNPs) and to identify significant associations of candidate gene-associated SNPs with OC.

We used genomic DNA from 211 horses including 104 foals, 99 mares and eight stallions of 14 paternal half-sib families that were not closely related (Table S1). A detailed description of these families and the scheme for diagnosis of OC and OCD can be found elsewhere (Dierks et al. 2007).

We supplemented the previously employed marker set for ECA2 with a total of 37 informative and evenly spaced microsatellites. From previously published equine linkage maps, 10 microsatellites could be selected for improvement of the marker set. For a uniform coverage of ECA2, 27 microsatellites had to be newly developed (Table S2).

Genotyping of microsatellites was performed as described by Dierks et al. (2007), and details can be found in Table S2. Screening and genotyping for candidate gene-associated polymorphisms was performed by sequencing all PCR products in both directions on a MegaBACE 1000 capillary sequencer (GE Healthcare) for all horses. The sequencing reactions were carried out using the DYEnamic ET Terminator Cycle Sequencing kit (GE Healthcare). A total of 64 SNPs and one indel mutation were identified within 16 positional candidate genes (Tables S3 and S4). These 16 genes were chosen according to their function in cartilage metabolism and their map position to cover the highly linked region at 20–30 Mb with SNPs in 1–2 Mb intervals.

After confirming Mendelian inheritance and correctness of marker transmission in the available pedigrees, multipoint non-parametric linkage analysis was performed using the MERLIN software (multipoint engine for rapid likelihood inference, version 1.1.2) (Abecasis et al. 2002) for all
informative markers on ECA2, including 62 microsatellites, 43 SNPs and one indel mutation. Zmeans and LOD scores were used to test for the proportion of alleles shared identical by descent (IBD) by affected individuals for the considered marker loci (Kong & Cox 1997). The maximum (minimum) achievable Zmean was 12.75 (~2.86) for OC in fetlock and hock joints. The corresponding maximum (minimum) value for the LOD score was 8.38 (~0.44). For OCD in fetlock and hock joints, the maximum (minimum) achievable Zmean was 7.62 (~2.81), and the corresponding values for LOD scores were 5.47 and ~0.50. These Zmeans and LOD scores were high enough to achieve genome-wide error probabilities. Chromosome-wide error probabilities were obtained using a permutation procedure as described by Dierks et al. (2007). Genome-wide probabilities were obtained by applying a Bonferroni correction with $P_{\text{genome-wide}} = 1 - (1 - P_{\text{chromosome-wide}})^{1/r}$, where $r =$ length of ECA2 (120.86 Mb) divided by the total equine genome length in Mb (2680 Mb). The data were evaluated using the ALLELE, CASECONTROL and HAPLOTYPE procedures of SAS/Genetics (Statistical Analysis System, Version 9.2; SAS Institute, 2009) to determine the observed heterozygosity (HET), polymorphism information content (PIC), Hardy–Weinberg equilibrium and to test for genotypic and allelic association with OC. Pairwise linkage disequilibrium (LD) among SNPs was calculated and pictured using HAPLOVIEW 4.0 (http://www.broad.mit.edu/mpg/haploview/1) (Barrett et al. 2005). We used the Tagger algorithm $r^2 \geq 0.8$ (de Bakker et al. 2005) to detect SNPs with strong LD among alleles.

The mean distance among the markers on ECA2 was 1.11 Mb. The 62 microsatellites had on average 6.74 alleles, a mean HET of 0.639 and a mean PIC of 0.588.

The non-parametric multipoint linkage analysis showed chromosome-wide significant Zmeans and LOD scores for fetlock and hock OC at 20.08–30.94 Mb. For OCD in fetlock and hock joints, chromosome-wide significant linkage was reached at 26.89–29.47 Mb (Table S5). Chromosome-wide significantly linked markers for fetlock OC and fetlock OCD spanned from 15.65 to 30.94 Mb and 21.15–31.91 Mb. Genome-wide error probabilities for fetlock OC were reached at 20.08–30.94 Mb, with peak values at 23.80 Mb (Fig. 1, Table S6). For hock OC, chromosome-wide significant linkage was found at 26.89–33.05 Mb (Table S7).

The most frequent mutation of the SNPs was the C > T transition motif, with a frequency of 0.45. The rarest mutation was the A > T transversion motif, with a frequency of 0.05. Minor allele frequencies (MAF) were between 0.04 and 0.49. In total, 43 SNPs were determined to not be in the same linkage phase as another SNP by tagging with threshold $r^2 \geq 0.8$. Associations between OC-traits and SNPs within the genes NCDN, FCN3 and MECR were significant (Table S8). Significant marker-OC-trait associations could be shown for the NCDN-associated haplotype, whereas for the other candidate gene-associated haplotypes, associations were far from significance. The NCDN-associated SNP showed consistent significant effects for all OC- and OCD-traits. Using a logit model, including the genotype of the NCDN-associated SNP and age at examination, sex and season of birth (winter, spring vs. summer) as fixed effects and including a random sire effect, confirmed the significant influence ($P < 0.02–0.003$) of this SNP on OC-traits. Therefore, we conclude that NCDN may be a suitable functional and positional candidate gene for OC, even if we cannot rule out genes adjacent to NCDN as possible positional candidates. Neurochondrin (NCDN) is a negative regulator of calcium-/calmodulin-dependent protein kinase II phosphorylation in mice. The expression of this gene in chondral tissues was seen in the chondrocytes of growth plate and fibrocartilage cells (Ishiduka et al. 1999). In mice that were heterozygous mutants for a disrupted mutation in the NCDN gene, a disturbance of chondrocyte proliferation and differentiation in growth plates was observed, and based on these results, the NCDN

![Figure 1](http://example.com/figure1.png)
gene appears to be involved in chondrocyte differentiation, proliferation and survival (Mochizuki et al. 2003). The present linkage and association analyses also support a possible role of NCDN in equine OC. Therefore, further mutation and expression analyses of the equine NCDN gene should clarify the role of this gene in the aetopathogenesis of OC in horses.

Acknowledgements

The study was supported by grants of the German Research Council, DFG, Bonn (DI 333/12-2). The authors thank Heike Klippert-Hasberg and Stefan Neander for expert technical assistance.

Conflict of interest

The authors have not declared any potential conflicts.

References


Supporting information

Additional supporting information may be found in the online version of this article.

**Table S1** Number of families analysed, their sizes and prevalences of osteochondrosis (OC), osteochondrosis dissecans (OCD), osteochondrosis in fetlock (OC-F) and hock (OC-H) joints and OC dissecans in fetlock (OCD-F) and hock (OCD-H) joints by family and in total.

**Table S2** Characteristics of the microsatellites used in this study.

**Table S3** Location of 16 positional candidate genes on *Homo sapiens* chromosome (HSA) 1, their equine gene symbol and their position on horse chromosome (ECA) 2 of the horse genome assembly EquCab2.

**Table S4** Primer sequences, product size, annealing temperature and the corresponding location in the human gene, polymorphism information content, heterozygosity and for 65 SNPs (43 SNPs genotyped in this study), one indel mutation, RFLP information about nine intragenic single nucleotide polymorphisms.

**Table S5** Multipoint chromosome-wide test statistics with their chromosome-wide significant error probabilities (P2, P3) and map positions of microsatellites and SNPs linked with osteochondrosis (OC) and osteochondrosis dissecans (OCD) in Hanoverian warmblood horses.

**Table S6** Multipoint chromosome-wide test statistics with their genome-wide and chromosome-wide significant error probabilities (P2, P3) and map positions of microsatellites and SNPs linked with fetlock osteochondrosis (OC-F) and fetlock osteochondrosis dissecans (OCD-F) in Hanoverian warmblood horses.

**Table S7** Multipoint chromosome-wide test statistics with their chromosome-wide significant error probabilities (P2, P3) and map positions of microsatellites and SNPs linked with hock osteochondrosis (OC-H) in Hanoverian warmblood horses.

**Table S8** Number of affected and controls, odds ratios, minor allele frequencies and significant results of the association analyses of SNPs within candidate genes on horse
chromosome 2 with osteochondrosis in Hanoverian warmblood horses.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Interspersed repeats in the horse (*Equus caballus*); spatial correlations highlight conserved chromosomal domains

D. L. Adelson*, J. M. Raison†, M. Garber‡ and R. C. Edgar§

*School of Molecular and Biomedical Science, University of Adelaide, North Terrace, Adelaide, South Australia, Australia 5005. †eResearchSA, University of Adelaide, North Terrace, Adelaide, South Australia, Australia 5005. ‡Broad Institute, Boston, MA, USA. §Tiburon, CA, USA.

Summary

The interspersed repeat content of mammalian genomes has been best characterized in human, mouse and cow. In this study, we carried out de novo identification of repeated elements in the equine genome and identified previously unknown elements present at low copy number. The equine genome contains typical eutherian mammal repeats, but also has a significant number of hybrid repeats in addition to clade-specific Long Interspersed Nuclear Elements (LINE). *Equus caballus* clade specific LINE 1 (L1) repeats can be classified into approximately five subfamilies, three of which have undergone significant expansion. There are 1115 full-length copies of these equine L1, but of the 103 presumptive active copies, 93 fall within a single subfamily, indicating a rapid recent expansion of this subfamily. We also analysed both interspersed and simple sequence repeats (SSR) genome-wide, finding that some repeat classes are spatially correlated with each other as well as with G+C content and gene density. Based on these spatial correlations, we have confirmed that recently-described ancestral vs. clade-specific genome territories can be defined by their repeat content. The clade-specific Short Interspersed Nuclear Element correlations were scattered over the genome and appear to have been extensively remodelled. In contrast, territories enriched for ancestral repeats tended to be contiguous domains. To determine if the latter territories were evolutionarily conserved, we compared these results with a similar analysis of the human genome, and observed similar ancestral repeat enriched domains. These results indicate that ancestral, evolutionarily conserved mammalian genome territories can be identified on the basis of repeat content alone. Interspersed repeats of different ages appear to be analogous to geologic strata, allowing identification of ancient vs. newly remodelled regions of mammalian genomes.

Keywords equine, horse, mobile element, repeat, retroposon, transposon.

Introduction

Mammals vary widely in their appearance and physiology, yet are very similar based on comparisons of their genes. The core mammalian genome consists of approximately 20 000 protein-coding genes, with the vast majority conserved across species (Lander *et al.* 2001; Venter *et al.* 2001; Metzker *et al.* 2004; Lindblad-Toh *et al.* 2005). However, these protein-coding genes account for only about 1.5% of a typical mammalian genome. The rest of the genome is non-protein coding and, for the most part, is not transcribed [van Bakel *et al.* (2010)]. While there is still debate on how much of the genome is in fact transcribed, almost half of a typical mammalian genome is repetitive and was dubbed by some as ‘junk DNA’, much of which is derived from self-propagating mobile elements and retroviruses (Jurka *et al.* 2007). Interspersed repeats are the largest class of sequences in mammals, accounting for 40–50% of the total length of these genomes (Smit 1996; Lander *et al.* 2001; Waterston *et al.* 2002; Jurka *et al.* 2007). The most common interspersed repeats are derived from retro-transposons, also known as retroposons or retrotransposable elements (RTs), which replicate and jump throughout the genome in a manner similar to retroviruses.
While many RT are common to all mammals and are thus presumably of ancestral origin (Smit & Riggs 1995), every species/clade seems to have one or more unique kind of Short Interspersed Nuclear Element (SINE) which contributes heavily to species-specific genome sequences (Jurka et al. 2007). While many RTs are no longer active, species- and lineage-specific repeats serve to remodel genomes by interrupting and often outnumbering ancestral repeats during their phase of rapid transposition/expansion (Deininger et al. 2003; Kazazian 2004; Giordano et al. 2007). Active RTs are believed to be responsible for 10% of mutations in rodents (Kazazian 1998), while less active RTs in humans appear to account for a small fraction of new mutations (Deininger & Batzer 1999). The accumulation of RTs within or near genes has been studied (Birney et al. 2007), and there is evidence that insertions within or near promoters can alter gene expression, while insertions into exons are often incorporated into existing protein-coding genes (Krugl et al. 2007). Recently it has become clear that evolution has also made use of these repetitive sequences to wire new regulatory circuits (Mikkelsen et al. 2007). This has resulted from the incorporation of RTs into promoters, miRNA precursors and coding exons (Babushok et al. 2007; Gentles et al. 2007). Species-specific RTs can also contain regulatory elements such as the P53 tumour suppressor binding motif, and thus influence transcriptional regulatory networks genome-wide (Wang et al. 2007). Therefore, while the protein-coding genetic complement of mammals is >80% orthologous or homologous (Elsik et al. 2009), the remainder of these genomes is both highly repetitive and variable. It is therefore apparent that RTs are major drivers of genome evolution.

In mammals, LINE L1 repeats are the dominant RT type both in the common ancestor and in extant species (Lander et al. 2001; Waterston et al. 2002; Lindblad-Toh et al. 2005). Few mammals have active non-long terminal repeat (LTR) LINEs other than L1 that contribute significantly to repeat composition, with the exception of the LINE RTE (BovB) repeats in ruminants (Adelson et al. 2009) and marsupials (Gentles et al. 2007). Short Interspersed Nuclear Elements require LINEs for their transposition. In primates, LINE L1 repeats encode the machinery to transpose SINE Alu repeats (Dewannieux et al. 2003). Ancestral L2 LINEs, which were incapable of retrotransposition prior to the divergence of eutheria, are believed to have encoded the machinery to transpose SINE MIR (Jurka et al. 2007).

In this report, we describe de novo repeat identification, the generation of repeat consensus sequences and an analysis of the overall repeat content of the equine genome. We also show that there is evidence for spatial accumulation/segregation of repeats based on pairwise correlations of repeat abundance. These results confirm the existence of ancestral genomic territories on the basis of repeat content.

Materials and methods

De novo repeat identification and annotation

Equine genome assembly v.2.0 was used for repeat identification as described in (Adelson et al. 2009; Wade et al. 2009).

Identification of, and tree construction for, intact LINE elements

Coordinates for the intact L1 were retrieved from the equine genome assembly using PALS (Edgar & Myers 2005), with a minimum length of 90% of the query sequence, and a minimum of 70% identity. Sequences were globally aligned using MUSCLE (Edgar 2004) and the alignments used to create maximum likelihood trees using RAxML (Stamatakis 2006) with the GTRCAT substitution model, and an initial 500 bootstraps followed by a thorough maximum likelihood search. Putative active L1 were identified based on conserved ORF1 and ORF2 sequences.

Correlation analysis

This step was performed according to the procedure described in (Adelson et al. 2009). The SINEs ERE1 and 2 were amalgamated into a single group (ERE1_2) for the purpose of this analysis. This analysis was carried out by partitioning the genome into bins, counting the repeats in repeat group in each bin and calculating Spearman’s rank correlations for all pairs of repeat groups, and for the repeat groups and segmental duplications, gene count, and G+C content. As well as the 1.5 Mbp bins used in (Adelson et al. 2009), the analysis was repeated with 20, 50, 100, 150, 500 Kbp, 1, and 7.5 Mbp to see the impact of bin size on the analysis.

Identification of extreme density bins and repeat content analysis

The bins were classified as having low, medium or high MIR/L2 density. The cut-off between groups was the 2-tail 10% significance level cut-off for the sum of the MIR and L2 ranks. For the ERE1/2, ERE3 and L1 repeat groups, the statistical package R (The R Foundation for Statistical Computing, 2009) was used to perform Wilcoxon rank sum tests with continuity correction between the high and low-density groups. The high-density MIR/L2 bins for human (hg18) were obtained as for the equine, except that the RepeatMasker library for human was used. These human bins were mapped to equine bins based on full genomic alignments of repeat masked sequence from the horse genome to the human hg18, using PatternHunter 10 (Ma et al. 2002). Following established methods (Waterston et al. 2002; Lindblad-Toh et al. 2005), we identified co-linear clusters of the identified synteny anchors, which were used to form
larger syntenic segments in a hierarchical fashion. Segments that were larger than a given size in both genomes, and were comprised of at least four anchors at a given stage in the merging process, defined a resolution-dependent, pairwise synteny map between the two genomes.

**Results**

Our *ab initio* repeat identification and annotations (Table 1) indicate that the equine genome has a comparable repeat content to other eutheria (~47%), with LINE I (L1) RT the...
major class of interspersed repeats. The major SINE class in horse is the perrisodactyl specific ERE type (Gallagher et al. 1999). There are however many unclassified repeats that account for 11% of the equine genome. Most of these unclassified repeats are made up of a number of fragments of known types, and can best be described as possessing a chimeric or recombinant sequence. Only a minority of these unclassified sequences cannot be annotated using currently available RepeatMasker or RepBase repeat databases. The complexity of the chimeric repeats has proved to be resistant to a simple classification scheme, and will not be discussed further here.

We have characterized the intact L1s because they are the most common type of RT in the equine genome, and they are the only non-LTR LINEs that have the potential for autonomous retrotransposition. For L1s to have autonomous activity, they must be full length and encode two functional ORFs. We identified 1115 full length L1s by aligning our improved L1 consensus sequences against the Equine v2 assembly and extracting all full-length (≥90% or ≤110%) matching sequences that were ≥70% identical. These intact L1s were used to construct a maximum likelihood tree (Fig. 1). The tree topology reflects the five known L1 equine subfamilies (Smit et al. 1996–2004; Jurka et al. 2005). Active L1s should have two ORFs encoding 40- and 150-kDa proteins (Goodier et al. 2007). Only 103 of our full length L1s could be classified as potentially active on this basis, with 93 of those in the L1_1 subfamily and the remaining in the L1_2 subfamily.

In order to identify spatial correlations between repeats and between repeats and other genome features such as gene models, segmental duplications and G+C content, we carried out a comprehensive, pairwise correlation analysis of repeat types as performed by (Adelson et al. 2010). The results of this analysis are summarized in Fig. 2.

The effects of bin sizes on the correlations are shown in the right hand diagonal half of Fig. 2 and summarized in Table 2. In general, the use of a 1 Mbp bin size gave the strongest correlations for most pairs, with little additional strengthening at larger bin sizes. Most pairs (61%) had correlations that strengthened as a function of increasing bin size, indicating no bin size dependency for those associations. However, 34% of the pairs had a bin size-dependent response of correlation to bin size. This indicated that some associations appeared to be specific for certain scales/genomic distances. Only a few pairs showed their strongest correlations at a small scale (≤50 Kbp), but almost 10% of pairs had correlations that changed sign as a function of bin size. Some exemplars of these bin size effects are shown in Fig. 3. The scale dependence of some correlations suggests a potential biological effect, perhaps related to effects on gene regulation. Because most correlations were close to their maxima at 1.5 Mbp and had large enough sample sizes to keep the standard error manageable, we settled on this scale as the representative set of correlations shown in the left hand diagonal half of Fig. 2.

Based on these results we were able to identify a small number of very strong correlations, some of which we had observed previously in cow and human (Adelson et al. 2009). Specifically, we wish to draw attention to a very strong correlation between the fossil RT LINE 2 (L2) and SINE MIR (r > 0.8, see Fig. 4), and that between clade-specific SINE ERE1/2 and ERE3 sub-families. We also wish to draw attention to a very strong correlation between L1 and gene models, and a weaker correlation between L1 and segmental duplications. The strong pair-wise correlations for L2 and MIR and ERE1/2 and ERE3 were reminiscent of the relationships we had previously uncovered (Adelson et al. 2009; Elsik et al. 2009) in cow and suggested to us that similar spatial associations might exist in horse. We were particularly interested in the L2/MIR correlation, as these are an ancestral LINE/SINE pair, and their strongly correlated ranks indicated that they were being conserved in the same genomic regions. Similarly, the relationship between the clade-specific ERE1/2 and ERE3 indicated that these repeats appeared to favor integration in similar regions.

By plotting the ranks for L2 and MIR for every 1.5 Mbp bin (see Fig. 4), it is clear that not only is the correlation coefficient very high, but the distribution of ranks is consistent, symmetric and highly constrained. Because rank acts as a proxy for repeat count or density, we were able to select the highest and lowest repeat density bins for further study by partitioning the bins on the basis of the 10% tails.

![Figure 1](image-url) L1 phylogeny/active repeats. The maximum likelihood tree derived from the global alignment of all 1115 intact/full length L1 sequences. Red lozenges indicate the 103 putative active L1s, 93 of these were L1_1 and 10 were L1_2.
from an expected random distribution of bins. These partitions are shown by the lines in Fig. 4, with the 5% extreme higher density L2/MIR bins in the upper-right hand corner, and the 5% extreme lower density L2/MIR bins in the lower-left hand corner. Because each bin can be positioned on the chromosome scaffolds, it was then possible to determine the spatial distribution of the extreme density L2/MIR regions and to do the same for the extreme density clade-specific ERE1/2/ERE3 regions (Fig. 5).

The distribution of extreme-density ancestral repeats (L2/MIR) and clade-specific repeats were clearly different, with large contiguous regions of high-density ancestral repeats compared to much more evenly distributed bins for the other extreme-density bins. Furthermore, there was little overlap between the high-density ancestral regions and the high-density clade-specific bins. This was also evident from Table 3, where there were significantly fewer ERE3 RT found in high-density ancestral bins than in low-density ancestral bins. The same was true for L1, which are also of recent origin compared to L2 and MIR.

In order to determine if these ancestral repeat-rich regions were evolutionarily conserved, we repeated our correlation analysis on the human genome assembly (hg18) and found that L2/MIR were highly correlated in human (r = 0.86) as well. We identified the high-density L2/MIR bins for the human genome and then converted their coordinates to equine genome locations using our horse/human whole-genome alignment. Figure 5b shows the equine and human ancestral domains and reveals that they have a large degree of overlap that is particularly striking for the larger contiguous sets of bins that define large territories.

**Discussion**

Our novel methodology for identifying, annotating and analysing repetitive DNA has yielded a number of interesting results. While we did not identify many novel repeats, we did find a large number of hybrid or chimeric repeats. Chimeric repeats have been reported before (Buzdin et al. 2003), but at a much lower frequency than we have
observed. These repeats could represent novel, recombinant and evolving RT or they could be satellite sequences that have arisen from RT insertion (Kapitonov et al. 1998), or composite satellites amplified in a chromosome-specific manner. If they are composite satellites, we expect their distribution to be chromosome specific and peri-centromeric or sub-telomeric. While preliminary analyses indicate that some of these repeats are probably chromosome specific, it is too early to conclude that all of these chimeric repeats are composite satellites.

Our comprehensive spatial analysis of repeat groups has confirmed that ancient fossil L2 and MIR RT are highly correlated, as they are in cow and human (Adelson et al. 2009; Elsik et al. 2009). In this report we extended our spatial correlation analysis to investigate the effect of bin size on correlation. While most correlations strengthened as a function of bin size, for some, choice of bin size determined the sign and magnitude of the correlation. A specific example of this is the relationship between L1 and G+C content (Fig. 3). When the human and mouse genome sequences (Lander et al. 2001; Waterston et al. 2002) were analysed using 50 Kbp sequence windows, L1 were observed to be negatively correlated with G+C content. Our small-bin observations were consistent with these results, but at larger window/bin sizes the relationship changed. An arbitrary choice of window or bin size can therefore result in incomplete understanding of spatial relationships. Comparing SINE repeat groups correlations with G+C content in

Figure 3 Exemplars of correlation vs. bin size patterns. Various correlation responses to changing bin size are illustrated.
horse gave an opposite result to what was reported in mouse and human, where SINE were positively correlated with G+C content. Furthermore, our analysis of bovine repeat correlations (Adelson et al. 2009) showed that SINE families paired with LINE_RTE were negatively correlated with G+C, while one that was probably paired with L1 was positively correlated with G+C content.

We also noted positive correlations of L1 and LTR ERV1 with segmental duplications. Others have also reported associations of RT with segmental duplication/copy number variation (Bailey et al. 2003; She et al. 2008), but only the latter report implicated L1 RT.

Our analysis also confirmed that fossil RT densities (L2 and MIR) define conserved, syntenic ancestral genome domains. Because L2 and MIR have been inactive since the mammalian radiation, the persistence of such domains can only be explained by two alternate scenarios: (i) negative selection that preserved ancestral territories or (ii) protection from new retrotransposition events. There is evidence that many L2 and MIR RT have undergone strong negative selection because they have been co-opted to regulate gene expression (Silva et al. 2003; Lowe et al. 2007). This suggests that the conserved ancestral repeat-enriched genome territories we have discovered here are the result of
purifying selection or of chromatin structural constraints and are probably therefore of functional significance.

Author contributions

DLA: designed research, performed research, analysed data, wrote paper; JMR: performed research, analysed data, wrote paper; MG: performed research; RCE: contributed analytic tools, wrote paper.

Conflicts of interest

The authors have not declared any potential conflicts.

References


Pleiotropic effects of pigmentation genes in horses

R. R. Bellone
Department of Biology, University of Tampa, 401 W. Kennedy Blvd., Tampa, FL 33606, USA

Summary
Horses are valued for the beauty and variety of colouration and coat patterning. To date, eleven different genes have been characterized that contribute to the variation observed in the horse. Unfortunately, mutations involving pigmentation often lead to deleterious effects in other systems, some of which have been described in the horse. This review focuses on six such pleiotropic effects or associations with pigmentation genes. These include neurological defects (lethal white foal syndrome and lavender foal syndrome), hearing defects, eye disorders (congenital stationary night blindness and multiple congenital ocular anomalies), as well as horse-specific melanoma. The pigmentation phenotype, disorder phenotype, mode of inheritance, genetic or genomic methods utilized to identify the genes involved and, if known, the causative mutations, molecular interactions and other susceptibility loci are discussed. As our understanding of pigmentation in the horse increases, through the use of novel genomic tools, we are likely to unravel yet unknown pleiotropic effects and determine additional interactions between previously discovered loci.

Keywords coat colour, equine, pleiotropy, spotting patterns.

Introduction
Coat colour and spotting patterns have long fascinated animal breeders and geneticists alike. The pigmentation genes of mice were one of the first genetic systems to be elucidated and over 130 loci and 1000 mutations have been characterized (Steingrímsson et al. 2006). Pigmentation was one of the first systems to be studied, originally, because of the ease with which the traits could be identified as well as the variability in coat colour resulting from selective breeding by ‘mouse fanciers’ in the 19th century.

Similarly to the mouse fanciers of the 19th century, breeders selectively breed horses for their beauty and variety of coat colour and patterning (among other traits). While it has been suggested that a particular coat colour might be essential for survival and or have a selective advantage under certain environmental conditions (such as white horses being less attractive to the horse-fly), it is believed that most of the variation observed in breeds of horses today is a result of domestication and selective breeding (Ludwig et al. 2009; Horváth et al. 2010).

While the horse falls short of the 130 loci discovered in mice, currently as many as 31 loci have been postulated to be involved in coat colour variation and patterning in domestic horse breeds (Sponenberg 2009). With the utility of the equine genome sequence, many more are likely to be discovered in the near future (Wade et al. 2009). Seventeen different coat colour phenotypes have been investigated at the molecular level, and causative mutations have been identified and/or associations have been reported for eleven different genes (Table 1). Variation in horse coat colour is most often because of mutations that affect functioning of mature melanocytes. The base coat colour of the horse (black, bay, chestnut or seal brown) is determined by two loci (MC1R and ASIP) that control the switch between eumelanin and phaeomelanin production (Marklund et al. 1996; Rieder et al. 2001; Sponenberg 2009). Mutations in four additional genes cause dilutions in coat colouring, resulting in cream (SLC45A2), silver (SILV), champagne (SLC36A1), and pearl (SLC45A2) dilutes as well as lavender foal (MYO5A) (Mariat et al. 2003; Brunberg et al. 2006; Cook et al. 2008; Sponenberg 2009; Brooks et al. 2010). Patterning has been shown to be controlled by genes involved in the migration, proliferation and survival of melanocyte precursor cells, such as KIT (sabino-1, tobiano and dominant white) and endothelial receptor type B.
gene (EDNRB) (frame overo), and is also caused by mutations affecting melanocyte stem cell populations such as in the STX17 gene (grey) (Metallinos et al. 1998; Santschi et al. 1998; Yang et al. 1998; Brooks & Bailey 2005; Brooks et al. 2007; Haase et al. 2007; Pielberg et al. 2008). Other dilutions and patterning traits have been mapped, but the causative mutations remain unknown; these include dun, roan, leopard complex, and white face and leg markings (Table 1, Reviewed in Rieder 2009).

Melanocytes are derived from embryonic cells of neural crest origin. Neural crest-derived melanocytes are found in the skin, hair, certain layers of the eye (uveal melanosomes), the inner ear and leptomeninges. In addition, embryonic stem cells of neural crest origin also give rise to bone, cartilage, adipose, endocrine cells and several types of neurons and glia (Le Douarin & Kalcheim 1999). Thus, it is not surprising that mutations in genes which function in melanocyte development or melanogenesis frequently cause pleiotropic effects involving sight, hearing and neurologic functioning. Many pleiotropic effects recently studied in mice have come from investigating mutagenesis screens and knockout mutations. However, the broad variation in colour and patterning in and among horse breeds that has occurred as a result of domestication makes the horse an excellent model to unravel yet unknown causes of disease that are associated with pigmentation, several of which are currently being investigated. The coat colour mutations that have been reported to date were recently reviewed by Rieder (2009). Thus, this review focuses only on those pigmentation genes investigated in the horse, which either have known pleiotropic effects or have been associated with a disorder but are still under investigation.

### Table 1 Coat colour genes characterized in the horse.

<table>
<thead>
<tr>
<th>Coat colour phenotype</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chestnut</td>
<td>MC1R</td>
</tr>
<tr>
<td>Bay</td>
<td>MC1R and ASIP</td>
</tr>
<tr>
<td>Black</td>
<td>MC1R and ASIP</td>
</tr>
<tr>
<td>Seal brown</td>
<td>MC1R and ASIP</td>
</tr>
<tr>
<td>Cream</td>
<td>SLC45A2</td>
</tr>
<tr>
<td>Pearl</td>
<td>SLC45A2</td>
</tr>
<tr>
<td>Silver</td>
<td>SILV</td>
</tr>
<tr>
<td>Champagne</td>
<td>SLC36A1</td>
</tr>
<tr>
<td>Lavender foal</td>
<td>MYO5A</td>
</tr>
<tr>
<td>Sabino-1</td>
<td>KIT</td>
</tr>
<tr>
<td>Tobiano</td>
<td>KIT</td>
</tr>
<tr>
<td>Dominant white</td>
<td>KIT</td>
</tr>
<tr>
<td>Frame overo</td>
<td>EDNRB</td>
</tr>
<tr>
<td>Grey</td>
<td>STX17</td>
</tr>
<tr>
<td>Leopard complex</td>
<td>Associated with TRPM1</td>
</tr>
<tr>
<td>Roan</td>
<td>Associated with KIT</td>
</tr>
<tr>
<td>White face and leg markings</td>
<td>Associated with KIT, MC1R, and MITF</td>
</tr>
</tbody>
</table>

---

**Pleiotropic pigmentation genes**

Lethal white foal syndrome (LWFS) and the endothelin receptor type B gene (EDNRB)

Lethal white foal syndrome was the first pigmentation-related pleiotropic effect to be identified at the molecular level in the horse. Lethal white foal syndrome has been associated with the frame overo pattern (Fig. 1a) (McCabe et al. 1990). The term ‘frame’ comes from the phenotypic appearance of this spotting pattern in which the pigmentation ‘frames’ the horse. Therefore, the white spotting usually occurs in the middle of the sides of the flank, and neck as well as ventrally, but rarely dorsally, unless other pigmentation gene mutations are involved. Variation in the amount of white spotting does occur and can complicate proper classification. Frame horses often have partially depigmented irides involving either one or both eyes (blue or partially blue eyes). However, to date, no eye disorders have been associated with this phenotype.

The frame pattern is inherited as a dominant allele (Bowling 1994). As such, all frame horses are heterozygous for the mutation. However, the result of the homozygous condition is LWFS. This syndrome is characterized by foals born with a complete or near-complete white coat (Fig. 1b). These horses are also affected with intestinal aganglionosis, which is the absence of enteric ganglion cells (Hultgren 1982 and McCabe et al. 1990). Because of the lack of proper enervation of the intestine, foals develop intestinal obstruction within 24 h of birth, as food cannot be passed through the lower digestive tract. This obstruction quickly results in death. Three research groups independently identified the mutation responsible for this disease using a candidate gene approach (Metallinos et al. 1998; Santschi et al. 1998; and Yang et al. 1998). Mutations in the endothelin signalling pathways, involving both EDNRB and endothelin 3 (EDN3), cause similar phenotypes in humans (Hirschsprung disease) and in Ednrb and Edn3 knockout (KO) mice. In addition to aganlionosis, humans homozygous (but not heterozygous) for a mutation in EDNRB have hypopigmentation, and KO mice for EDNRB but not EDN3 are almost completely white (Hosoda et al. 1994; Puffenberger et al. 1994). Thus, EDNRB was chosen as the candidate gene to sequence in the horse. A dinucleotide missense mutation in the first exon of this gene was identified as the cause (NM_001081837.1:c.353_354delinsAG, resulting in an amino acid substitution of lysine for isoleucine (p.Ile118Lys) (Metallinos et al. 1998; Santschi et al. 1998; and Yang et al. 1998) (Table 2). The substitution to a charged amino acid is predicted to impact on the function of the first transmembrane domain of this seven transmembrane G-coupled receptor protein. The exact mechanism of how this substitution affects function, either in receptor localization or in signalling, remains to be determined.
Similarly to melanocytes, enteric ganglia cells are derived from the neural crest and thus endothelin signalling is important for the proliferation, migration and differentiation of both cell populations. While both melanocytes and enteric ganglia are derived from neural crest precursors, they migrate along two different pathways; cells destined to become enteric ganglia migrate along a dorso-ventral pathway while those destined to become melanocytes migrate along a dorso-lateral pathway (Pla & Larue 2003). Thus, the timing and expression of endothelin signalling appears to have several effects on these two cell types. EDN3/EDNRB signalling represses differentiation of early-stage enteric neural crest cells, thus allowing for migration of these cells and sensitivity to mitogenic effects by other molecules (Pla & Larue 2003). Later in development, EDNRB-mediated signalling is required after the establishment of enteric neuroblasts for correct innervations (McCallion & Chakravarti 2001). Similarly, in melanocytes, endothelin signalling plays a role in migration and terminal differentiation; however, dorso-lateral migration appears to begin later (Reid et al. 1996; Pla & Larue 2003).

While the precise intracellular signalling pathways for melanocytes that are activated by EDN3 binding to EDNRB are not known, signalling through Gq coupling of EDNRB, involving diacylglycerol, inositol triphoshpate and calcium messengers, has been shown to be involved in enteric glia development (Imamura et al. 2000). Determining the intracellular signalling cascade for melanocytes in both humans and horse would help resolve whether signalling through EDNRB is different in the two tissue types. Furthermore, if different intracellular signalling pathways are observed for horse and human, this could explain why, in the heterozygous condition, mutations in EDNRB cause white spotting phenotypes in horses but not in humans (Metallinos et al. 1998; McCallion & Chakravarti 2001).

As mentioned previously, variability in the amount of white exists for frame horses. Furthermore, in a study conducted by Santschi et al. 2001, it was shown that the association with p.Ile118Lys and the frame phenotype was not 100%; non-frame horses had the mutation, and horses with a phenotype of frame had a genotype of p.118Ile (wild type). Thus, it is likely that this mutation is not fully penetrant in the heterozygous condition and can be masked by modifier genes (Santschi et al. 2001). Additionally, it is also likely that mutations in other genes may cause a similar frame phenotype. Based on work in mice and humans, a sex effect and the tyrosine kinase receptor KIT, its ligand KITLG and transcription factors SOX10 and PAX3 may be important modifier genes for further exploration (McCallion & Chakravarti 2001). Ten per cent of the tobiano horses (a white spotting pattern different from frame overo in which the white crosses the top line) tested were heterozygous for the p.Ile118Lys mutation, but did not display a recognizable overo pattern (Santschi et al. 2001). This

Figure 1 Pigmentation phenotypes in the horse associated with pleiotropic effects. (a) Frame overo, which in the homozygous condition causes lethal white foal syndrome. (b) Lethal white foal resulting from a homozygous mutation in EDNRB. (c) Splash white is often associated with deafness; however, the genetic mutation has not been determined. (d) Lavender foal caused by a homozygous mutation in MYO5A. (e) Flea bitten grey. (f) Dapple grey. (g) Grey horse with melanoma in atypical area. (h) ‘Leopard’ pattern, which is heterozygous for LP and unaffected for congenital stationary night blindness (CSNB). (i) ‘Few spot’ pattern, which is homozygous for LP and been associated with CSNB, and is likely caused by a mutation in TRPM1. (j) Silver dapple caused by a mutation in PMEL17 and associated with multiple congenital ocular anomalies.

© 2010 The Author, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 100–110
strongly suggests that KIT is likely a modifier gene, and this interaction should be explored in the horse.

**Deafness and the endothelin receptor type B gene (EDNRB)**

Humans homozygous for a mutation in EDNRB with a specific form of Hirschsprung disease, known as Shah–Wardenburg syndrome (WS4), also have sensorineural deafness and bicoloured irides, in addition to aganglionosis and hypopigmentation (Puffenberger et al. 1994). A 318 bp deletion in EDNRB was developed as a mouse model for WS4. This mouse also had pigmentation anomalies, deafness, and megacolon (Matsushima et al. 2002). The neurosensory deafness associated with WS4 in the mouse model and humans is believed to be caused by the lack of melanocytes in the inner ear. These melanocytes are thought to be required for the formation and integrity of the stria vascularis and maintenance of the endocochlear potential (Pla & Larue 2003). White spotting has also been associated with deafness in dogs and cats, but EDNRB has not yet been associated (Geigy et al. 2007; Strain et al. 2009).

Prior to the identification of the mutation for LWFS, deafness was anecdotally associated with this disorder. However, it was not known whether all lethal white foals were affected (McCabe et al. 1990). Recently, a study by Magdesian and colleagues investigated deafness in American Paint Horses, including those with the NM_001081837.1:c.353_354delinsAG mutation in EDNRB (Magdesian et al. 2009). Among those horses that were confirmed deaf by brainstem auditory-evoked responses, and suspected deaf by their owners, most had the coat patterning of either splashed white (causative gene mutation not yet known, Fig. 1c, Table 2) or framesplashed white blend. Three lethal white foals were

<table>
<thead>
<tr>
<th>Coat colour or pattern</th>
<th>Gene</th>
<th>Locus</th>
<th>DNA mutation</th>
<th>Predicted function of mutation</th>
<th>Pleiotropic effects observed</th>
<th>Gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame overo pattern</td>
<td>EDNRB</td>
<td>ECA17</td>
<td>NM_001081837.1:c.353_354delinsAG</td>
<td>Amino acid substitution (p.Ile118Lys) disrupting function</td>
<td>Lethal white foal syndrome in homozygotes (absence of enteric ganglion cells); deafness susceptibility</td>
<td>Known to affect proliferation, migration, and differentiation of neural crest-derived melanocytes and enteric ganglia cells.</td>
</tr>
<tr>
<td>Splash white pattern</td>
<td>Unk</td>
<td>Unk</td>
<td>Unk</td>
<td>Unk</td>
<td>Unk</td>
<td>Unk</td>
</tr>
<tr>
<td>Lavender coat dilution</td>
<td>MYO5A</td>
<td>ECA1</td>
<td>ECA1g, 138235715del</td>
<td>Frame shift resulting in a truncated protein</td>
<td>Homozygotes have neurological symptoms leading to death Melanoma in which homozygotes are more susceptible</td>
<td>Proposed to disrupt trafficking of melanosome in melanocytes and synapse molecules in neurons. Proposed to promote melanocyte proliferation, causing melanoma in the dermis and depletion of stem cell melanocyte populations associated with hair follicles.</td>
</tr>
<tr>
<td>Grey pattern</td>
<td>STX17</td>
<td>ECA25</td>
<td>ECA25g, 6575277_657982dup</td>
<td>cis-acting regulatory element that upregulates both STX17 and NR4A3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leopard complex pattern</td>
<td>TRPM1</td>
<td>ECA1</td>
<td>Unk</td>
<td>Downregulate TRPM1</td>
<td>Homozygotes are affected with congenital stationary night blindness</td>
<td>Shown to be the selective cation transduction channel in the ON bipolar cell pathway for vision in low light conditions. Also proposed to play a role in the storage of melanin and potentially in melanocyte migration.</td>
</tr>
<tr>
<td>Silver dilution</td>
<td>SILV</td>
<td>ECA6</td>
<td>DQ665301:g.1457C&gt;T</td>
<td>Amino acid substitution (p.Arg618Cys) disrupting function</td>
<td>Association of silver with multiple congenital ocular anomalies</td>
<td>Shown to be involved in biogenesis of premelanosome and perhaps specifically the eumelanosome. Work in dogs suggests functional role in eye development</td>
</tr>
</tbody>
</table>

Unk, unknown.

Table 2 An overview of the genes involved in pigmentation and their pleiotropic effects studied in the horse.

© 2010 The Author, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 100-110
confirmed deaf prior to death. Furthermore, 91% of the horses confirmed or suspected to be deaf had the NM_001081837.1:c. 353_354delinsAG mutation (Magdesian et al. 2009). This association suggests that EDNRB may play a role in deafness; however, similar to the frame coat patterning, other genes are likely to be involved. One of these is probably the gene and/or genes responsible for splash white. Splash white has not been molecularly characterized in the horse. Unravelling the genetic mechanism of splash will likely help elucidate the interaction of splash and frame patterning and the association of deafness with these phenotypes. Additionally, cases of WS4 in humans have been shown to also be caused by mutations in EDN3 and SOX10; thus, these candidate genes should be investigated for their association with deafness in the horse (Price & Fisher 2001). Furthermore, while all of the lethal white foals in Magdesian study were deaf, the sample size was small (N = 3). Thus, it cannot be concluded that this is the case for all lethal white foals.

Lavender foal syndrome (LFS) and the myosin VA gene (MYO5A)

The most recently elucidated pleiotropic effect involves LFS, also known as coat colour dilution lethal. Lavender foal syndrome is inherited as a lethal recessive disorder in the Arabian breed, which is primarily of Egyptian descent. This disease was first described by Bowling as light-coloured foals that failed to nurse and died shortly after birth (Bowling 1996). The coats of these animals are often described as a lavender-like colour, and thus the disorder was named for this unique characteristic (Fig. 1d). However, the coat can range from pale grey to light chestnut. The syndrome is characterized by several neurological signs involving tetany (involuntary muscle contractions), opisthotonus (hypermobility of the head and neck), nystagmus (involuntary eye movements), and paddling leg movement (Bowling 1996; Fanelli 2005).

The genetic mutation causing this disorder has very recently been discovered by an SNP-based whole-genome association approach (Brooks et al. 2010). The trait mapped to a 10.5 Mb region on ECA1 containing the candidate genes myosin VA (MYO5A) and ras-associated protein RAB27A (RAB27A), which cause similar disorders in mice and humans (Griscelli syndrome). In addition to neurological abnormalities, mutations in RAB27A often cause immunological disruptions. Because such disruptions are not detected in LFS, MYO5A was thus chosen as the candidate gene to investigate further. A single-base deletion in exon 30 (ECA1 g.138235715del) is suspected to cause a frame shift leading to a premature stop codon in a highly conserved region of the gene (Brooks et al. 2010) (Table 2). Thus, the authors have proposed that this prematurely terminated protein would not be able to effectively bind cargo for intracellular transport. In the case of melanocytes, unlike those mutations that disrupt melanocyte migration and differentiation (such as EDNRB described above), this deletion would alter the function of a mature melanocyte, in that melanosome trafficking to the periphery of the cell for transfer to the keratinocyte would be disrupted (Marks & Seabra 2001; Brooks et al. 2010). Likewise, in nerve cells, glutamate receptors and secretory granules would not be transported properly, and thus this could explain the various neurological defects observed (Goda 2008; Brooks et al. 2010). Additional work is needed to confirm the underlying biochemical mechanisms. Furthermore, while a dilution in pigmentation has not been observed in heterozygotes with this deletion (Samantha Brooks personal communication), there is speculation that LFS carriers may have a mild survivable epileptic condition (Fanelli 2005). Future studies will likely investigate pigment density and epileptic condition in heterozygotes.

Melanomas and the syntaxin 17 gene (STX17)

Previous studies have reported that approximately 80% of horses with the grey coat pattern will develop melanomas (Valentine 1995; Sutton & Coleman 1997; Johnson 1998). Grey is a progressive coat colour phenotype that occurs in many breeds of horses, and in some breeds, such as Lipizzaner and Andalusian, it is the predominant colour. Grey horses progressively acquire white hairs throughout the coat as they age, and the greying process (rate and location) varies from horse to horse. Some horses will grey first in the mane and tail hairs while others will lose pigment last in the mane and tail (Sponenberg 2009). Additionally, some horses will retain small flecks of pigmented hairs to display a ‘flea-bitten’ appearance (Fig. 1e), while others will retain pigmented hairs that outline areas of depigmented hairs to display a ‘dappled’ appearance (Fig. 1f). Furthermore, some grey horses will develop skin depigmentation (similar to vitiligo in humans) in addition to hair depigmentation (Sutton & Coleman 1997). It is speculated that some of the horse to horse variation in the greying process is likely due to modifier genes that have not yet been investigated. ‘Grey’ is said to be epistatic to all other coat colours and patterns, as horses that inherit the grey duplication will eventually lose pigment as they age, and any underlying coat colour or pattern will be masked by ‘grey’.

‘Grey’ is inherited as dominant gene. The ‘grey’ condition was mapped to ECA25 by three independent groups (Hanner et al. 2002; Locke et al. 2002; Swinburne et al. 2002) (Table 2). The status of the horse genome at that time did not allow for the identification of candidate genes from that region. Therefore, Pielberg and colleagues identified SNPs in genes on ECA25 to refine the map position. They successfully defined a region that corresponds to approximately 6.9 Mb on human chromosome 9q (Pielberg et al. 2005). However, again, no obvious candidate genes causing either
pigmentation defects or melanoma susceptibility were detected in this region. The availability of the horse genome and SNPs detected during the genome sequencing effort made the further refinement and candidate gene investigation possible. Eighteen SNPs in the 6.9-Mb region were used to define a 350-kb critical interval, and four genes were investigated from this region, none of which had been previously implicated in pigmentation defects or melanoma susceptibility. Pielberg et al. (2008) identified a 4.6-kb duplication in intron 6 of syntaxin 17 (STX17) as the cause of grey (ECA25_6575277_6579862dup, as is identifiable with GenBank accession numbers EU606026 and EU606027). The authors proposed that this duplication is a cis-acting regulator mutation that upregulates both STX17 and nuclear receptor subfamily 4, group A, member 3 (NR4A3), another gene in the critical interval. Upregulation of STX17 and/or NR4A3 is thought to promote melanocyte proliferation, which in dermal melanocytes leads to predisposition to melanoma development and in hair follicle melanocytes leads to hyperproliferation and a depletion of stem cells. This depletion of the stem cell pool causes the hair to grey as new hairs replace those that are lost. While it is still likely that variation in the greying process and incidence of melanomas is determined by unknown modifiers, Pielberg and coauthors showed that some of this variability is because of the incomplete dominant nature of this mutation. Horses homozygous for the duplication greyed at a faster rate, had more skin depigmentation and less speckling in the coat and also had a higher incidence of melanoma (Pielberg et al. 2008).

The melanomas that appear on horses typically occur as black-pigmented nodules in the dermis of hairless skin, usually under the tail, perianal and genital regions, as well as around the lips and eyelids, although they can occur in other areas (Fig. 1g) (Seltenhammer et al. 2004). Two cases of neoplasms of the vertebral column in grey horses have also been reported (Schott et al. 1990). Although melanomas in grey horses show less malignancy to those of solid-coloured horses (Seltenhammer et al. 2003), metastases do occur in grey horses, and the most common sites are in the lymph nodes, liver, spleen, skeletal muscle, lungs, and surrounding or within blood vessels (MacGillivray et al. 2002). Genetic mechanisms controlling metastasis in the horse has not been characterized. However, melanoma susceptibility is increased in grey horses that also have a loss of function mutation in agouti signaling protein (ASIP) (Pielberg et al. 2008). ASIP is an antagonist of the melanocortin 1 receptor (MCIR) and is involved in the switch from eumelanogenesis to phaeomelanogenesis. This suggests that in the horse, increased MCIR signalling influences melanoma development. This is in contrast to what has been shown in humans, in which loss of function polymorphisms in MCIR and not ASIP contribute to melanoma susceptibility (Fargnoli et al. 2006; Stratigos et al. 2006; Brudnik et al. 2009). Although loss of function mutations in MCIR and melanoma susceptibility have not yet been studied in the horse, studying these interactions may help to more clearly discern the differences in the biology and metastasis of equine melanoma from that of human. Additionally, other genes in humans have been shown to play a role in melanoma susceptibility. For example, one polymorphism (NP_000266.2:p.Arg419Glu) in OCA2 is associated with an increased risk of developing malignant melanoma (Fernandez et al. 2009). In the horse, an SNP in the coding region of OCA2 (DQ454071.1:c.346A>G) has been identified, but no association with coat colour phenotype has been determined (Bellone et al. 2006a). Furthermore, this SNP and others in OCA2 have not yet been investigated in the horse for melanoma susceptibility. Similarly, an amino acid substitution in MYO7A (NP_000251.3:p.Ser1666Cys) was found to be associated with melanoma risk in humans; this gene has not been investigated in the horse as either a cause of a coat colour phenotype or a melanoma risk factor (Fernandez et al. 2009).

Congenital stationary night blindness (CSNB) and the transient receptor potential cation channel, subfamily M, member 1 gene (TRPM1)

Congenital stationary night blindness has been associated with homozygosity for leopard complex spotting (LP, also known as appaloosa spotting) in the Appaloosa breed (Sandmeyer et al. 2007). Leopard complex spotting is characterized by patterns of white in the coat that tend to be symmetrical and centred over the hips (Spomember et al. 2009). The extent of white patterning varies widely among individuals, and, like the variation already discussed with frame patterning and grey, this variation is in part because of modifier genes (Miller 1965; Spomember et al. 1990; S. Archer and R. R. Bellone, unpublished data). The term ‘leopard’ is derived from one of these patterns in which oval spots of pigment are found in the pattern of white extending over most of the body (Fig. 1h). In addition to the patterning in the coat, LP is associated with four other pigmentation traits: striped hooves, readily visible non-pigmented sclera around the eye, mottled pigmentation around the anus, genitalia, and muzzle, and LP-specific roaning (Spomember et al. 2009). Horses that are homozygous for leopard complex spotting (LP/LP) tend to have fewer pigmented spots than heterozygotes in the white patterned areas (Spomember et al. 1990; Lapp & Carr 1998) (Fig. 1i). In addition to the Appaloosa breed, many other breeds have leopard complex spotting (Knabstrupper, Noriker, Pony of the Americas, American Miniature, British Spotted Pony, and Australian Spotted Pony, among others); however, association of CSNB with homozygosity for LP has not yet been documented in these other breeds.

LP is inherited as an incompletely dominant gene that was mapped to a 6-cM region on ECA1 (Terry et al. 2004).
The positional and functional candidate *TRPM1* was implicated as the genetic cause (Bellone et al. 2006b, 2008, 2010a,b) (Table 2). Similarly to STX17 and NR4A, the role of TRPM1 in pigment production has not been elucidated. However, it was shown that TRPM1 is downregulated in highly metastatic melanoma cells, suggesting that this protein plays an important role in maintaining normal melanogenesis (Duncan et al. 1998). The extent of TRPM1 involvement in horse melanoma has not been determined. Most recently, work in humans has demonstrated that TRPM1 expression correlates directly with melanin concentration, suggesting a potential role for TRPM1 in the storage of melanin (Oancea et al. 2009). TRPM1 belongs to the Ca\(^{2+}\) transient receptor potential superfamily. Ca\(^{2+}\) signalling and sensation have obvious roles in both cell migration and signalling. TRPM1 therefore may also play a role in melanocyte migration, but this remains to be determined.

Congenital stationary night blindness is characterized by a congenital and non-progressive scotopic (low light condition) visual deficit and was first characterized in an Appaloosa horse in 1977 (Witzel et al. 1977). Affected animals occasionally manifest a bilateral dorso-medial strabismus (improper eye alignment) and nystagmus (involuntary eye movement) (Sandmeyer et al. 2007). Congenital stationary night blindness is diagnosed by a ‘negative ERG’, which is a dark-adapted electroretinography in which the b-wave is absent and there is a depolarizing a-wave (Witzel et al. 1977). This is similar to the Schubert–Bornshein type of human CSNB (Schubert & Bornshein 1952; Witzel et al. 1978). The ‘negative ERG’ is indicative of a defect in depolarizing the ON bipolar cells (the next cells involved in night vision after the rod photoreceptors). The synapse between the ON bipolar cell and the photoreceptor, in causing the depolarizing event, involves the binding of glutamate to its receptor, metabotropic glutamate receptor (MGLuR6), which couples to the closure of a cation-selective transduction channel (Nomura et al. 1994; Nakanishi et al. 1998). Until very recently, the cation channel involved in ON bipolar cell signalling was not determined. Differential gene expression of *TRPM1* in CSNB-affected Appaloosa horses provided evidence that this gene was the cation-selective channel (Bellone et al. 2008). Work by Shen et al. further supports TRPM1 as the cation channel, as TRPM1 knockout mice exhibit a similar ‘negative ERG’ (Shen et al. 2009). Furthermore, very recently several different mutations in TRPM1 have been shown to cause CSNB in humans (Audo et al. 2009; van Genderen et al. 2009; Li et al. 2009).

qRT-PCR analyses showed that *TRPM1* mRNA expression is significantly downregulated in both the skin of homozygotes (LP/LP) (downregulated by about 200-fold) and the retina of CSNB-affected Appaloosas (downregulated by over 1800-fold), whereas four other linked genes had unaltered mRNA expression (Bellone et al. 2008). This suggested that *TRPM1* was involved in both pigmentation and CSNB. The putative coding region of this gene was investigated for mutations that could explain this difference in expression; none were identified (Bellone et al. 2010a). Thus, to localize the causative mutation, 70 SNPs spanning over 2 Mb encompassing the *TRPM1* gene were utilized, and a single 173-kb haplotype associated with LP and CSNB (ECA1: 108 197 355–108 370 150) was identified. Illumina resequencing of 300 kb surrounding this haplotype identified six candidate SNP variants for further investigation, which is currently underway and is described in this issue (Bellone et al. 2010a,b). None of the causative mutations in *TRPM1* in humans have been associated with pigmentation differences in the skin. Determining the exact mechanism of LP and CSNB in the horse may support melanogenesis processes specific to horse.

**Multiple congenital ocular anomalies (MCOA) and the silver homolog gene (SILV)**

The association of MCOA with the silver colouration in the horse was first documented in 1999 by Ramsey and colleagues. These authors noted a high proportion of affected Rocky Mountain horses who possessed a white mane and tail and dark chocolate body colour (Ramsey et al. 1999). While the precise association of MCOA with the silver phenotype is still not clear, the mutation in *SILV* causing the silver coat colour has since been determined.

The silver colouration occurs frequently in many breeds, including the Rocky Mountain Horse and closely related breeds Kentucky Mountain Saddle Horse, and Mountain Pleasure Horse as well as the Icelandic horse and American Miniature horse, among others. The prevalence of MCOA and its association with silver colouration have been examined in purebred Rocky Mountain horses in the United States as well as purebred and cross-bred Rocky Mountain Horses in Canada (Ramsey et al. 1999; Grahn et al. 2008). The association of MCOA with silver in other breeds such as Icelandic horse and the American Miniature has not yet been reported.

Horses with the silver coat colour dilution (also known as silver dapple) are most often associated with a diluted body coat that is ‘silver’ or ‘chocolate’ in appearance, as well as a white mane and tail (Fig. 1)). In addition, horses with silver colouring frequently have dappled areas in their coat, with darker pigment outlining lighter areas, and hence the term ‘silver dapple’ is frequently used to denote this coat colour. The presence of the silver dapple allele is undetectable in horses that are chestnut in colouration, as this mutation only dilutes eumelanin and not phaeomelanin.

‘Silver’ is inherited as a dominant allele. Similar pigmentation phenotypes in mice, chicken, dogs, zebrafish and humans suggested SILV (silver homolog also referred to as *pre-melanosomal protein 17* or *PMEL17*) as a plausible genetic cause of silver in the horse. *SILV* was mapped to ECA6 (Rieder et al. 2000) and subsequently sequenced by
two separate groups. Two single-base substitutions were reported to be in complete association with the silver colouration, one within intron 9 and the other in exon 11 (Brunberg et al. 2006; Reissmann et al. 2007). The exonic SNP causes a missense mutation at the fifth base of the exon (DQ665301:g.1457C>T), which results in substitution of arginine for a cysteine in the cytoplasmic region of the protein (p.Arg618Cys), and it is thus suspected to be the causative mutation (Brunberg et al. 2006). This claim is well supported, as chickens that are dilute in colour (dun) have the same amino acid substitution. As its alias name suggests, PMEL17 is thought to be involved in the biogenesis of the premelanosome (Yasumoto et al. 2004; Hoashi et al. 2005). Given that phaeomelanin pigmentation is not disrupted by a mutation in SILV, as evidenced in horse, it is likely that SILV is involved specifically with eumelanosome formation.

Equine MCOA is characterized by a diverse set of ocular phenotypes. The predominant phenotype is large cysts (fluid-filled sacs) that originate from the temporal ciliary body or peripheral retina, and additional phenotypes include abnormalities of the cornea, iris, lens and iridocorneal angle (Ewart et al. 2000; Grahn et al. 2008). Similar to the silver mutation, MCOA has been documented to be controlled by a dominant gene, with some reports demonstrating a co-dominant mode of inheritance and incomplete penetrance (Ewart et al. 2000; Andersson et al. 2008; Grahn et al. 2008). Homozygotes are thought to be more severely affected, having multiple abnormalities, while heterozygotes have cysts only (Andersson et al. 2008), although this may not always be the case (Grahn et al. 2008).

Incomplete penetrance of this disorder has made studying the molecular mechanism behind these eye phenotypes difficult. Individuals carrying the causative mutation that are phenotyped as normal may either have cysts that were too small to detect or be true cases of non-penetrance (Andersson et al. 2008; Grahn et al. 2008). Multiple congenital ocular anomalies has recently been mapped to a 4.9-Mb interval on ECA 6 containing SILV (ECA6: 70 589 360–75 475 262). Re-sequencing this gene in horses diagnosed with MCOA did not lead to the identification of any additional mutations; thus, because SILV is in the critical interval, this gene and specifically the p.Arg618Cys substitution cannot be ruled out as the cause. However, Grahn and colleagues reported exceptions that could indicate linkage and not causality. Specifically, there was one silver offspring who upon ocular examination was described as having a normal phenotype (Grahn et al. 2008). Non-penetrance may explain this particular case, and thus further investigation is needed before SILV can be definitely ruled out as the cause. In the dog, a SINE insertion in SILV causes the merle phenotype; merle-patterned dogs have various ophthalmic disorders including microphthalmia (abnormally small eyes), cataracts and colobomas (hole

in one of the structures of the eye) (Dausch et al. 1978; Gelatt et al. 1981). Thus, it is possible that either a mutation in SILV or a cis-acting regulatory molecule effecting PMEL17 may be involved in MCOA. Investigating the eyes of silver horses from other breeds, such as the Icelandic horse, could add to our understanding of the interaction of SILV with MCOA. Moreover, dogs with the SINE insertion have a higher prevalence of deafness, and homozygotes are more significantly affected (Strain et al. 2009). Unlike EDNRB, SILV and/or MCOA association with deafness has not yet been investigated in the horse.

Conclusion

Pleiotropic effects in the horse related to pigmentation include lethal nervous system defects (LWFS and LFS), melanomas (grey pattern phenotype), eye and vision abnormalities (MCOA and CSNB), and well as deafness (frame overo patterning and potentially splash patterning). It is interesting to note that in all cases reported, the homozygotes have a more severe effect. In LWFS, LFS and CSNB, only homozygotes are affected with the disorder, whereas in the case of melanomas and MCOA, it appears that homozygotes are more severely affected. As demonstrated by LFS, the availability of new molecular tools (such as horse genome sequence and 56K SNP chip) has aided the speed at which Mendelian traits can be unravelled. Thus, it is likely that many additional equine-specific pigmentation mutations and pleiotropic effects will be discovered shortly. Of considerable interest are those genes involved in modifying colour and pattern, as they will likely add to our understanding of susceptibility to several of the pleiotropic effects, such as melanoma and deafness, that are discussed in this review. Additionally, as we learn more about these pigmentation genes, we will understand more about the biochemical pathways connecting these genes, such as TRPM1 involvement in melanomas in the horse. Furthermore, as we gain more insight into pleiotropic effects in the horse, we are likely to better understand the differences in melanogenesis associated with the hair follicles and those associated with the keratinocytes, as demonstrated by the grey condition. We will also learn about horse-specific pigmentation differences. Because of the nature and biochemical function of pigmentation genes, it is also likely that additional pleiotropic effects, such as neuroendocrine or immunological, will be observed.

Acknowledgements

I graciously acknowledge the fruitful discussions with many colleagues concerning various coat colour traits. I specifically acknowledge three of my collaborators, Samantha Brooks, Sheila Archer and Phil Spoenenberg, who provided photographs and helpful comments during the preparation of this manuscript. In addition, I acknowledge Dr Tamara
Lightbody, who provided the lethal white foal picture, and Dr Yael Grinboim, for providing the lavender foal picture.

Conflicts of interest
The author has declared no potential conflicts.

References


Bellone R., Lear T., Adelson D.L. & Bailey E. (2006b) Comparative mapping of ocuolucataneous albinism type II (OCA2), transient receptor potential cation channel subfamily M member 1 (TRPM1) and two equine microsatellites, ASB08 and 1CA43, among four equid species by fluorescence in situ hybridization. *Cytogenetic and Genome Research* 114, 93A.


Bellone R.R., Archer S., Wude C.M., Cuka-Lawson C., Haase B., Leeb T., Forsyth G., Sandmeyer L. & Grahn B. (2010b) Association analysis of candidate SNPs in TRPM1 with leopard complex spotting (LP) and congenital stationary night blindness (CSNB) in horses. *Animal Genetics* 41, 207.


van Genderen M.M., Bijveld M.M.C., Claassen Y.B. et al. (2009) Mutations in TRPM1 are a common cause of complete congenital stationary night blindness. *American Journal of Human Genetics* 85, 730–6.


night blindness in the Appaloosa and the association with the leopard complex. *Veterinary Ophthalmology* 10, 368–75.


Genome-wide association analysis of osteochondrosis of the tibiotarsal joint in Norwegian Standardbred trotters

S. Lykkjen*, N. I. Dolvik†, M. E. McCue‡, A. K. Rendahl†,§, J. R. Mickelson¶ and K. H. Roed*

1The Norwegian School of Veterinary Science, Department of Basic Sciences and Aquatic Medicine, Post-box 8146 Dep., N-0033 Oslo, Norway. 2The Norwegian School of Veterinary Science, Department of Companion Animal Clinical Sciences, Section for Equine Medicine and Surgery, Oslo, Norway. 3Department of Veterinary Population Medicine, University of Minnesota, St Paul, MN, USA. 4School of Statistics, University of Minnesota, Minneapolis, MN, USA. 5Department of Veterinary and Biomedical Sciences, University of Minnesota, St Paul, MN, USA.

Summary

Osteochondrosis (OC), a disturbance in the process of endochondral ossification, is by far the most important equine developmental orthopaedic disease and is also common in other domestic animals and humans. The purpose of this study was to identify quantitative trait loci (QTL) associated with osteochondrosis dissecans (OCD) at the intermediate ridge of the distal tibia in Norwegian Standardbred (SB) using the Illumina Equine SNP50 BeadChip whole-genome single-nucleotide polymorphism (SNP) assay. Radiographic data and blood samples were obtained from 464 SB yearlings. Based on the radiographic examination, 162 horses were selected for genotyping; 80 of these were cases with an OCD at the intermediate ridge of the distal tibia, and 82 were controls without any developmental lesions in the joints examined. Genotyped horses descended from 22 sires, and the number of horses in each half-sib group ranged from 3 to 14. The population structure necessitated statistical correction for stratification. When conducting a case–control genome-wide association study (GWAS), mixed-model analyses displayed regions on chromosomes (Equus callabus chromosome – ECA) 5, 10, 27 and 28 that showed moderate evidence of association (P £ 5 · 10^{-5}; this P-value is uncorrected i.e. not adjusted for multiple comparisons) with OCD in the tibiotarsal joint. Two SNPs on ECA10 represent the most significant hits (uncorrected P = 1.19 · 10^{-5} in the mixed-model). In the basic association (chi-square) test, these SNPs achieved statistical significance with the Bonferroni correction (P = 0.038) and were close in the permuted logistic regression test (P = 0.054). Putative QTL on ECA 5, 10, 27 and 28 represent interesting areas for future research, validation studies and fine mapping of candidate regions. Results presented here represent the first GWAS of OC in horses using the recently released Illumina Equine SNP50 BeadChip.

Keywords equine, Equine SNP50, Genome-wide association, osteochondrosis, quantitative trait loci.

Introduction

Osteochondrosis (OC) is a common and clinically important disease that affects developing joints in horses as well as cattle, pigs, dogs, poultry and humans (Olsson & Reiland 1978). OC is defined as a disturbance in the process of endochondral ossification (Rejnö & Strömberg 1978). At the epiphyseal growth cartilage, the disturbance can lead to the formation of partially or completely detached fragments (osteochondrosis dissecans (OCD), fissures, or subchondral bone cysts (Rejnö & Strömberg 1978; Ytrehus et al. 2007). Lesions occur at predilection sites that are specific to the species and joint in question (Ytrehus et al. 2007). By causing joint inflammation and osteoarthritis, OC is a frequent cause of pain, lameness and reduced performance in young athletic horses (McIlwraith 1993). dogs (Harrairi 1998) and humans (Schenck & Goodnight 1996). In swine, OC is regarded as the most important cause of leg weakness (Jørgensen et al. 1995).

Address for correspondence

S. Lykkjen, Department of Basic Sciences and Aquatic Medicine, Norwegian School of Veterinary Science, Post-box 8146 Dep., N-0033 Oslo, Norway.
E-mail: sigrid.lykkjen@nvh.no
Accepted for publication 21 September 2010

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 111–120
Although intensively researched, the aetiopathogenesis of OC is not fully understood. Opinions have differed as to whether the disease is caused by trauma (Pool 1993), dyschondroplasia and abnormal chondrocyte differentiation (Henson et al. 1997; Singleton et al. 1997) or ischaemic necrosis of growth cartilage secondary to vascular failure (Carlson et al. 1995; Ytrehus et al. 2004a,b; Olstad et al. 2008). Evidence has been given that the cartilage canal vessels play an important role in the pathogenesis of the disease in pigs (Carlson et al. 1991; Ytrehus et al. 2004a,b) and horses (Olstad et al. 2007, 2008).

In horses, the most commonly affected joints are the femur-opatellar joint, the femorotibial joints, the tibiotarsal joint and the metacarpo- and metatarsophalangeal joints (McIlwraith 1993). In the tibiotarsal joint, commonly affected predilection sites include the cranial apex of the distal intermediate ridge of the tibia, the distal end of the lateral trochlear ridge of talus and the medial malleolus of the distal tibia (McIlwraith 1993). Previous studies reported the prevalence of osteochondrotic lesions in the tibiotarsal joint; in Standardbreds (SB) between 10.5% and 26.2% (Hoppe & Philipsson 1985; Schougaard et al. 1990; Grondahl 1991; Sandgren et al. 1993), in Swedish Warmblood (WB) 15.2% (Hoppe & Philipsson 1985), in the Marmenano horse 9.2% (Pieramati et al. 2003), in Hanoverian WB horses 9.6% (Stock et al. 2005a), in Dutch WB horses 31.4% (Van Grevenhof et al. 2009a), in Thoroughbreds (TB) between 4.0% and 4.4% (Kane et al. 2003; Oliver et al. 2008) and in South German Coldblood (SGC) horses 40.1% (Wittwer et al. 2006). Few reports of OC in ponies exist. In a survey of degenerative joint disease in the distal tarsal joints in 614 Icelandic horses, no radiographic signs of OC in the tibiotarsal joint were detected (Björnsdóttir et al. 2000). Incidence differences in breeds as well as in progeny groups suggest a genetic predisposition for OC. Heritability for OC in the tibiotarsal joint has been estimated in SB trotters ranging from \( h^2 = 0.24 \) to 0.52 (Schougaard et al. 1990; Grondahl & Dolvik 1993; Philipsson et al. 1993), in Hanoverian WB \( h^2 = 0.37 \) (Stock et al. 2005b), in Dutch WB \( h^2 = 0.36 \) (Van Grevenhof et al. 2009b) and in SGC \( h^2 = 0.04 \) (Wittwer et al. 2007a).

Microsatellite-based whole-genome scans in Hanoverian WB and SGC have identified a moderate number of quantitative trait loci (QTL) on different chromosomes linked to OC and OCD in tibiotarsal, metacarpo- and metatarsophalangeal joints (Dierks et al. 2007; Wittwer et al. 2007b). A QTL on Equus caballus chromosome (ECA) five for fetlock OC in Hanoverian WB has been refined (Lampe et al. 2009), and candidate genes associated with OCD in fetlock joints and OC in fetlock and hock joints in SGC are reported on ECA 4 and ECA 18, respectively (Wittwer et al. 2008, 2009). In this study, we performed a whole-genome scan, using the newly developed Illumina Equine SNP50 BeadChip® (San Diego, CA, USA), to identify loci associated with osteochondrosis dissecans at the intermediate ridge of the distal tibia in Norwegian SB trotters. We hypothesized that the high-density BeadChip offered a good opportunity to identify QTL associated with OCD status.

**Materials and methods**

**Study population and phenotypes**

**Animals**

Radiographic data and blood samples were obtained from 363 Norwegian SB trotter yearlings in 2007 and from 101 yearlings in 2008, representing 50% and 14% of the total cohort, respectively. Offspring of sires that had 10 or more progeny in 2006 were chosen as the sample population. The combined 464 horses were offspring of 22 different sires and 424 dams. and the number of animals per sire varied from 9 to 55. At examination, the animals had a mean (±SD) age of 12.1 (±1.8) months, a minimum age of 8.0 months and a maximum age of 17.9 months.

The horses were reared by their breeders, and feeding, housing and exercise levels varied among the animals. Because of extensive use of artificial insemination, the examined progeny of most sires were geographically well distributed in Norway and thus representative of the Norwegian SB population. All veterinary interventions were in accordance with the Norwegian Animal Welfare Regulations.

**Radiography**

Horses were examined radiographically at the Norwegian School of Veterinary Science (NSVS) (n = 110) and by staff from NSVS at nine regional equine clinics (n = 354). Digital radiography was the predominant examination technique, but conventional radiography was performed at two clinics (n = 107). Before examination, the horses were given de-tomidine (Domosedan vet®, Orion: 1 mg/100 kg I.V.) and butorphanol (Torbugesic®, Fort Dodge; 1 mg/100 kg I.V.).

The radiologic examination comprised ten views to reveal OC, OCD and other bony fragments in the metacarpo- and metatarsophalangeal joints and the tibiotarsal joint. Lateral-medial (90°) views were used for the metacarpophalangeal joint. The metatarsophalangeal joints were examined by dorso (45°) proximal (35°) lateral-plantarodistomedial and dorso (45°) proximal (35°) medial-plantarodistolateral views. The tibiotarsal joints were examined by dorso (45°) medial-plantarolateral oblique and dorso (30°) lateral-plantaromedial views.

**Phenotypic traits – Predilection sites and interpretation**

All radiographs were scrutinized by two of the authors, a professor in equine surgery and an experienced equine veterinary surgeon, for the presence of all types of bony fragments and OC. In the tibiotarsal joints, three predilection sites were evaluated bilaterally for OC and OCD: the intermediate ridge of the distal tibia, the lateral trochlear ridge of the distal tibia in Norwegian SB trotters. We hypothesized that the high-density BeadChip offered a good opportunity to identify QTL associated with OCD status.
Criteria for OC (Butler et al. 2008) included (i) the presence of irregular texture of the bone with variable radiopacity (e.g. radiolucency of the subchondral bone) and (ii) changes of the regular bone contour such as an irregularly flattened, smaller or larger concavity at the predilection site. OCD was diagnosed when isolated radiodense areas (osteochondral fragments) in the joint space were visible at the predilection sites. A corresponding defect in the underlying bone often accompanied the OCD lesion.

The metacarpophalangeal and metatarsophalangeal joints in particular were evaluated for OC/OCD at the dorsal sagittal ridge of the third metacarpal bone, and osteochondral lesions at the dorsoproximal rim of the proximal phalanx, the palmar and plantar processes of the proximal phalanx, the attachment sites of the short distal sesamoid ligaments, and also at the proximal sesamoid bones.

Genotyping

The genome-wide association analysis (GWAS) encompassed 162 of the examined yearlings (Table S1), including 80 quite uniform cases (39 females, 41 males), preferably with OCD just at the intermediate ridge of the distal tibia, and 82 controls (40 females, 42 males) without any developmental lesions in the joints examined. One hundred and twenty-eight of these yearlings were born in 2006 and 34 in 2007. They were offspring of 22 sires and 158 dams, and the number of animals per sire varied from 3 to 14. For genotyping, 5–10 μg genomic DNA (100 μl, 50–100 ng/μl) was isolated from 400 μl of proteinase-K-treated EDTA blood using an animal blood and tissue kit (QIAGEN). The DNA was measured using a Nanodrop® spectrophotometer ND-1000 (Thermo scientific, Delaware, CO, USA). The Equine SNP 50 BeadChip® (Illumina) was used for genotyping. The chip includes 54 602 evenly distributed SNPs (average probe spacing of 43.2 kb across all autosomes) derived from the EquCab2.0 SNP collection (http://www.broad.mit.edu/mammals/horse/), which were discovered in light coverage sequencing of a single horse of each of the following breeds: Arabian, Andalusian, Akhal-teke, Icelandic, SB, TB and Quarter horse. In SB, the BeadChip has an average call rate of 99.55%, 45,715 polymorphic loci (MAF > 0.05) and average MAF of 0.20.

The SNP array analysis of the 162 Norwegian SB trotter yearlings was performed by the Mayo Clinic’s Shared Genotyping Resource (Rochester, MN, USA). The laboratory protocols were all according to the manufacturers’ instructions and used Illumina’s software to call the genotype data.

Genotype quality assurance

All data were subject to quality control procedures, which were as follows. First, only samples with a minimum call rate of 95% were included. With an average genotype call rate of 99.18%, all samples met this criterion. Second, we discharged the SNPs with: (i) a call rate < 95% in the total sample (n = 470); (ii) those deviating from Hardy–Weinberg equilibrium in cases and controls (P < 0.001), or having a differential case/control missing of P < 0.01 (n = 88); and (iii) MAF < 0.05 in the total sample (n = 13 265). A total of 41 170 SNPs passed our quality control criteria.

Data analysis

All GWAS analyses and plots were performed using the whole-genome association analysis toolset PLINK (Purcell et al. 2007; Version 1.07), R and the pedigreemm package of the R environment (R Development Core Team 2008). The PLINK toolset and the R packages are freely available at http://pngu.mgh.harvard.edu/~purcell/plink/, http://www.r-project.org/ and http://r-forge.r-project.org/projects/pedigreemm/.

Osteochondrosis dissecans cases and healthy controls were compared by a basic association (chi-square) test in the whole-genome association analysis toolset PLINK; the Bonferroni correction was initially applied to address the problem of multiple comparisons. When deviation from expected distribution of P-values was demonstrated with quantile–quantile plots, suggesting inflation of the chi-square statistic because of population stratification, association with the disease phenotype was evaluated using logistic regression with sire treated as a covariate, using Cochran–Mantel–Haenszel (CMH) test 2 × 2 × K (K = 16) and finally by using mixed-model analyses (Yu et al. 2006).

Logistic regression analysis in PLINK was followed by 10 000 t-max permutations. The CMH test in PLINK was based upon 16 clusters generated by complete linkage agglomerative pairwise identity-by-state (IBS) clustering. The clustering process was modified by restrictions based on significance distance (P = 0.01) and a phenotype criterion (i.e. all clusters must contain at least one case and one control). Mixed-model analyses were performed in R using the pedigreemm package. The model includes a correction for population structure based on individual inbreeding coefficients as well as pairwise values of kinship. The relationship matrix was constructed by using the option “know pedigree relationships” as implemented in pedigreemm.

Quantile–quantile (QQ) plots showing the difference between expected and observed test statistics, and whole-genome association plots of significance, were generated in R. The association plots display the results for each chromosome as a negative log10 of the P-value. For genome-wide association studies, uncorrected P-values <5 × 10−7 provide strong evidence of association, whereas uncorrected P-values between 5 × 10−5 and 5 × 10−7 are considered to provide moderate evidence (Wellcome Trust Case Control Consortium 2007). When applying the Bonferroni correction or permutation, P-values ≤0.05 were considered to
provide strong evidence of association (Hirschhorn & Daly 2005). Genes and homologue regions in equine and human genomes were identified by the NCBI map viewer and Basic Local Alignment Search Tool (BLAST) (http://www.ncbi.nlm.nih.gov/MapView/), the Ensemble genome browser (http://www.ensembl.org/index.html) and the UCSC genome browser (http://genome.ucsc.edu/).

Results

The mixed-model statistics of the GWAS analysis revealed seven SNPs that showed moderate evidence of association with OCD status \( \left( P^* < 5 \times 10^{-5}; \log P^* \geq 4.30 \right) \); the asterisk (*) denotes an uncorrected P-value i.e. not adjusted for multiple comparisons); these SNPs were located on ECA 5, 10, 27 and 28 (Fig. 1). Table 1 gives significant test results \( \left( P \leq 5 \times 10^{-5} \right) \) with different test statistics, SNP positions on the chromosome, minor allele frequencies (MAF), odds ratio (OR) and genes within1 Mb of actual SNPs.

On ECA 5, a single SNP \((BIEC2-744792)\) showed moderate significant association in the mixed-model, the chi-square test and in the logistic regression statistics (Table 1).

Two SNPs on ECA 10 \((BIEC2-132748\) and \(BIEC2-132753\) represent the most significantly associated hits in both the mixed-model \((P^* = 1.19 \times 10^{-5}\) ), the chi-square statistics \((P = 9.31 \times 10^{-7}\); \(P = 0.038\) after Bonferroni correction) and also in the logistic regression statistics \((P^* = 1.28 \times 10^{-5}\); \(P = 0.054\) after t-max permutation) (Table 1). For both SNPs, the minor allele had an odds ratio (OR) of 0.23 (95% confidence interval (CI): 0.13–0.43) and allele frequencies of 0.10 and 0.32 in cases and controls, respectively (Table 1). The SNPs on ECA 10 were neighbours and flanked by two less significantly associated SNPs (Table 1).

Two SNPs on ECA 27 \((BIEC2-721410\) and \(BIEC2-722382\) showed moderate evidence of association with OCD status in the mixed-model, but not in the chi-square or the logistic regression statistics (Table 1). The associated SNPs on ECA 27 were located approximately 1 Mb apart, and \(BIEC2-721410\) was flanked by two less significantly associated SNPs (Table 1).

The neighbouring SNPs on ECA 28 \((BIEC2-744792\) and \(BIEC2-744794\) displayed moderate evidence of association with OCD status in the mixed-model \((P^* = 4.66 \times 10^{-6}\) ), but not in the chi-square or the logistic regression statistics (Table 1). Downstream, these SNPs were flanked by two less significantly associated SNPs (Table 1).

In the chi-square statistics, an additional seven SNPs on ECA 1, 3, 5, 10, 18 and 28 were found to be moderately associated with OCD status (Table 1).

The chi-square and logistic regression models showed evidence that the results were influenced by population substructure (identified by QQ-plots showing differences between expected and observed results, Fig. 2). This was not unexpected because of uneven distribution of cases and controls in the different half-sib family groups. The mixed-model and CMH analyses showed partial and almost full accounts for population substructure (Fig. 2).

Discussion

After conducting a case-control genome-wide association study to identify loci associated with OCD status in Norwegian SB trotters, we identified seven SNPs located on ECA 5, 10, 27 and 28 (Table 1) that in the mixed-model showed moderate evidence of association \((P^* < 5 \times 10^{-5}; \log P^* \geq 4.30 \) equivalents \(P^* \leq 5 \times 10^{-5}\) \). The moderate number of QTL on different chromosomes found for OCD in the tibiotarsal joint, suggests that several genes are possibly involved in the development of the condition (Table 1).

SNPs on ECA 10 \((BIEC2-132748\) and \(BIEC2-132753\) represent the most significant hits in the mixed-model, as well as in the other statistical models (Table 1). The SNPs achieved statistical significance in the basic association (chi-square) test with Bonferroni correction, and they also were close in the permuted logistic regression test (Table 1). The

![Mixed model](image-url)
## Table 1 Genomic regions associated with osteochondrosis dissecans at the distal intermediate ridge of the tibia in Norwegian Standardbred trotters.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr: position (bp)</th>
<th>Chi-square-test (P-value)</th>
<th>Logistic regression (P-value)</th>
<th>Permutated logistic regression (P-value)</th>
<th>CMH test (P-value)</th>
<th>Mixed-model (P-value)</th>
<th>Allele</th>
<th>MAF in cases</th>
<th>MAF in controls</th>
<th>OR (95% CI)</th>
<th>Genes within 1 Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIEC2-61415</td>
<td>ECA 1: 139 675 202</td>
<td>2.41E-05</td>
<td>1.06E-04</td>
<td>0.511</td>
<td>4.18E-03</td>
<td>8.00E-05</td>
<td>G/A</td>
<td>0.582</td>
<td>0.348</td>
<td>2.62 (1.67–4.12)</td>
<td>BIEC2-61415</td>
</tr>
<tr>
<td>BIEC2-810282</td>
<td>ECA3: 113 503 327</td>
<td>8.78E-06</td>
<td>8.20E-05</td>
<td>0.414</td>
<td>6.43E-04</td>
<td>6.24E-05</td>
<td>A/G</td>
<td>0.287</td>
<td>0.531</td>
<td>0.36 (0.23–0.57)</td>
<td>BIEC2-810282</td>
</tr>
<tr>
<td>BIEC2-870341</td>
<td>ECA 4: 76 971 434</td>
<td>1.76E-03</td>
<td>1.15E-03</td>
<td>1.000</td>
<td>7.44E-05</td>
<td>8.42E-05</td>
<td>A/G</td>
<td>0.250</td>
<td>0.116</td>
<td>2.54 (1.40–4.62)</td>
<td>BIEC2-870341</td>
</tr>
<tr>
<td>BIEC2-907168</td>
<td>ECA 5: 42 454 658</td>
<td>1.94E-05</td>
<td>1.69E-04</td>
<td>0.697</td>
<td>3.66E-03</td>
<td>1.47E-04</td>
<td>G/A</td>
<td>0.375</td>
<td>0.165</td>
<td>3.04 (1.81–5.13)</td>
<td>BIEC2-907168</td>
</tr>
<tr>
<td>BIEC2-920265</td>
<td>ECA 5: 77 424 986</td>
<td>1.60E-05</td>
<td>4.38E-05</td>
<td>0.222</td>
<td>1.24E-02</td>
<td>4.26E-05</td>
<td>A/G</td>
<td>0.119</td>
<td>0.317</td>
<td>0.29 (0.16–0.52)</td>
<td>BIEC2-920265</td>
</tr>
</tbody>
</table>

**Nearby genes:** CLCA1, CLCA2, ODF2L, COL24A1

**Important SNPs flanking significant SNPs on ECA 10, 27 and 28 are also included in the table. Abbreviations:** Chr, Chromosome; ECA, Equus callabus chromosome; CMH test, Cochran–Mantel–Haenszel test; MAF, minor allele frequency; OR, odds ratio; 95% CI, 95% confidence interval. Bold numbers represent significant test results.

1Bonferroni significance, when applicable, in brackets.

2The former allele represents the minor allele.
fact that the SNPs on ECA 10 are neighbours and also flanked by two less significantly associated SNPs (Table 1) increases the likelihood that they represent a true QTL associated with OCD in the tibiotarsal joint.

Disease-causing variants are seldom directly typed in GWAS studies, and the pattern of linkage disequilibrium (LD) is of importance when looking for candidate genes. LD in the horse is moderate, with long-range haplotype sharing among breeds (Wade et al. 2009). Table 1 lists candidate genes within 1 Mb of the significantly associated SNPs, although it is possible that genes further away can be involved, as the length of haplotypes in the population (or in different areas of the genome) is as yet unknown. LOCI00073151 (protein-coding gene similar to serum/glucocorticoid regulated kinase) located 0.15 Mb upstream of SNP BIEC2-132748 (Tables 1 & 2), represents a predicted gene closest to the associated SNPs. In a homologous region on human chromosome 6, the serum/glucocorticoid regulated kinase 1 gene (SGK1) can be identified (Table 2). SGK1 encodes a serine/threonine protein kinase that covers a wide variety of physiological functions and plays an important role in cellular stress response (Table 2). In rat brain, an increased level of SGK1 gene expression has been demonstrated after both focal and global brain ischaemia (Lu et al. 2003; Nishidaa et al. 2004). Ischaemic necrosis seems to play an important role in the aetiopathogenesis of OC (Carlson et al. 1995; Ytrehus et al. 2004a,b; Olstad et al. 2008), implying that SGK1 is potentially relevant to the pathogenesis of the disease.

On ECA 5, a single SNP (BIEC2-920265) located within the chloride channel, calcium activated, family member 4 (CLCA4) gene yielded moderate significant associations in all test statistics (Tables 1 & 2). Lampe et al. (2009) have identified collagen type XXVII alpha1 (COL24A1) as a potential candidate gene responsible for fetlock OC in WB horses (Table 2). COL24A1 is located 0.36 Mb downstream of 

Figure 2 Quantile–quantile (QQ) plots of results from the basic association (chi-square) test, the logistic regression with sire as covariate, the Cochran–Mantel–Haenzel test and the mixed-model analyses. Under the null hypothesis of no association at any locus, the points would be expected to follow the slope line. Deviations from the slope line correspond to loci that deviate from the null hypothesis. The slope line represent the negative log_{10} of expected versus expected P-values. Points represent the negative log_{10} of expected versus observed P-values.

<table>
<thead>
<tr>
<th>Candidate gene</th>
<th>Gene function</th>
<th>Gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC100073151 (protein-coding gene similar to serum/glucocorticoid regulated kinase)</td>
<td>In the syntenic region on human chromosome 6, the homologue Serum/glucocorticoid regulated kinase 1 (SGK1) can be identified (position; 134,490,387-134,639,196bp). SGK1 encodes a serine/threonine protein kinase that plays an important role in cellular stress response. This kinase activates certain potassium, sodium and chloride channels, suggesting an involvement in the regulation of processes such as cell survival, neuronal excitability and renal sodium excretion. The gene encodes a calcium-sensitive chloride conductance protein (chloride channel), reported to be located 0.36 Mb downstream of COL24A1 (Table 1 &amp; 2), represents a predicted gene encoding a chloride channel, calcium activated, family member 4 (CLCA4)</td>
<td>LOC100073151 (protein-coding gene similar to serum/glucocorticoid regulated kinase)</td>
</tr>
<tr>
<td>Collagen, type XXIV, alpha1 (COL24A1)</td>
<td>COL24A1 encodes collagen type XXIV. COL24A1 is found to be expressed in the forming skeleton of the mouse embryo and is also transcribed in the trabecular bone and periosteum of the newborn mouse. (Matsuo et al. 2008).</td>
<td>COL24A1 encodes collagen type XXIV. COL24A1 is found to be expressed in the forming skeleton of the mouse embryo and is also transcribed in the trabecular bone and periosteum of the newborn mouse. (Matsuo et al. 2008).</td>
</tr>
<tr>
<td>F-box protein 25 (FBXO25)</td>
<td>FBXO25 encodes a member of the F-box protein family. F-box proteins are determinant in ubiquitin-mediated proteolysis and are positioned as key regulators in many pathways of cell signalling, transcription and cell cycle.</td>
<td>FBXO25 encodes a member of the F-box protein family. F-box proteins are determinant in ubiquitin-mediated proteolysis and are positioned as key regulators in many pathways of cell signalling, transcription and cell cycle.</td>
</tr>
<tr>
<td>TBC1 domain family, member 22A (TBC1D22A)</td>
<td>The gene encodes a GTPase-activating protein that regulates small GTPases (guanine nucleotide-binding proteins). GTPases are proteins involved in cellular signal transduction. The small GTPases regulate a wide variety of processes in the cell, including growth, cellular differentiation, cell movement and lipid vesicle transport.</td>
<td>The gene encodes a GTPase-activating protein that regulates small GTPases (guanine nucleotide-binding proteins). GTPases are proteins involved in cellular signal transduction. The small GTPases regulate a wide variety of processes in the cell, including growth, cellular differentiation, cell movement and lipid vesicle transport.</td>
</tr>
</tbody>
</table>
BIEC2-920265, which might also reflect correspondence between this gene and OCD in tibiotarsal joints of Norwegian SB.

The two associated SNPs on ECA 27 (BIEC2-721410 and BIEC2-722382) are located upstream of the gene F-box protein 25 (FBXO25) (Table 2). The SNPs displayed moderate evidence of association with OCD status in the mixed-model, but are not supported with corresponding results in the remaining test statistics (Table 1). BIEC2-721410, being flanked by two less significantly associated SNPs, contributes supporting evidence that the SNPs on ECA 27 represent a true QTL associated with the disease (Table 1).

Moderately associated SNPs on ECA 28 (BIEC2-744792 and BIEC2-744794) are located within the gene TBC1 domain family, member 22A (TBC1D22A) (Table 2). Their statistics are quite similar to those encountered for SNPs on ECA 27: moderate evidence of association is only displayed in the mixed-model, but the fact that the SNPs are neighbours and flanked downstream by two less significantly associated SNPs suggests that they represent a true QTL for OCD.

Previous studies have reported QTL, potential candidate genes and differentially expressed genes related to OC in different horse breeds, pigs and humans. Dierks et al. (2007) performed a genome-wide search for microsatellite markers associated with OC/OCD in Hanoverian WB horses. Although the study identified a moderate number of QTL on different chromosomes, only two QTL on ECA5 represent a putative correspondence between the previous and the present studies. In a follow-up study, Lampe et al. (2009) reported refinement of a QTL on ECA5 and identification of COL24A1 as a potential candidate gene for OCD in fetlock joints. As previously mentioned, COL24A1 might also be involved in the pathogenesis of OCD at the intermediate ridge of the distal tibia in Norwegian SB. Wittwer et al. (2007b) mapped QTL for OC, OCD and palmar/plantar osseous fragments in fetlock joints of SGC horses. Numerous QTL were identified by the use of 250 microsatellite markers, but they were all different from the ones observed in the present study. An expression study in SB foals identified two upregulated genes, the equine tousled-like kinase 2 (Tlk2) and an unknown gene (EST: CD465746.1) (Austbo et al. 2010). The genes on ECA 11 and 13 do not correspond to QTL identified in the present study.

In pigs, Andersson-Eklund et al. (2000) identified QTL for OC on Sus scrofa chromosomes (SSC) 5 and 13. The QTL on SSC 13 did not correspond to any QTL identified in horses here. The QTL on SSC 5 is homologous to human chromosome 12q14-q24, which harbours the ALX homeobox 1 gene (ALX1). A BLAST of the human ALX1-region identifies ALX1 on ECA 28 at position 12.03–12.05 Mb. The large interval (approximately 30 Mb) between ALX1 and the QTL on ECA 28 observed in the present study gives no support to a putative correspondence. Laenoi et al. (2010) found downregulation of the matrix Gla protein gene (MGP) in OC compared to healthy cartilage and suggested that MGP might play an important role in the pathogenesis of OC in pigs. MGP on SSC 5 did not correspond to any QTL identified in horses here.

In humans, a missense mutation in the aggrecan C-type lectin domain gene disrupts extracellular matrix interactions in cartilage and causes dominant familial osteochondritis dissecans (Stat tin et al. 2010). The region on human chromosome 15 harbouring the mutated gene aggrecan (ACAN) shares homology with ECA 1, but does not correspond to QTL identified in the present study.

A possible explanation for the differences in the QTL identified between studies could be that the trait analysed is associated with different loci in different species and breeds. However, the fact that a genomic region is associated only in one horse breed does not necessarily mean the region is not a general risk factor. Rather the genomic markers may have different levels of penetrance in different breeds. In addition, a lack of statistical power caused by too few animals and/or markers could result in important markers falling below the significance thresholds, or smaller risk factors being missed (Hirschhorn & Daly 2005). A huge difference in marker coverage in microsatellite-based versus SNP chip-based genome scans represents a third explanation for the unequal study results. The equine SNP chip provides an evenly distributed and approximately 250-fold increase in genomic coverage and should therefore be well suited to detect QTL in GWAS (Wade et al. 2009). On the other hand, thousands of markers represent a multi-testing problem, possibly resulting in more false positive results (Balding 2006). Various phenotypic criteria applied when selecting cases can be another explanation for the different results, as lesions at different predilection sites are possibly influenced by different genes. An apparent advantage of the present study is that the GWAS sample came from a homogenous population containing cases with a stringent phenotype (McCarthy et al. 2008). Also, as radiographic identification of OC and OCD lesions can be challenging, including only horses with OCD at the intermediate ridge of the distal tibia might facilitate correct identification of cases.

The results of gene expression studies in OC are especially difficult to interpret. As long as the aetiologypathogenesis, timeline and sequence of events in OC have not been clarified, it is difficult to know whether the differentially expressed genes reflect events prior to the disturbance in enchondral ossification, the disease process itself, or secondary repair processes within the bone. Subsequently, a lack of correlation between the present study and past expression studies might simply reflect different stages in the disease process.

The goal of GWAS studies is to identify patterns of polymorphisms that vary systematically between individuals with different disease states and could therefore represent the effects of risk-enhancing or protective alleles (Balding...
2006). However, the genome is so large that patterns that are suggestive of causal polymorphism could well arise by chance. To help distinguish causal from spurious signals, tight standards for study design, quality control and statistical approach need to be established. A potential limitation of our study is that the GWAS results may have been affected by discrete population structure, as most QQ-plots (Fig. 2) showed some difference between observed and expected results (Hirschhorn & Daly 2005; McCarthy et al. 2008). Dealing with association within but not between individual strata, the IBS-clustered CMH test successfully accounted for population substructure, but resulted in less significant SNP associations. As mixed-model analyses are considered to be better corrections of population stratification in GWAS (Balding 2006; Yu et al. 2006), the interpretation of our results is primarily based on these statistics.

In conclusion, this is the first report of applying the Equine SNP50 BeadChip® (Illumina) for genome-wide association analysis of OCD in horses. The study identified several SNPs that showed evidence of association with OCD in the tibiotarsal joint in Norwegian SB trotters. Putative QTL on ECA 5, 10, 27 and 28 represent the most significant hits and would be interesting areas for future research. Further studies must include replication and validation of these study results by including a larger number of animals or joint (meta-) analysis of data from comparable GWAS. When confirmed, refinement of candidate regions will eventually follow.

Acknowledgements

The current study was funded by the University of Minnesota, the Norwegian School of Veterinary Science and the Norwegian Horse Center. Funding for the equine SNP chips was provided by the US Department of Agriculture (National Research Initiative Competitive Grants Program) and the Morris Animal Foundation. The authors are indebted to Prof. Theo Meuwissen and Morten Mattingsdal for assistance in selecting samples for genotyping and interpretation of analyses. We gratefully acknowledge the generous assistance of the staff at the Section for Equine Medicine and Surgery, Radiology and Genetics of the Norwegian School of Veterinary Science, as well as staff at the Department of Veterinary and Biomedical Sciences of the University of Minnesota.

Conflicts of interest

The authors have declared no potential conflicts.

References


associated with osteochondrosis in fetlock and hock joints of South German Coldblood horses. *Journal of Heredity* 100, 481–6.


**Supporting information**

Additional supporting information may be found in the online version of this article.

**Table S1** Radiographic findings in the 162 genotyped horses.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Structural annotation of equine protein-coding genes determined by mRNA sequencing

S. J. Coleman*, Z. Zeng†, K. Wang‡, S. Luo‡, I. Khrebtukova‡, M. J. Mienaltowski*†,1, G. P. Schroth‡, J. Liu† and J. N. MacLeod*

*Department of Veterinary Science, Maxwell H. Gluck Equine Research Center, University of Kentucky, Lexington, KY 40546, USA.
†Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA.
‡Illumina Inc., 25861 Industrial Boulevard, Hayward, CA 94545, USA

Summary

The horse, like the majority of animal species, has a limited amount of species-specific expressed sequence data available in public databases. As a result, structural models for the majority of genes defined in the equine genome are predictions based on ab initio sequence analysis or the projection of gene structures from other mammalian species. The current study used Illumina-based sequencing of messenger RNA (RNA-seq) to help refine structural annotation of equine protein-coding genes and for a preliminary assessment of gene expression patterns. Sequencing of mRNA from eight equine tissues generated 293,758,105 sequence tags of 35 bases each, equalling 10.28 gbp of total sequence data. The tag alignments represent approximately 207× coverage of the equine mRNA transcriptome and confirmed transcriptional activity for roughly 90% of the protein-coding gene structures predicted by Ensembl and NCBI. Tag coverage was sufficient to refine the structural annotation for 11,356 of these predicted genes, while also identifying an additional 456 transcripts with exon/intron features that are not listed by either Ensembl or NCBI. Genomic locus data and intervals for the protein-coding genes predicted by the Ensembl and NCBI annotation pipelines were combined with 75,116 RNA-seq-derived transcriptional units to generate a consensus equine protein-coding gene set of 20,302 defined loci. Gene ontology annotation was used to compare the functional and structural categories of genes expressed in either a tissue-restricted pattern or broadly across all tissue samples.

Keywords: gene structure annotation, transcriptome, gene expression, Equus caballus, RNA-seq.

Introduction

After completion of genomic DNA sequencing, the structural annotation of protein-coding genes in an organism is generally accomplished by cDNA and protein sequence alignment, ab initio gene prediction, and inter-species sequence comparison. While in silico strategies have improved substantially, the gold standard for gene identification and structural annotation remains the alignment of full-length cDNA sequence (Brent 2008). Analyses of human and mouse mRNA structures are supported by large EST data sets numbering 8,301,471 and 4,852,146, respectively (NCBI dbEST, http://www.ncbi.nlm.nih.gov/dbEST). Many organisms, however, are at a disadvantage in this regard. From 196 animal species with genome sequencing projects (Entrez Genome Project, http://www.ncbi.nlm.nih.gov/genomes/leukcs.cgi), 80% are supported by fewer than 100,000 ESTs, with a median of only 17,435 ESTs (NCBI dbEST Summary, http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). The horse is an example. Equine-specific expressed sequence data available in public databases are limited to 35,702 ESTs and 1,236 mRNA sequences (UCSC Genome Browser, http://genome.ucsc.edu/). As a result, structural annotations for the majority of protein-coding genes in the current equine gene

Sequencing of mRNA (RNA-seq) has been used to investigate the transcriptomes of human, mouse, yeast and Arabidopsis (Cloonan et al. 2008; Lister et al. 2008; Morin et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Pan et al. 2008; Rosenkranz et al. 2008; Sultan et al. 2008; Wang et al. 2008; Wilhelm et al. 2008) and is clearly an important new tool for transcriptome research (Wang et al. 2009). Data generated in these experiments generally exhibit a high degree of concordance with established gene annotations. Even so, a surprising number of exons, mRNA splicing patterns and novel genes have been detected that were not previously identified in these highly studied species. The current report uses the draft assembly of the equine genome (Wade et al. 2009) and RNA-seq data from eight equine tissue samples to refine the structural annotation of protein-coding genes in the horse and to generate a preliminary assessment of tissue-specific expression patterns.

Materials and methods

Experimental samples

Eight equine tissue samples were studied. Articular cartilage and synovial membrane samples with and without pro-inflammatory stimuli using bacterial lipopolysaccharide (LPS) were collected from the left and right carpal joints of a 3-year-old male pony. Antemortem, the left carpal joints received four LPS injections (0.5 ng) over 8 days, while the right carpal joints received control injections of PBS (MacLeod et al. 1998). A cerebellum sample was collected from a 2-year-old female thoroughbred, and a testis sample was collected from a 4-year-old thoroughbred. A placental villous sample was collected immediately post-partum from a 2-year-old female thoroughbred, and a testis sample was collected from a 4-year-old thoroughbred. A whole embryo sample was obtained from a 34-day-old male thoroughbred conceptus, after non-surgical recovery from the uterus (Antczak et al. 1987).

RNA isolation

Total RNA was isolated from the eight samples using variations of guanidinium-based protocols. RNA was purified from the articular cartilage samples as described in MacLeod et al. (1996), from control synovial membrane, cerebellum, testis and placenta villous using acid guanidinium thiocyanate/phenol/chloroform extraction with alcohol and salt precipitations (Chomczynski & Sacchi 1987), and from LPS-stimulated synovial membrane and whole embryo with RNeasy mini kits (Cat# 74104: Qiagen). All samples were processed with a final RNeasy (Qiagen) silica gel spin column clean up, ethanol precipitation and solubilization in nuclease-free water (Cat# AM9937: Ambion). Processing total RNA through RNeasy columns has been shown to remove a substantial majority of small non-coding RNAs (Mraz et al. 2009). Quantitative and qualitative parameters of the RNA preparations were assessed using a NanoDrop ND-1000 and Bioanalyzer 2100 (Agilent. Eukaryotic Total RNA Nano Series II). All samples had 260/280 ratios ≥1.93, 260/230 ratios ≥1.75 and an Agilent RNA integrity number (RIN) ≥8.1.

Library preparation and sequencing

Library preparation and subsequent nucleotide sequencing was performed according to Illumina’s standard mRNA-seq kit protocol (Transcriptome Analysis: mRNA-seq, http://www.illumina.com/pages illumina?D = 291). Messenger RNA was selected with PolyT capture beads from 3.5 to 10.0 µg of total RNA and fragmented by zinc metal ion hydrolysis. RNA fragments were converted into cDNA using random hexamer primers and made double stranded using RNase H and DNA polymerase. The cDNA fragments were ligated to sequencing adaptors using Illumina’s Genomic DNA sample prep. Fragments between 100 and 300 bp were size selected by gel electrophoresis, applied to sequencing flow cells and enriched by PCR to grow fragment clusters. Sequencing of 35-bp reads was performed on an Illumina Genome Analyzer. Data were captured and processed with version 0.3 of Illumina’s standard analysis pipeline (Genome Analyzer Pipeline, http://www.illumina.com/documents/products/datasheets/datasheet_genome_analyzer_software.pdf). Tag sequences have been deposited in NCBI’s Gene Expression Omnibus (Edgar et al. 2002; Barrett et al. 2009) and are accessible through GEO series accession number GSE 21925 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21925).

Sequence tag alignment

Sequencing tags were mapped to the reference equine genome sequence (EquCab2: Wade et al. 2009) using ELAND (efficient local alignment of nucleotide data – Illumina analysis pipeline version 0.3). Mapping coordinates were assigned using Illumina’s default alignment parameters (best co-linear alignment, allowing up to two mismatches with no insertions or deletions, and an alignment score of 4 or greater). Sequence tags that did not map by Illumina’s default parameters were re-evaluated to determine if they would align with a single gap to the reference genome (Wang et al. 2010). A maximum of two nucleotide mismatches were allowed for the gapped alignments. To minimize spurious junction tag alignments with the initial assessment, a minimum of 10 bases on either side of the gap was required. From the resulting putative splice sites, synthetic junction sequences were then generated by concatenating 31 nucleotides from each side of the gap.
Unmapped tags were re-assessed to these junction sequences with stringency relaxed to a minimum of four bases on each side of the splice junction. To identify novel exons and transcriptional units, overlapping targeted regions in the equine genome were searched sequentially. Each region corresponded to a known gene locus and both the adjacent 5’ and 3’ intergenic regions. The smallest gap allowed was 50 nucleotides, based on parameters defining the minimum size of introns (Deutsch & Long 1999). The maximum gap allowed was limited only by the length of the chromosome.

Annotation of protein-coding gene structures with RNA-seq data

Tag alignment cluster (TAC) was the term given to a linear segment of nucleotides in the genome where the depth (number of tag hits) at each individual base was higher than a defined threshold. To reduce ambiguity and increase confidence in the structural annotation derived from the tag alignments, a minimum depth threshold of 15 was used. Based on intron size limits (Deutsch & Long 1999) and some tag depth oscillations, adjacent regions above the threshold were grouped together into a single TAC if separated by <50 nucleotides. TACs were then structurally organized into TAC-contigs if connected by at least two splice junction tags with the same gap in genomic sequence alignment positioned at different points along their 35 base lengths. Nucleotides at the start and end of the alignment gap defined exon splicing boundaries with single-base resolution. The lowest and highest nucleotide coordinates, based on alignment to EquCab2.0 (Wade et al. 2009), defined the start and end points of individual TACs and TAC-contigs.

Statistical analyses

Overall concordance of the 5’ and 3’ extent of RNA-seq data to Ensembl equine gene predictions at the same loci and orthologous human or mouse mRNA transcripts was determined by calculating and comparing the variance of available distributions. First, positional differences were calculated by subtracting the 5’ and 3’ basepair position of the RNA-seq data from the corresponding positions of either the Ensembl gene prediction or the human or mouse transcript. The calculated differences for each set were plotted as a distribution around the RNA-seq defined positions on two graphs, one for the 5’ extent and the other for the 3’ extent. Then, for each of the plots generated, Levene’s test was performed to assess the equality of variance between the two sets of samples. The null hypothesis was that the population variances were equal.

A Satterthwaite’s t-test was used to determine if there was a significant difference in tag densities between adjacent TAC-contigs to assist in assessing if they represented two distinct gene structures.

Consensus gene models

Consensus equine gene models were generated by consolidation of the in silico predictions by Ensembl (Ensembl Horse Genome Browser, http://www.ensembl.org/Eq-us_caballus/) and NCBI (NCBI Equine Genome Page, http://www.ncbi.nlm.nih.gov/projects/genome/guide/horse/) with the structural annotation derived from the RNA-seq experiments. First, the genomic locus data and intervals for the genes predicted by Ensembl and NCBI were combined in one list and sorted to identify loci supported only by Ensembl or NCBI annotation and the concordant overlapping loci identified by both groups. When the intervals of two predictions (from both the Ensembl and NCBI gene lists) were concordant and overlapping, the intersection (overlap length) and union (combined length) of their respective coding sequences were calculated. These predictions were considered to represent the same gene when the intersection of their coding regions was at least 50% of the union. Next, structural boundaries for the resulting loci were compared to the TACs and TAC-contigs defined by the RNA-seq tag alignments in the current dataset. Experimentally derived RNA-seq annotation superceded the in silico predictions in reaching consensus gene models (Fig. S1a). For loci containing discordant predictions from Ensembl and NCBI where there were insufficient RNA-seq tag alignments to refine expressed structures, the longest consensus model generated from the in silico predictions was used (Fig. S1b). Gene symbols were assigned to the consensus models based on the established Ensembl and NCBI annotations.

Gene expression analysis

Relative gene expression levels between tissues were determined by counting the number of tags aligned over the defined coding regions of each consensus gene model. Expression values (tag density) were calculated by converting tag counts into hit counts (individual basepair coverage, 1 tag = 35 hits) and dividing the total hit count by the gene’s cDNA length. Relative expression data between tissues were normalized to the total number of hits generated for each RNA sample. A threshold level for annotating expression of 3 hits per 10 bases was identified empirically by the comparison of tag density between the coding and intergenic regions of the genome to maximize the detection of expression (Fig. S2). In an independent study, a normalized density value of 0.3 was also selected as an expression threshold value to minimize false discovery and false negative rates (Ramsköld et al. 2009). Gene ontology terms were extracted and summarized using the Generic Gene Ontology Term Mapper (http://go.princeton.edu/cgi-bin/GOTermMapper) and the Gene Ontology Annotation generic human GO slim (http://www.ebi.ac.uk/GOA/).
Results

Structural annotation analysis

Deep sequencing of eight equine tissues (normal and bacterial lipopolysaccharide (LPS)-stimulated articular cartilage, normal and LPS-stimulated synovial membrane, placental villous, testis, cerebellum and a 34-day embryo) on an Illumina Genome Analyzer generated 293 758 105 tags, each 35 nucleotides in length (Table 1). Of the total tags generated, 54% (158 787 455) mapped to the second build of the equine genome sequence (EquCab2) at a single unique position with a contiguous 35 bp alignment. Tags that did not map to the genome sequence by default parameters (up to two mismatches with an alignment score of 4 or greater) were reassessed to determine whether they mapped with a gapped alignment. Such alignment gaps would be generated with tags that cross adjacent exons at a splice junction. In this fashion, 43 316 466 additional tags (14.7%) were mapped to the genome with a gapped alignment and are termed splice junction tags. Computational analyses of the RNA-seq tag alignments identified 169 170 tag alignment clusters (TACs). Sixty-seven per cent (113 432) were linked together by splice junction tags, generating 19 378 TAC-contigs.

The RNA-seq tag alignments confirmed the majority of the in silico protein-coding gene predictions made in the equine genome sequence, while also enabling refinements of the structural annotation details. Approximately 89% of the exon predictions (Ensembl: 148 879/169 073; NCBI: 177 825/196 513) and 89% of the gene predictions (Ensembl: 15 654/18 039; NCBI: 15 665/16 983) from single gene loci (genes that do not co-localize with another gene or pseudogene structure, for example on the opposite DNA strand) had tag alignments that directly demonstrate expression (Table 2). The remaining 11% of non-overlapping equine gene predictions (Ensembl: 2385/18 039; NCBI: 1319/16 983) did not have aligned RNA-seq tags, likely reflecting to a large extent incomplete tissue diversity and developmental time points in the samples analysed. Four hundred and sixty-six TAC-contigs were identified that did not co-localize with any in silico gene prediction. BLASTN analysis (discontinuous megablast; e-value ≤-5, bit score ≥100) reduced this total to 456 transcriptional units by linking 19 of the TAC-contigs to nine genes. Concordance at the gene level of the RNA-seq data with the Ensembl and NCBI in silico gene prediction lists is illustrated in Fig. 1.

Structural refinements

Nucleotide coordinates defining equine gene boundaries as indicated by RNA-seq data were compared to Ensembl predictions. At both the 5’ and 3’ ends, RNA-seq data displayed higher concordance (P < 0.01) with orthologous human or mouse transcripts than with the Ensembl models (Fig. 2). In silico predictions frequently did not fully extend to the transcript boundaries defined by tag alignments, visualized as lower peaks at the end points and a broader shoulder to the curves within the gene coordinates. The number of Ensembl predictions that have endpoints located within 20 bases of the transcript boundaries defined by RNA-seq data (0 point on the abscissa) was approximately 60% lower than the number of orthologous human or mouse transcripts within the same 20 base interval. In addition, the 3’ ends of transcripts annotated by RNA-seq data correlated highly with the 18 625 equine ESTs in dbEST that contain a polyadenylation signal (Fig. S3). Thus, analysed on a transcriptome level, the RNA-seq data appear to provide a more accurate assessment of the 5’ and 3’

Table 1 Summary of RNA-seq tag alignments and structural annotation.

<table>
<thead>
<tr>
<th>Tag mapping summary</th>
<th>Number of sequence tags</th>
<th>BLASTN analysis bit score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total 35-bp tag sequences generated</td>
<td>293 758 105</td>
<td>-</td>
</tr>
<tr>
<td>Tags aligned to a single genomic locus with ≥94% (33/35 bases) nucleotide identity</td>
<td>158 787 455</td>
<td>-</td>
</tr>
<tr>
<td>Tags aligned to a single genomic locus with 100% (35/35 bases) nucleotide identity</td>
<td>144 031 041</td>
<td>-</td>
</tr>
<tr>
<td>Tags that mapped only once to genome with a single alignment gap</td>
<td>43 316 466</td>
<td>-</td>
</tr>
<tr>
<td>Tags not used in the gene structure annotation (multiple alignments, low alignment score)</td>
<td>91 654 184</td>
<td>-</td>
</tr>
<tr>
<td>Tag alignments in clustered groups</td>
<td>169 170</td>
<td>-</td>
</tr>
<tr>
<td>Number of sequence tags in TACs</td>
<td>142 686 959</td>
<td>-</td>
</tr>
<tr>
<td>TACs linked by tags with single alignment gap (splice junction tags)</td>
<td>113 432</td>
<td>-</td>
</tr>
<tr>
<td>Number of linked TAC groupings (TAC-contigs)</td>
<td>19 378</td>
<td>-</td>
</tr>
<tr>
<td>Number of sequence tags in TAC-contigs</td>
<td>118 204 109</td>
<td>-</td>
</tr>
</tbody>
</table>

Of the 293 758 105 tags generated, 68.8% (158 787 455 + 43 316 466) mapped to the genome at a single locus and were used for the annotation analysis. Tag alignment clusters (TACs) that rose above a depth threshold of 15 were used to annotate exon structure. Tags with a single gap in their genomic alignment that connected two TACs and suggested adjoining exons in mRNA structure were called splice junction tags. A string of two or more TACs linked by splice junction tags were termed TAC-contigs. Results of BLAST analyses using TAC-contig sequences to nucleotide and protein databases from other mammalian species were used in some cases to support the grouping of more than one TAC-contig to a single equine gene.

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 121–130
boundaries that reflect the combined set of mRNA transcript variants of a particular gene.

Individual examples of predicted equine gene structures that can be refined by RNA-seq data are illustrated in Fig. 3. The first category is the identification of additional exons. Three additional 5' exons were present on the TAC-contig corresponding to the ribosomal protein, large, P0 (RPLP0) gene locus (Fig. 3a). Across the equine genome, additional exons were identified for at least some of the transcripts for 2344 genes relative to Ensembl predictions (Table S1). In the second category, nucleotide coordinates defining the start and end points of individual exons can be better determined. The 3' end of at least some transcripts that encode type II collagen (COL2A1) extend farther than the in silico prediction, consistent with equine EST data in GenBank (Fig. 3b). Protein-coding genes for which RNA-seq tag alignments extend the boundaries of the first (3095) or last (2991) exon are listed in Table S2. The last category is comprised of novel transcriptional units or separate genes not predicted by in silico methods. Ensembl annotated a 55 000 nucleotide region on ECA16 as containing a single protein-coding gene. However, both the mRNA structure predicted by the position of splice junction tags and the differential tag densities of individual TACs indicate that this region actually contains two equine genes, transferrin (TF) and signal recognition particle receptor, B subunit (SRPRB) (Fig. 3c). Average tag density of the TF exons is 706 ± 335, compared to 189 ± 75 for the SRPRB exons (mean ± SD; P < 0.01).

Consensus gene models

A consensus set of equine protein-coding genes was defined by consolidation of gene sets predicted by Ensembl and NCBI and the expressed structures annotated from the RNA-seq data. Consolidation of the two in silico gene sets resulted in the definition of 20 087 loci. RNA-seq structures with tag hits above the minimum depth threshold of 15 overlapped the intervals of 11 356 loci and were used to refine structural annotation for the coding region of those genes, while 5914 relied on the longest consensus exon models derived from the Ensembl and NCBI in silico predictions. The final 2817 loci that included overlapping structures representing multiple genes were not analysed because they lacked sufficient information to resolve their internal structure. Two hundred and fifteen of the 456 transcriptional units annotated from the RNA-seq data that are not predicted as equine genes by Ensembl or NCBI.

Table 2 RNA-seq tag alignments to in silico structural predictions of equine protein-coding genes.

<table>
<thead>
<tr>
<th>Ensembl</th>
<th>NCBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of in silico protein-coding gene predictions</td>
<td>20 322</td>
</tr>
<tr>
<td>Protein-coding gene predictions that do not have other overlapping gene or pseudogene structures</td>
<td>18 039</td>
</tr>
<tr>
<td>Non-overlapping gene predictions with aligned RNA-seq data</td>
<td>15 654</td>
</tr>
<tr>
<td>Non-overlapping gene predictions without aligned RNA-seq data</td>
<td>2385</td>
</tr>
<tr>
<td>Total number of tags aligned to non-overlapping gene predictions</td>
<td>87 214 512</td>
</tr>
</tbody>
</table>

Only genes annotated as protein-coding were included in the analysis.

Figure 1 Venn Diagrams illustrating the number of Ensembl (a) and NCBI (b) in silico equine gene predictions supported by RNA-seq tag alignments. The analysis considered only the 18 039 (Ensembl) and 16 983 (NCBI) genes annotated as protein coding in each gene set that do not co-localize with another gene or pseudogene structure. Values reported in the intersecting regions with bold font reflect the number of predictions from one set which overlap with the genomic position of a prediction in the other set and/or gene expression by RNA-seq. Four hundred and fifty-six transcriptional units with exon/intron structure were identified by the RNA-seq data that are not predicted as equine genes by Ensembl or NCBI.
Gene expression was analysed for six of eight equine samples. The LPS-stimulated articular cartilage and synovial membrane samples were not included to avoid confounding patterns of gene expression resulting from the inflammatory response. Relative expression levels between tissues were determined for 17,270 of the consensus genes that do not structurally overlap with other protein-coding genes in the equine genome. A description of the gene set used for the expression analysis is presented in Table 4. The number of genes expressed by individual tissues ranged from 9716 (articular cartilage) to 12,038 (34-day embryo), or roughly 56–70% of the genes analysed (Table 5). Analysis of the expressed gene sets identified 8884 genes with expression above a normalized density threshold of 0.3 in all tissues. The majority of these genes were expressed in the same order of magnitude in all tissue samples. The same analysis found 1230 genes with expression in only one of the six tissues. A summary of tissue expression support for the 17,270 genes analysed is presented in Fig. 4.

Gene ontology (GO) annotation data related to biological process, molecular function and cellular component were used to evaluate the functional and structural categories of the genes expressed either in all six tissues (stable) or in a tissue-restricted pattern. Gene ontology data were also available for 717 of the 1230 (58.3%) genes with a tissue-restricted pattern of expression. GO annotation summaries are presented in Fig. 5.
Table 3: Summary of consensus equine protein-coding genes.

<table>
<thead>
<tr>
<th>Structure support</th>
<th>in silico only</th>
<th>Refined by RNA-seq annotation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loci with structures predicted by both Ensembl and NCBI</td>
<td>3184</td>
<td>10 009</td>
<td>13 193</td>
</tr>
<tr>
<td>Loci with structures predicted by Ensembl only</td>
<td>1947</td>
<td>1038</td>
<td>2985</td>
</tr>
<tr>
<td>Loci with structures predicted by NCBI only</td>
<td>783</td>
<td>309</td>
<td>1092</td>
</tr>
<tr>
<td>Loci with structures supported by RNA-seq only</td>
<td>–</td>
<td>215</td>
<td>215</td>
</tr>
<tr>
<td>Loci requiring additional data to resolve gene structure</td>
<td>2817</td>
<td>–</td>
<td>2817</td>
</tr>
<tr>
<td>Total</td>
<td>8731</td>
<td>11 571</td>
<td>20 302</td>
</tr>
</tbody>
</table>

Structural annotation refinement with RNA-seq data was only considered for gene loci where tag alignment depth was above 15.

Table 4: Summary of protein-coding gene loci used for gene expression analysis.

<table>
<thead>
<tr>
<th>Annotation category</th>
<th>No. of Loci</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>In silico</em> gene predictions with structural annotation refined by RNA-seq data</td>
<td>11 356</td>
</tr>
<tr>
<td><em>In silico</em> gene predictions that did not have RNA-seq tag alignments of sufficient depth to enable structural annotation refinement</td>
<td>5914</td>
</tr>
<tr>
<td>Total</td>
<td>17 270</td>
</tr>
</tbody>
</table>

Table 5: Total genes expressed by tissue sample.

<table>
<thead>
<tr>
<th>Tissue sample</th>
<th>Genes expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articular cartilage</td>
<td>9716</td>
</tr>
<tr>
<td>Synovial membrane</td>
<td>11 190</td>
</tr>
<tr>
<td>Placental villous</td>
<td>10 891</td>
</tr>
<tr>
<td>Testes</td>
<td>12 013</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>11 129</td>
</tr>
<tr>
<td>34-day Embryo</td>
<td>12 038</td>
</tr>
</tbody>
</table>
to have functions related to molecular binding, catalytic activity and metabolic processes. By contrast, genes with a tissue-restricted pattern of expression were more likely to be localized in the extracellular region or on the cell membrane and have functions related to signal transduction and membrane transport involved in the processes of cell communication, differentiation and response to stimulus.

**Discussion**

Application of next-generation sequencing technology to the study of transcriptomics is quickly becoming a standard approach used for the annotation of protein-coding genes. Techniques like RNA-seq are capable of producing high-quality sequence data at coverage depths sufficient for the annotation of gene structure and evaluation of expression patterns at a genome-wide level. This technology is particularly useful for newly sequenced genomes that lack the critical mass of species-specific expression data required for accurate characterization of gene structures. The current report demonstrates the application of RNA-seq to the equine mRNA transcriptome, in which in silico-derived gene annotations have been generated for the high-quality draft equine genome sequence but with very limited input from horse-specific expression data.

Computational analyses of the RNA-seq tag alignment patterns identified 19,378 TAC-contigs and 55,738 TACs that were used to refine the exon structure for 11,356 genes. The experimentally derived RNA-seq structures proved to more accurately reflect mRNA transcript structure, including the untranslated regions. The in silico predictions, by comparison, seemed to more closely reflect protein structure. This makes sense given that protein structure and not nucleotide sequence would more routinely be used to make gene structure predictions based on projection from other species, which is the case for the majority of equine gene predictions. Structures annotated from the tag alignments identified additional or extended exons for more than one-third of the genes analysed. The majority of the 466 RNA-seq transcriptional units with no corresponding in silico prediction will most likely be equine homologues of known genes.

Gene expression was analysed at 17,270 discrete loci. The average number of genes expressed by individual tissues was 11,163 (64.6%). As expected, the 34-day embryo had the most complex transcriptome and articular cartilage the least. This likely reflects differences in cell type complexity between samples. The embryo represents the full spectrum of developing tissues with diverse functional roles and gene expression patterns. Articular cartilage, on other hand, is composed of a single cell type with a focused functional role. The average number of genes expressed by tissues in this study is in close agreement with values determined for a number of human and mouse tissues (Ramsköld et al. 2009), though the range is somewhat lower. These differences likely reflect an analysis of the full protein-coding gene set (~22,600) in the mouse and human data.

Fifty-one per cent (8884) of the genes demonstrated expression above the 0.3 threshold in all six tissues studied. A similar analysis by Ramsköld et al. (2009) identified 7897 genes which they called ubiquitously expressed (expressed in all 24 human tissues analysed). Their analysis of more tissues explains the lower number of genes in the ‘all tissues’ category. Of the remaining 8386 genes analysed, 3471 were expressed in different subsets of the samples analysed, 1230 were expressed in only one tissue and 3685 had tag densities below the 0.3 threshold. The GO annotation analysis showed the expected results. Biologically, it makes sense that diverse cell types exhibit shared expression of genes involved in primary metabolic, transcriptional and translational processes. The basic details of these processes are required by all cell types. It could also be expected then that genes with tissue-restricted patterns of expression would have roles in the regulation of these basic processes, which would dictate different responses to biological stimuli, helping to distinguish different tissue types. While tissue-restricted genes did include a higher portion of response genes, it was somewhat surprising that there was not a corresponding increased representation of regulatory genes when compared to the genes expressed by all tissues. It is also logical that the functional and structural roles of genes with a tissue-restricted expression pattern be over-represented within the cell membrane and extracellular regions and have functions related to transport and communication. Unique differences in these areas relate to tissue morphology and have dramatic effects on how cells perceive and react to their immediate microenvironment, contributing to their differentiated phenotype.

The total number of nucleotides in the equine genome expressed in various patterns to generate the protein-coding
mRNA transcriptome is estimated at 34,203,688 bases (1.28% of the whole genome). Equine EST and mRNA data currently available in public databases approximate only 0.59× coverage of the horse transcriptome. This value actually overestimates the extent of coverage, because many of the ESTs overlap at the 3′ end of transcripts. The 202,103,921 RNA-seq tags with unique single alignments to the equine genome that were generated in the current project achieve roughly 207× coverage of the horse mRNA transcriptome, an increase of about 350× over the previous 0.59×. This translates into experimentally supported structural annotation for 11,356 equine genes and expression support for almost 90% of predicted gene structures. Although clearly not evenly distributed because of different levels of gene expression in the tissues studied, the RNA-seq data yielded a very substantial increase in equine-specific expression data. These results reinforce the utility of this approach for annotation and characterization of protein-coding genes in genomes lacking extensive EST and cDNA resources.

**Acknowledgements**

We thank Jan Prins for his participation in the development of algorithms to map splice junction tags and Rebekah Cosden for assistance with the preparation of RNA isolates from the experimental samples. We thank Ernie Bailey, Arne Bathke, Teri Lear, Tim McClintock, Martha Peterson, Jan Prins, Brett Spear, and Arny Stromberg for critical review of the manuscript. This work was funded by the University of Kentucky.

Figure 5 GO annotation summaries by (a) Cellular component, (b) Biological process and (c) Molecular function.
Gluck Equine Research Foundation, the Morris Animal Foundation, the National Science Foundation (EF-0850237), and the Kentucky Biomedical Research Infrastructure Network (KBRIN) supported in part by National Institutes of Health grant KY-INBRE P20 RR16481.

Conflicts of interest

SJ, IK and GPS are employees of Illumina and own stock in the company. The remaining authors declare no conflicts of interest.

References


Supporting information

Additional supporting information may be found in the online version of this article.

**Figure S1** Hypothetical examples showing the derivation of consensus structures at (a) loci with primary RNA-seq data and supporting *in silico* predictions and (b) loci supported only by *in silico* predictions.

**Figure S2** Comparison of tag alignments between genomic regions annotated as protein-coding genes, pseudogenes, or intergenic.

**Figure S3** Positional comparison of RNA-seq data to 3′ equine ESTs.

**Table S1** Ensembl genes with additional exons suggested by RNA-seq data.

**Table S2** Ensembl genes with extended exons suggested by RNA-seq data.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
IgE, IgGa, IgGb and IgG(T) serum antibody levels in offspring of two sires affected with equine recurrent airway obstruction

A. Scharrenberg*, V. Gerber*, J. E. Swinburne†, A. D. Wilson‡, J. Klukowska-Rötzler§, E. Laumen* and E. Marti§

Department of Clinical Veterinary Medicine, Vetsuisse Faculty, University of Berne, 3001 Berne, Switzerland. †Animal Health Trust, Lanwades Park, Kentford, Newmarket, Suffolk CB8 7UU, UK. ‡Division of Pathology Infection and Immunity, School of Clinical Veterinary Science, University of Bristol, Langford House, Langford BS40 5DU, UK. §Department of Clinical Research VPH, Vetsuisse Faculty, University of Berne, 3001 Berne, Switzerland.

Summary

Equine recurrent airway obstruction (RAO) is a chronic lower airway disease of the horse caused by hypersensitivity reactions to inhaled stable dust, including mould spores such as Aspergillus fumigatus. The goals of this study were to investigate whether total serum IgE levels and allergen-specific IgE and IgG subclasses are influenced by genetic factors and/or RAO and whether quantitative trait loci (QTL) could be identified for these parameters. The offspring of two RAO-affected sires (S1: n = 56 and S2: n = 65) were grouped by stallion and disease status, and total serum IgE levels and specific IgE, IgGa, IgGb and IgG(T) levels against recombinant Aspergillus fumigatus 7 (rAspf7) were measured by ELISA. A panel of 315 microsatellite markers covering the 31 equine autosomes were used to genotype the stallions and their offspring. A whole-genome scan using half-sib regression interval mapping was performed for each of the IgG and IgE subclasses. There was no significant effect of disease status or sire on total IgE levels, but there was a significant effect of gender and age. rAspf7-specific IgGa levels were significantly higher in RAO-affected than in healthy horses. The offspring of S1 had significantly higher rAspf7-specific IgGa and IgE levels than those of S2. Five QTLs were significant chromosome-wide (P < 0.01). QTLs for rAspf7-specific IgGa and IgE were identified on ECA 1, for rAspf7-specific IgGa and IgGb on ECA 24 and for rAspf7 IgGa on ECA 26. These results provide evidence for effects of disease status and genetics on allergen-specific IgGa and IgE.

Keywords horse, serum antibody levels, genetics, QTL.

Introduction

Recurrent airway obstruction (RAO), or heaves, is one of the most common diseases of the lower respiratory tract of mature horses housed in a conventional stable environment and is caused by hypersensitivity reactions to inhaled stable dust, including mould spores, such as Aspergillus fumigatus (Robinson et al. 1996; Robinson 2001). A genetic predisposition for this asthma-like disease has been demonstrated (Martí et al. 1991; Ramseyer et al. 2007; Gerber et al. 2009).

IgE as well as IgG-mediated reactions have been suggested to be involved in RAO, although the immunological basis of RAO has still not been elucidated (Lawson et al. 1979; Asmundsson et al. 1983; Robinson 2001). Eder et al. (2000) showed that horses with RAO have higher serum IgE and IgG levels against pure recombinant Aspergillus fumigatus allergens than control animals. Furthermore, they also showed that the environment as well as genetic factors exerts significant effects on Aspergillus fumigatus-specific serum IgE levels, displaying a heritability of 0.3 (Eder et al. 2001). However, in this study, the clinical lung status of the horses was not known.

Ramseyer et al. (2007) investigated data from over 300 offspring of two severely RAO-affected Swiss Warmblood sires. In that study, a strong genetic predisposition to RAO was demonstrated, and QTLs associated with RAO were then identified (Swinburne et al. 2009). Furthermore, an association was found between the RAO phenotype and...
microsatellite markers around the interleukin 4 receptor gene (IL4R) in one family but not in the other (Jost et al. 2007). As IL4R is a positional and functional candidate gene for atopy and has been shown to be associated with total IgE levels in humans (Bottini et al. 2002), the first aim of our study was to investigate whether IL4R is associated with total IgE levels in the offspring of these two stallions. As total IgE levels are of limited use in detecting atopic predisposition in animals, where they are mainly influenced by the degree of parasitic infection (Pernthaner et al. 2005), we wanted to test whether genetic effects could be identified for allergen-specific IgE and IgG subclass levels in these two half-sibling groups. We selected rAspf7 as model antigen, because sire 1 clearly had higher IgE and IgG subclass levels against this antigen than sire 2, and because a previous study had shown that RAO-affected horses have higher IgE levels against rAspf7 than healthy horses (Eder et al. 2000).

The aims of our study were therefore to investigate whether total serum IgE levels and allergen-specific IgE and IgG subclass levels are influenced by genetic factors and/or RAO and whether quantitative trait loci (QTLs) for antibody levels can be identified within the families of these two sires.

**Material and methods**

**Horses**

The horses belong to the same two families described in Ramseyer et al.’s (2007) and Swinburne et al.’s (2009) study and were offspring of two RAO-affected unrelated Warmblood sires. The horses were graded with a horse owner-assessed respiratory signs index (HOARSI) and classified into HOARSI-1 (non-affected), -2 (mildly), -3 (moderately) and -4 (severely) affected with RAO (Ramseyer et al. 2007; Swinburne et al. 2009). Moderately and severely affected horses (HOARSI-3 and -4) were combined and were further classified as being in clinical exacerbation (HOARSI-3/4-E), i.e. they showed typical clinical signs of the disease, or in clinical remission (HOARSI-3/4-R), when they had a history indicating HOARSI-3/4, but did not show any clinical signs of the disease at the time of examination. As the degree of exposure to mouldy hay may influence mould-specific antibody levels, the group of horses in remission (i.e. not fed hay) was analysed separately from those in clinical exacerbation (i.e. fed hay).

Sera of 121 horses, 56 offspring of sire 1 and 65 of sire 2, as well as the sera of the sires themselves were available for our study. The number of horses categorized by sire and HOARSI group is given in Table 1. The age of the horses ranged from 7 to 14.

**Determination of total serum IgE levels**

Total serum IgE levels were determined with antigen-capture enzyme-linked immunoassay (ELISA) according to Wilson et al. (2006). Briefly, plates were coated by overnight incubation at 4 °C with 1 µg/ml monoclonal anti-IgE 1C12 in 0.05 M carbonate buffer pH 9.6. Following blocking with PBS, Tween 0.5%, BSA, sera (diluted 1:40 in blocking buffer) and standard samples were added to the plate, twofold serially diluted and incubated for 1.5 h at RT. A biotinylated anti-equine IgE mAb was added and incubated for 2 h, followed by incubation with extra-avidin alkaline phosphatase (Sigma-Aldrich, http://www.sigma-aldrich.com). The ELISA was developed with p-nitrophenyl phosphate. Plates were read when the OD 450 nm of the highest standard reached 1.00. Duplicate samples of six doubling dilutions of a horse serum with a known IgE concentration ranging from 128 to 4 ng/ml were used as standards. The OD of the test samples were interpolated from the standard curve to give their IgE concentration in ng/ml. The concentration of the original test samples was calculated as the average of the values from those dilutions where the concentration fell within the standard curve. All samples were measured on the same day. The samples were randomly distributed on seven plates.

**Determination of allergen-specific serum IgE, IgGa, IgGb and IgG(T) levels**

Allergen-specific serum IgE, IgGa, IgGb and IgG(T) levels against the recombinant Aspergillus fumigatus 7 allergen

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Number of offspring from sires 1 and 2 with recurrent airway obstruction (RAO), classified according to their lung status [horse owner-assessed respiratory signs index (HOARSI)] and their gender.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOARSI</td>
<td>Offspring of sire 1</td>
</tr>
<tr>
<td></td>
<td>Female</td>
</tr>
<tr>
<td>HOARSI-1</td>
<td>19</td>
</tr>
<tr>
<td>HOARSI-2</td>
<td>4</td>
</tr>
<tr>
<td>HOARSI-3/4-R</td>
<td>2</td>
</tr>
<tr>
<td>HOARSI-3/4-E</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
</tr>
</tbody>
</table>

HOARSI-1, healthy; HOARSI-2, mildly affected with RAO; HOARSI-3/4-R, moderately to severely affected with RAO in remission; HOARSI-3/4-E, moderately to severely affected with RAO in exacerbation.
(rAsp7; Kodzius et al. 2003) were also determined by ELISA. This allergen was selected because a preliminary screening had shown that sire 1 had higher IgE levels against this allergen than sire 2. Immulon 2 polystyrene microtiter plates (Dynatech Labs) were coated with 1 μg/ml rAsp7 in 0.2 M carbonate–bicarbonate buffer pH 9.6 (Pierce Biotechnology, http://www.piercenet.com) at 37 °C for 2 h. After washing the plates in PBS pH 7.4 (Calbiochem AG, http://www.calbiochem.com) containing 0.05% Tween, non-specific binding sites were blocked with blocking buffer (PBS pH 7.4, 5% milk powder, 5% Tween 20) for 30 min at 37 °C. All dilutions of sera and antibodies were made in blocking buffer. Following washing, horse sera were added at a dilution of 1:5 for IgE, 1:10 for IgGa, 1:25 for IgGb and 1:50 for IgG(T) levels, followed by four serial twofold dilutions in the plate. After an overnight incubation at 4 °C and washing, anti-equine IgE mAb 3H10 (0.5 μg/ml) or mouse anti-equine IgGa (1:1000; detecting equine IgG1 according to the new nomenclature: Lewis et al. 2008), IgGb (1:2000; detecting IgG4 and 7) or IgG(T) (1:2000; detecting IgG3 and IgG5) mAbs (AbD Serotec, http://www.ab-direct.com) were added, and the plates were incubated for 2 h at RT on a shaker. After washing, an alkaline phosphatase–conjugated affinity-purified goat anti-mouse IgGa antibody (1:2000; Jackson ImmunoResearch Inc., http://www.jacksonimmuno.com) was added. Following incubation at RT for 1.5 h while shaking, and then washing, the ELISA plates were developed with p-nitrophenyl phosphate (Sigma-Aldrich) in diethanolamine (Fluka Chemie AG) buffer, pH 9.6. Absorbance readings were measured at 405 nm with a Molecular Devices Reader (Molecular Devices) and converted to relative ELISA units (REU). For this purpose, a horse serum positive for the tested allergen (IgE or the different IgG subclasses, respectively) was used as reference serum and assigned a value of 100 REU. Serial dilutions of the reference serum were used to generate a standard curve, and the test sample results were calculated from the curve using an ELISA software program (SOFTmax® 2.31 for Windows™; Molecular Devices Co.).

Transformation of data and statistical analyses

The data were not normally distributed, and therefore, all values were log-transformed. Values below the detection limit of the assay were replaced by a value of 0.1 REU. Statistical analyses were carried out with the statistical software package NCSS 2001 (NCSS).

An analysis of variance (ANOVA) routine was used. First, we fitted a model that included the effects of the sire (fixed), HOARSI (fixed), gender (fixed), age at blood sampling (covariate) and all possible two-way interactions between these factors. In a second step, the same analysis was repeated, but this time, all non-significant interactions were omitted from the model. Correlations between age and total or allergen-specific IgE or allergen-specific IgG subclasses were calculated with the Spearman correlation coefficient. P values ≤0.05 were considered statistically significant.

Genotyping

All 123 horses had been genotyped previously with 315 microsatellite markers, and this is described in detail elsewhere (Swinburne et al. 2009).

QTL analysis

A whole-genome scan using half-sib regression interval mapping was performed with the QTLEXPRESS software (http://qtl.cap.ed.ac.uk/; Seaton et al. 2002; Knott et al. 1996). Five phenotypic traits were included in the analysis: log of total IgE and of rAsp7-specific IgE, IgGa, IgGb and IgG(T). The level of roughage feeding, bedding type and gender were included as fixed effects, and age as a covariate. A single QTL model was fitted at 1-Mb steps along the chromosome, and chromosome-wide significance thresholds for the resulting F-statistics were obtained using 5000 random permutations of the data. The analysis was performed for each half-sibling group separately because of observed genetic heterogeneity between the two families (Jost et al. 2007; Swinburne et al. 2009).

Results

Effects of sire, HOARSI, gender and age on serum antibody levels

Comparison of both sires revealed that sire 1 had ten times higher total serum IgE levels than sire 2 and between 1.5 and 5 times higher rAsp7-specific IgE, and IgG subclass antibody levels against rAsp7 than sire 2.

The mean total serum IgE level in the horse sera was 7291 ng/ml and ranged from 20 ng/ml to 43 μg/ml. The ANOVA showed that ‘HOARSI’ and ‘sire’ exerted no significant effects on total IgE levels. However, mares had significantly higher mean log IgE levels than geldings. Additionally, total IgE levels were significantly influenced by the age of the horse (Table 2). The age of the horse was negatively correlated with total IgE levels (Spearman rank correlation coefficient = −0.28; P < 0.01).

The sire exerted a significant effect on rAsp7-specific IgE levels, while HOARSI, gender and age did not (Table 2). Independently of the HOARSI, the offspring of sire 1, which had himself higher rAsp7-specific IgE than sire 2, had significantly higher mean IgE levels against this antigen than the offspring of sire 2.

Significant HOARSI (P < 0.001) and sire (P < 0.01) effects on rAsp7-specific IgGa levels were found (Table 2). The offspring of sire 1 had significantly higher rAsp7-specific IgGa levels than the offspring of sire 2, and HOARSI-3/
R and E horses had higher IgGa levels against this antigen than the HOARSI-1 control horses.

No significant effects of HOARSI, sire, age and gender on allergen-specific IgGb levels could be demonstrated (Table 2). The ANOVA revealed a significant interaction between HOARSI and gender (P = 0.04) on rAspf7-specific IgG(T) levels. These antibody levels were significantly influenced by gender (P = 0.04) and age (P = 0.03). No significant effects of HOARSI and sire on rAspf7-specific IgG(T) levels could be demonstrated (Table 2).

QTL analysis

Evidence for QTLs affecting total IgE and specific IgE and IgG subclasses was found on different equine chromosomes (ECA) as summarized in Table 3. Five QTLs were significant chromosome-wide at P < 0.01 (Table 3). In the family of sire 1, a QTL for rAspf7-specific IgGa was identified on ECA 1 (Fig. 1). This was the only QTL that almost reached genome-wide significance at P < 0.05. An additional QTL for rAspf7-specific IgE was also identified on this chromosome in the same family. In the family of sire 2, two QTLs influencing rAspf7-specific IgGa and IgGb were found on ECA 24 (Fig. 1), and one QTL for rAspf7 IgGa was also found on ECA 26 (Fig. 1). Additional QTLs significant chromosome-wide at P < 0.05 are listed in Table 3, including two QTLs on ECA 13 in the family of sire 1 (Fig. 1). This QTL overlaps the QTL for HOARSI described for this same family in Swinburne et al.’s (2009) study. All QTLs were only significant in one or the other family.

Discussion

The presence of a genetic predisposition for RAO is now well documented (Marti et al. 1991; Ramseyer et al. 2007). Because RAO has similarities to human asthma, genes that are associated with human asthma are good candidates for RAO susceptibility. In a recent study, Jost et al. (2007) investigated whether the region of ECA 13 that carries IL4R, which is one of the genes associated with asthma susceptibility in humans (Ober et al. 2000), is linked to the RAO phenotype in the horse. They found that microsatellites closely linked to this gene were associated with RAO in the offspring of one RAO-affected sire (sire 1) but not in the offspring of another (sire 2). As IL4R is also associated with total IgE levels in humans (Bottini et al. 2002), we tested whether this could also be the case in these families. We could not identify an association between IL4R and total IgE levels within the offspring of sire 1 or sire 2 (data not shown), nor find differences between offspring of sire 1 or sire 2 in total IgE levels. Neither did the HOARSI have a significant effect on total IgE levels (Table 2), probably because total IgE levels are mainly influenced by the degree of infection with endoparasites (Pernthaner et al. 2005). This may also explain why total IgE levels decrease with age.

Table 2: Mean total IgE and IgE, IgGa, IgGb and IgG(T) ab levels (Log10) against recombinant Aspergillus fumigatus 7 (rAspf7) in sera of 121 offspring of two recurrent airway obstruction (RAO)-affected sires grouped by horse owner-assessed respiratory signs index (HOARSI), sire, gender and age.

<table>
<thead>
<tr>
<th></th>
<th>Log antibody levels – mean (standard error)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HOARSI</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Total IgE</td>
<td>3.790</td>
</tr>
<tr>
<td>IgE rAspf7</td>
<td>1.266</td>
</tr>
<tr>
<td>IgGa rAspf7</td>
<td>1.754</td>
</tr>
<tr>
<td>IgGb rAspf7</td>
<td>1.960</td>
</tr>
<tr>
<td>IgG(T) rAspf7</td>
<td>0.823</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ns, not significant.
Table 3  QTL significant at the chromosome-wide level in the half-siblings of the two sires S1 and S2, for total IgE levels and for recombinant *Aspergillus fumigatus* 7 (rAspf7)-specific IgE and IgG subclasses. The analysis was performed for the families of each sire separately.

<table>
<thead>
<tr>
<th>ECA</th>
<th>Trait</th>
<th>Family of:</th>
<th>Position of peak of signal (Mb)</th>
<th>F-statistic</th>
<th>Chromosome 5% threshold</th>
<th>Chromosome 1% threshold</th>
<th>Genome significance 5% threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Log rAspf7 IgGa</td>
<td>Sire 1</td>
<td>127</td>
<td>13.31**</td>
<td>7.86</td>
<td>11.87</td>
<td>13.53</td>
</tr>
<tr>
<td>1</td>
<td>Log rAspf7 IgE</td>
<td>Sire 1</td>
<td>141</td>
<td>11.87**</td>
<td>7.68</td>
<td>11.23</td>
<td>13.69</td>
</tr>
<tr>
<td>2</td>
<td>Log rAspf7 IgGa</td>
<td>Sire 1</td>
<td>28</td>
<td>8.64 *</td>
<td>6.86</td>
<td>10.47</td>
<td>13.53</td>
</tr>
<tr>
<td>2</td>
<td>Log total IgE</td>
<td>Sire 1</td>
<td>116</td>
<td>7.54*</td>
<td>6.64</td>
<td>10.05</td>
<td>14.61</td>
</tr>
<tr>
<td>4</td>
<td>Log total IgE</td>
<td>Sire 1</td>
<td>57</td>
<td>9.41*</td>
<td>6.81</td>
<td>10.30</td>
<td>14.61</td>
</tr>
<tr>
<td>5</td>
<td>Log rAspf7 IgG(T)</td>
<td>Sire 2</td>
<td>29</td>
<td>9.60*</td>
<td>6.92</td>
<td>10.98</td>
<td>14.43</td>
</tr>
<tr>
<td>6</td>
<td>Log rAspf7 IgE</td>
<td>Sire 2</td>
<td>29</td>
<td>5.30*</td>
<td>5.16</td>
<td>7.56</td>
<td>10.69</td>
</tr>
<tr>
<td>10</td>
<td>Log rAspf7 IgE</td>
<td>Sire 2</td>
<td>73</td>
<td>6.47*</td>
<td>5.86</td>
<td>8.60</td>
<td>10.69</td>
</tr>
<tr>
<td>13</td>
<td>Log rAspf7 IgE</td>
<td>Sire 1</td>
<td>26</td>
<td>6.92*</td>
<td>6.14</td>
<td>10.16</td>
<td>13.69</td>
</tr>
<tr>
<td>13</td>
<td>Log rAspf7 IgGa</td>
<td>Sire 1</td>
<td>29</td>
<td>7.57*</td>
<td>6.33</td>
<td>9.69</td>
<td>13.53</td>
</tr>
<tr>
<td>16</td>
<td>Log rAspf7 IgG(T)</td>
<td>Sire 2</td>
<td>18</td>
<td>7.87*</td>
<td>6.53</td>
<td>10.47</td>
<td>14.43</td>
</tr>
<tr>
<td>23</td>
<td>Log rAspf7 IgGb</td>
<td>Sire 1</td>
<td>47</td>
<td>5.50*</td>
<td>5.20</td>
<td>8.27</td>
<td>14.43</td>
</tr>
<tr>
<td>23</td>
<td>Log rAspf7 IgG(T)</td>
<td>Sire 1</td>
<td>16</td>
<td>10.14**</td>
<td>5.54</td>
<td>9.61</td>
<td>14.67</td>
</tr>
<tr>
<td>24</td>
<td>Log rAspf7 IgGb</td>
<td>Sire 2</td>
<td>26</td>
<td>10.24**</td>
<td>5.57</td>
<td>8.82</td>
<td>14.62</td>
</tr>
<tr>
<td>26</td>
<td>Log rAspf7 IgGa</td>
<td>Sire 2</td>
<td>24</td>
<td>11.09**</td>
<td>5.37</td>
<td>8.51</td>
<td>14.67</td>
</tr>
<tr>
<td>26</td>
<td>Log rAspf7 IgGb</td>
<td>Sire 1</td>
<td>30</td>
<td>5.98*</td>
<td>4.85</td>
<td>7.20</td>
<td>10.69</td>
</tr>
<tr>
<td>27</td>
<td>Log total IgE</td>
<td>Sire 2</td>
<td>20</td>
<td>8.49*</td>
<td>5.84</td>
<td>9.42</td>
<td>14.03</td>
</tr>
<tr>
<td>29</td>
<td>Log rAspf7 IgE</td>
<td>Sire 2</td>
<td>18</td>
<td>5.30*</td>
<td>4.89</td>
<td>8.20</td>
<td>13.69</td>
</tr>
<tr>
<td>31</td>
<td>Log total IgE</td>
<td>Sire 1</td>
<td>6</td>
<td>8.44*</td>
<td>4.42</td>
<td>7.97</td>
<td>14.61</td>
</tr>
</tbody>
</table>

** 1% chromosome-wide significance.

* 5% chromosome-wide significance.

Figure 1  F-statistic plots for significant antibody phenotypes on ECA 1, ECA 13, ECA 24 and ECA 26. Non-broken lines show plots for log rAspf7-specific IgE. Regularly broken lines show plot for log rAspf7-specific IgGa, and the irregularly broken line shows plot for log rAspf7-specific IgGb. The positions of informative markers for each respective family are shown with grey triangles. The position of genes referred to in the text is indicated.

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 131–137
because younger horses are more heavily infected with gastrointestinal parasites. Additionally, mares had significantly higher total IgE levels than geldings (Table 2). Interestingly, similar effects of age and sex on total IgE levels have been observed in a colony of beagle dogs (Racine et al. 1999).

We then tested whether genetic effects could have an influence on allergen-specific IgE and IgG levels. We found a statistically significant effect of the sire on rAspf7-specific IgE (Table 2), but IL4R showed no association with rAspf7-specific IgE levels (data not shown). However, the QTL analysis within the family of sire 1 showed significant signals (chromosome-wide \( P < 0.05 \), Fig. 1 and Table 3) on ECA 13 for rAspf7-specific IgE, peaking at 26 Mb. This was similar to the peak obtained for HOARSI in this same family (Swinburne et al. 2009). Although the \( F \)-statistic plot for rAspf7-specific IgE had a similar shape to the plot for HOARSI, the chromosomal region where IL4R is located (20 Mb) did not reach statistical significance for rAspf7-specific IgE, confirming that IL4R is most probably not the gene influencing IgE levels in this family. The QTL on ECA 13 for rAspf7-specific IgE had a lower significance level than the one for HOARSI (genome-wide, \( P < 0.05 \)), but the difference may be attributable to the smaller number of horses used in our study; serum was only available from 56 of the 130 offspring of sire 1 used by Swinburne et al. (2009). This finding suggests that, in this family, IgE against rAspf7 may indeed be one of the predisposing factors for RAO. QTLs with higher significance levels (\( P < 0.01 \), chromosome-wide) were found on three chromosomes: On ECA 1, a QTL which nearly reached genome-wide significance was identified for rAspf7-specific IgGa in the family of sire 1. This region contains one obvious candidate gene, SMAD3 (Fig. 1). SMAD proteins mediate TGFβ signalling, and TGFβ, together with IL-10, suppresses the production of IgE and induces non-inflammatory IgG isotypes (Åkdis 2006). A recent study showed that Smad3−/− mice have a severely impaired antigen-specific IgG2a response (Anthoni et al. 2010), demonstrating that this gene does indeed influence the production of certain IgG isotypes against specific antigens and is thus an interesting candidate for specific IgGa levels in the horse. A QTL for rAspf7-specific IgE, peaking at 141 Mb, was also found on ECA 1 in the same family. An interesting candidate gene in this area encodes GA-binding protein transcription factor beta subunit 1 (GABPB1). This transcription factor controls the induction of the IL-16 promoter in T lymphocytes (Bannert et al. 1999), and IL-16 inhibits IgE production in B lymphocytes (Trudelle et al. 2007). Most interestingly, in the family of sire 2, a QTL for rAspf7-specific IgGa was identified on ECA 26, and GABPA maps to this chromosome. Another candidate gene on this chromosome which could influence IgG isotypes is IL10RB. Finally, in the family of sire 2, two QTLs for rAspf7-specific IgGa (i.e. IgG1; Lewis et al. 2008) and IgGb were identified on ECA 24. Interestingly, IGHC1 maps to ECA 24 (Wagner et al. 2004) at 0.55 Mb as shown by BLAST alignment. Additionally, a marker associated with IgE levels was described in humans on a chromosomal region (14q13-24; Mansur et al. 2004) corresponding to ECA 24. However, additional polymorphic markers first need to be typed in the regions of interest on ECA 24 and ECA 26, as data from only five and three informative markers, respectively, were available for family 2 on these chromosomes. All identified QTLs were only significant in one or the other family. A possible explanation could be that many different genes influence antibody production and that QTLs which segregate in one family are homozygous (and therefore not detected) in the other sire.

In conclusion, our study shows that while total IgE levels are neither influenced by the HOARSI phenotype nor by the sire, IgE and IgGa antibody levels against specific allergens, such as rAspf7, seem to be influenced by disease status and genetic factors. The offspring of stallion 1, similar to their sire, had significantly higher IgE and IgGa levels against rAspf7 than the offspring of sire 2, and these were independent from environmental factors. Furthermore, some chromosomal regions showing linkage to the humoral immune response have been identified. Further studies to better characterize these regions are warranted.

Acknowledgements

We are grateful to Prof. R. Cramer, Swiss Institute of Allergy and Asthma Research, Davos, for providing rAspf7 and to Prof. Dr M. Doherr, Department of Clinical Research, for his advice on the statistical analyses. This work was supported by the Swiss National Science Foundation grants No. 31-116803/1 and 31000 – 116502/1 and by a grant of the Department of Clinical Veterinary Medicine, Vetsuisse Faculty, University of Bern, Switzerland. The genotyping work was supported by a Horse Trust (UK) grant No. G1306.

Conflicts of interest

The authors have not declared any potential conflicts.

References


Genomic structure, polymorphism and expression of ACCN1 and ACCN3 genes in the horse

X. Mata, A. Ducasse, A. Vaiman, M. Diribarne, A.-S. Fraud and G. Guérin
INRA, UMR1313, Unité de Génétique Animale et Biologie Intégrative, Centre de Recherche de Jouy, 78350 Jouy-en-Josas, France

Summary

A category of cation gate proteins was shown to be present in sensory neurons and act as receptors of protons present in tissues such as muscles. The Amiloride-sensitive Cation Channel, Neuronal (ACCN) gene family is known to play a role in the transmission of pain through specialized pH sensitive neurons. Muscles from horses submitted to strenuous exercises produce lactic acid, which may induce variable pain through ACCN differential properties. The sequences of the equine cDNAs were determined to be 2.6 kb in length with an open reading frame of 1539 bp for ACCN1 and 2.1 kb in length with an open reading frame of 1602 bp for ACCN3. The ACCN1 gene is 990 kb long and contains 10 exons, and the ACCN3 gene is 4.2 kb long and contains 11 exons. The equine ACCN1 and ACCN3 genes have an ubiquitous expression but ACCN1 is more highly expressed in the spinal cord. We identified one alternative ACCN3 splicing variant present in various equine tissues. These mRNA variants may encode two different protein isoforms 533 and 509 amino acids long. Ten single nucleotide polymorphisms (SNPs) were detected for ACCN1: five in the coding and five in the non-coding region, with no amino acid change, while the three SNPs identified in the coding region of the ACCN3 gene introduce amino acid changes. The equine in silico promoter sequence reveals a structure similar to those of other mammalian species, especially for the ACCN1 gene.

Keywords ACCN1, ACCN3, expression, gene, horse, polymorphism.

Introduction

Pain is caused by multiple factors such as mechanical, thermal or chemical stimuli that are processed by specialized neurons classified according to their characteristics. Those of the cation gates triggering the nerve impulse by depolarization of the cell membrane are particularly involved. The signal originates in the terminal part of neurons because of tissue damage and is conveyed to the brain, where it is interpreted as an uncomfortable perception. But pain is also a source of information for the living organism to react and protect tissue or body from alteration or destruction. The existence of interindividual differences in pain perception suggests the involvement of genetic factors (Mogil 1999, 2004). Pain research has undergone major evolution at the system, cellular, subcellular and molecular levels. One source of chemical stimulation triggering pain is the high concentration of hydrogen ions, which lowering the tissue pH in the cellular environment, and this can be due in part to lactic acid secretion.

A category of cation gates was shown to be present in sensory neurons and act as receptors of acidity. The Amiloride-sensitive Cation Channel, Neuronal (ACCN, also known as Acid Sensing Ion Channel, ASIC) gene family probably plays an important role in the transmission of pain through specialized pH sensitive neurons (for review see Kellenberger & Schild 2002; Mamet & Voilley 2002). The first gene of the family was cloned in a neurodegenerative mutant of Caenorhabditis elegans and was thus called degenerin (Driscoll & Chalfie 1992). To date, several other members of this degenerin/epithelial Na+ channel superfamily have been identified and cloned in mammals. The ACCN genes are differentially distributed in the central and/or peripheral nervous system, and four of them, ACCN1 (ASIC2) (Lingueglia et al. 1997), ACCN2 (Waldmann et al. 1997b; Chen et al. 1998), ACCN3 (ASIC3) (Waldmann et al. 1997a) and ACCN4 (Grunder et al. 2000), have been cloned. In addition, two isoforms are generated by alternative splicing for ACCN2 (Waldmann et al. 1997b; Ugawa...
et al. 2001) and three for ACCN3. Two initiation sites in ACCN1 trigger the alternative transcription of exon 1a or exon 1b in humans (Lingueglia et al. 1997) and in rats.

Amiloride-sensitive Cation Channel. Neuronal proteins have two transmembrane domains (TM1 and TM2), a large extracellular loop and the N and C termini in the intracellular space (Chen et al. 1998). The proteins associate into homo- or heterotetramers to form cation transmembrane channels with differences in electrophysiological properties. One of these is lactic acid, also released during heart pain (Inmike & McCleskey 2001a,b) and in intense muscle exercise periods, which is a situation encountered in human runners or swimmers when sprinting (Naves & McCleskey 2005). This should also apply to racing animals, such as horses and dogs, when cardiac and skeletal muscles turn to anaerobic metabolism at when the sustained speed is high. The characteristics of ACCN3 make it a good candidate for acidosis-related muscle pain sensitivity. Expression patterns and physical properties of the ACCN3 channel in rats identify ACCN3 as the detector of lactic acid (Naves & McCleskey 2005; Yagi et al. 2006), although it is also activated by other stimuli (Molliver et al. 2005). ACCN3 is specific to sensory neurons and is expressed in dorsal root ganglia of the DRG and absent in the brain (Waldmann et al. 1997a). Knockout mice for ACCN3 experience alterations in sensory perception, and this gene is postulated to modulate moderate- to high-intensity pain sensation (Price et al. 2001; Chen et al. 2002).

Similarly, ACCN1 mutations in the MDEG1 gene, which is an ACCN1 var 2 mammalian isoform, modify pH dependence, inactivation kinetics and amiloride sensitivity of this channel (Champigny et al. 1998). ACCN1 isoform 1 modulates the kinetics and pH sensitivity of ACCN2 and ACCN3 (Lingueglia et al. 1997; Yagi et al. 2006), and genetic alterations may also contribute to sensory variability. Mutations changing amino acids in the first pre-transmembrane region of ACCN1 and ACCN3 have been shown to affect the Na+/Ca+ selectivity and pH dependence of the ion channels (Coscøy et al. 1999).

Much work remains to be performed to elucidate the actual mechanisms by which the acid-gated channels function. However, when all the evidence is taken together, it appears that ACCN3 might be the preferential candidate for ischaemic pain, with a possible modulation of its functional properties by ACCN1.

In horses, the main cause of metabolic acidosis observed during both high-intensity and short-distance racing comes from anaerobic oxidation of glucose to lactic acid. In addition, plasma volume decreases because water moves from the vasculature to the intracellular and interstitial spaces at the onset of intense exercise (Hyypää & Poso 1998). Thus, one may assume that lactic acidosis hampers horse performance in short-distance flat racing and that ACCN functional heterogeneity can influence this individual variability. Here, we have investigated the genomic structure of the ACCN1 and ACCN3 genes in the horse and have explored expression in different tissues. Transcript structures were also analysed for single nucleotide polymorphism (SNP) identification in different breeds and for alternative splicing.

Materials and methods

Gene structure analysis

Total RNAs were extracted from different tissues using the RNA Now procedure (Biognex). Reverse transcriptions (RT) were performed on 5 μg of total RNA using the Superscript First Strand Synthesis System (Invitrogen) following the manufacturer’s instructions. RT on horse RNAs were performed using oligonucleotides (dT)-18 followed by PCR using the GoTaq® Flexi DNA Polymerase and reaction buffer (Promega).

The 3’ ends of the horse ACCN1 and ACCN3 cDNA were isolated using 3’RACE kit (Invitrogen). The 5’ ends of the horse ACCN3 cDNA were also cloned using 5’RACE kit (Invitrogen). We cloned and sequenced the most expected 5’ end by comparison with other species using proximal contiguous forward primers.

The cDNA sequencing of ACCN1 and ACCN3 was performed after RT-PCR amplification using respectively four and three primer pairs designed to amplify overlapping fragments 500–600 bp long covering all the exons of the genes (Table 1). All oligonucleotides of this study were designed using the primer 3 software (http://frodo.wi.mit.edu/primer3/). The PCR amplification reactions were optimized and carried out in a PTC-100 (MJ-Research) using the following cycling conditions: 94 °C for 5 min followed by 35 cycles of (94 °C for 30 s, annealing temperatures (Table 1) for 30 s, 72 °C for 2 min) and 72 °C for 5 min. The resulting PCR products were separated on a 2% agarose gel, purified using Wizard SV Gel and PCR Clean-Up System (Promega) and sequenced by Qiagen sequencing services. The resulting sequences were compared to the horse genome sequence in the NCBI database by means of the BLASTN software (http://www.ncbi.nlm.nih.gov) to deduce the intron/exon structure of the ACCN1 and ACCN3 genes. All ACCN1 and ACCN3 genomic and cDNA sequences generated during the course of this study have been submitted to GenBank databases (HM467835, HM467836, HM467837, HM567408 and HM567409).

Polymorphisms

All exons of the ACCN1 and ACCN3 genes were amplified using primers in the introns on both sides of each exon. The PCRs were performed on genomic DNA extracted from the blood of a panel of 15 horses among four breeds (French trotter, French saddlebred, Thoroughbred and Norman cob) using the Genisol™ Maxiprep kit (ABgene)
According to the manufacturer’s instructions, PCRs were optimized and DNA products were analysed as for the gene structure analysis. The primers used are listed in Table 1. The resulting sequences were introduced into the NOVOSNP software (Weckx et al. 2005) to find SNPs in the amplified sequences.

Alternative splicing and expression of the ACCN1 and ACCN3 gene

RNAs were isolated from different tissues (skeletal muscle, heart, liver, spleen, lung, kidney and spinal cord) in one adult healthy horse. Semi-quantitative RT-PCR was conducted on these RNA tissue samples from five adult horses for ACCN1 and ACCN3 mRNAs. RT-PCR amplifications for splice variant determination were performed on several tissue RNA samples (skeletal muscle, heart, liver, and spinal cord). The primers used are presented in Table 1. PCRs were optimized and DNA products were analysed as for the gene structure analysis. The results of semi-quantitative RT-PCR were normalized with the GAPDH horse gene. The PCR amplifications used the following cycling conditions: 94 °C for 5 min followed by 28 cycles of (94 °C for 30 s, annealing for 30 s (temperatures in Table 1), 72 °C for 2 min) and 72 °C for 5 min.

A comparison of the tissue expression of the two ACCN mRNAs was made with the non-parametric Kruskal–Wallis one-way analysis of variance (Kruskal & Wallis 1952). It is an extension of the Mann–Whitney U-test for three or more groups. The test does not assume a normal distribution population but does assume an identically shaped and scaled distribution for each group, except for any difference in medians.

### Promoter sequence analysis

The MatInspector (Search transcription factor binding sites) program, part of the GenomatixSuite (http://www.genomatix.de/en/produkte/genomatix-software-suite.html), was used for in silico proximal promoter analysis. About 1 kb upstream of the transcription start site (TSS) of ACCN1 was selected in the human, equine, canine, murine, and bovine species. For ACCN3, the sequences came from the Genomatix database. Stringent analysis parameters (matrix similar-

### Table 1

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence</th>
<th>Annealing temperature (°C)</th>
<th>Exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCN1_cDNA5'R</td>
<td>CTCTTCCACATACCCCTGGTGC</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_cDNA1'R</td>
<td>CACTACCCCTGGTGC</td>
<td>60</td>
<td>1a–2</td>
</tr>
<tr>
<td>ACCN1_cDNA2'R</td>
<td>GGAACTTTGACATATGTTA</td>
<td>60</td>
<td>2–6</td>
</tr>
<tr>
<td>ACCN1_cDNA3'R</td>
<td>CTCTGAATTCGCCCAAGGAC</td>
<td>60</td>
<td>6–10</td>
</tr>
<tr>
<td>ACCN1_cDNA4'R</td>
<td>CGGGCAGCTGTCATGAGAC</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>ACCN1_cDNA3'R</td>
<td>GGGCTTTATTTGTGCTTC</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>5'RACE Oligo</td>
<td>AAGACGGCTAGCCATGCTGTCCG</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>ACCN3_cDNA1'R</td>
<td>CCTCTAGGAGTGAAGGAGAG</td>
<td>60</td>
<td>1–2</td>
</tr>
<tr>
<td>ACCN3_cDNA2'R</td>
<td>CTTCACCTGATCCACG</td>
<td>60</td>
<td>2–7</td>
</tr>
<tr>
<td>ACCN3_cDNA3'R</td>
<td>GGCAAGACATTGTCAGGC</td>
<td>60</td>
<td>6–11</td>
</tr>
<tr>
<td>ACCN3_cDNA4'R</td>
<td>3'RACE Oligo</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext1F</td>
<td>ATTGCCTAAAGCTCTCTCT</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext2F</td>
<td>ATCTTGGCAACACCTCAAC</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext3F</td>
<td>GATTTCTGGGTCTCACAA</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext4F</td>
<td>ATTGGAGCCTAAATGTTA</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext5F</td>
<td>CACAGACGACGACGAC</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext6F</td>
<td>CTCCACGACGACGAC</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext7F</td>
<td>ACTCATGACCTCTGTGAC</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext8F</td>
<td>CTCCTTCCTTTCAGAGGTC</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext9F</td>
<td>CTTTCTGACCCAGACG</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN1_Ext10F</td>
<td>TGAATTTTGCCCAATGTTA</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN3_Ext1F</td>
<td>CAACCCAGGCTCTTATGC</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>ACCN3_Ext2-3F</td>
<td>GTTCTCTGGTGGGTGGTGT</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>ACCN3_Ext4F</td>
<td>CTTCACCAAGCTCTCTCT</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN3_Ext5F</td>
<td>GTTACCGGCTGCTGTGAC</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN3_Ext6F</td>
<td>CAGACCTCCCCTGTGAC</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN3_Ext7-9F</td>
<td>CTCTCGGCTCCTCTCCTCT</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>ACCN3_Ext10-11F</td>
<td>CTCCTTCCTTGCACCTCTCTCT</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 138–144
ity = 0.05) giving a score >0.80 were used to select the most likely conserved transcription factors. The EMBOSS CpGPlot (http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html) was used to detect CpG islands on the same sequences.

### Results

#### Sequence of the ACCN1 and ACCN3 cDNAs

The length of the ACCN1 ORF is 1539 nucleotides, with a predicted protein of 512 amino acids. Sequence analysis of this equine cDNA shows an identity of more than 92% compared to those of other mammalian species (Table 2 and Appendix S1). The ACCN3 ORF length is 1602 nucleotides, producing an expected 534 amino acid protein. The sequence identity between the equine ACCN3 cDNAs and other mammalian species is more than 83% (Table 2 and Appendix S2).

Sequencing clones from the 3′ ends of these two cDNAs does not reveal any multiple polyadenylation sites. Cloning and sequencing the 5′ extremity of ACCN1 and ACCN3 do not present any alternative initiation transcription sites. Despite the use of several 5′ primers designed in the horse, we were unable to amplify the exon homologous to the human ACCN1 var 1 by RT-PCR.

ACCN1 and ACCN3 gene and ORF lengths and sequence identities compared between species and results are reported in Table 2. The ACCN1 gene structure is highly conserved, with the same length of each coding exon between species. For the ACCN3 gene, the situation is more complex, with small differences in the ORF length. There are two species that have one less amino acid (human and bovine) and one species that has three amino acids less (mouse), while the dog has a protein two amino acids longer. These differences within the coding sequence are principally localized in exon 4 of the human and mouse sequences. In the bovine and canine, sequence differences take place in exons 4 and 9 for the bovine sequence and in exons 5 and 6 for the canine sequence in comparison with the horse genome.

#### Structure of the ACCN1 and ACCN3 genes

The precise positions of the intron–exon junctions were deduced from the sequencing of the equine ACCN1 (ECA11) and ACCN3 (ECA4) mRNA sequences and the whole genomic sequence (Ecab2). The ten exons of the ACCN1 gene were identified (with the exception of the human exon 1b) as were the 11 exons of the ACCN3 gene. Each junction was delimited by the GT – AG crucial sequence at the beginning and the end of each exon (HM567408 and HM567409).

#### Polymorphism of ACCN1 and ACCN3 genes

Single nucleotide polymorphisms were located in the coding region of the cDNA of both genes by sequencing. All exons were sequenced on a panel of horses from different breeds. Ten SNPs were detected for ACCN1, five in the coding region and five in the non-coding region (Table 3). The SNPs in the coding region do not introduce any amino acid

### Table 2 Length of ACCN1 and ACCN3 5′ and 3′UTR and ORFs, and percent identities in different species.

<table>
<thead>
<tr>
<th></th>
<th>Equus caballus</th>
<th>Homo sapiens</th>
<th>Mus musculus</th>
<th>Bos taurus (predicted)</th>
<th>Canis familiaris (predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5′UTR</td>
<td>274</td>
<td>274</td>
<td>92%</td>
<td>212</td>
<td>87%</td>
</tr>
<tr>
<td>ORF (nt)</td>
<td>1539</td>
<td>1539</td>
<td>95%</td>
<td>1539</td>
<td>92%</td>
</tr>
<tr>
<td>ORF (aa)</td>
<td>512</td>
<td>512</td>
<td>99%</td>
<td>512</td>
<td>99%</td>
</tr>
<tr>
<td>3′UTR</td>
<td>841</td>
<td>875</td>
<td>77%</td>
<td>892</td>
<td>73%</td>
</tr>
<tr>
<td>Total length</td>
<td>2654</td>
<td>2688</td>
<td></td>
<td>2643</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Equus caballus</th>
<th>Homo sapiens (var 2)</th>
<th>Mus musculus</th>
<th>Bos taurus (predicted)</th>
<th>Canis familiaris (predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5′UTR</td>
<td>407</td>
<td>368</td>
<td>60%</td>
<td>261</td>
<td>ND</td>
</tr>
<tr>
<td>ORF (nt)</td>
<td>1602</td>
<td>1596</td>
<td>88%</td>
<td>1593</td>
<td>83%</td>
</tr>
<tr>
<td>ORF (aa)</td>
<td>533</td>
<td>532</td>
<td>86%</td>
<td>530</td>
<td>85%</td>
</tr>
<tr>
<td>3′UTR</td>
<td>101</td>
<td>70</td>
<td>56%</td>
<td>88</td>
<td>63%</td>
</tr>
<tr>
<td>Total length</td>
<td>2110</td>
<td>2088</td>
<td></td>
<td>1942</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3 Position of single nucleotide polymorphisms (SNPs) on the cDNA sequences of ACCN1 (HM467836) and ACCN3 (HM467837).

<table>
<thead>
<tr>
<th>ACCN1</th>
<th>ACCN3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5′ UTR</td>
<td>c.-125G&gt;T</td>
</tr>
<tr>
<td></td>
<td>c.237G&gt;A</td>
</tr>
<tr>
<td></td>
<td>c.378G&gt;A</td>
</tr>
<tr>
<td></td>
<td>c.756C&gt;T</td>
</tr>
<tr>
<td></td>
<td>c.861G&gt;T</td>
</tr>
<tr>
<td></td>
<td>c.954G&gt;A</td>
</tr>
<tr>
<td>ORF</td>
<td>c.439G&gt;A</td>
</tr>
<tr>
<td></td>
<td>c.569A&gt;T</td>
</tr>
<tr>
<td></td>
<td>c.582C&gt;A</td>
</tr>
<tr>
<td>3′ UTR</td>
<td>c.*434G&gt;A</td>
</tr>
<tr>
<td></td>
<td>c.*606G&gt;A</td>
</tr>
<tr>
<td></td>
<td>c.*634T&gt;C</td>
</tr>
<tr>
<td></td>
<td>c.*709G&gt;A</td>
</tr>
</tbody>
</table>
changes. In contrast, all three SNPs identified in the coding region of the ACEN3 gene introduce changes: an Alanine to Threonine at position 439, a Lysine to Isoleucine at position 569 and finally an Asparagine into Lysine at position 582 of the protein sequence.

Alternative splicing of ACEN1 and ACEN3 genes

The expression of the human ACEN1 gene is reported to undergo a unique splicing involving only exons 1a and 1b. In fact, these two exons could not be found together in the transcripts of the same individual. In the horse muscle, RT-PCR could identify the homologous human exon 1a, but we were unable to identify exon 1b.

The situation is more diversified for ACEN3, where cDNA sequencing detects three alternative transcripts in muscle. One transcript contains all 11 exons, the second lacks the entire 72 bp exon 9 and the third contains an unspliced intron 8 (97 bp) in addition to the usual 11 exons. The lack of exon 9 in the second transcript does not change the frame, but it would be expected to shorten the protein by the small 24 amino acid domain encoded by exon 9. The last alternative transcript presents a shift of the open reading frame, with a new stop codon 96 bp further on, which would result in a longer protein of 597 amino acids with a different C-terminal. These expectations would only stand if these two cDNAs were translated, but we have no evidence that they do not come from an immature mRNA. We thus only report the exon 9 deletion in Fig. 1. No tissue-specific transcript was recognized in our RT-PCR amplifications of the tissues analysed.

ACEN1 and ACEN3 gene expression

Expression analysis of ACEN1 and three genes was performed on different tissues (skeletal muscle, heart, liver, spleen, lung, kidney and spinal cord) by semi-quantitative RT-PCR on five adult horse RNAs. The results showed significant differential expression of the ACEN1 gene in the spinal cord tissue with the Kruskal–Wallis test ($H = 18.578; P = 0.009$) (Fig. 2).

Promoter sequence analysis

No TATA or CAAT box could be identified in the promoters of the ACEN1 and ACEN3 genes. EMBOSS CpGPlot detected a 200–300 bp-long CpG island in ACEN1 of all five species. No CpG island could be found for ACEN3. Several potential transcription factor (TF) sites were detected in the promoter regions of the five species, but only those present in three of five species are reported (Fig. 3). For ACEN1, five TF sites were observed within 100 bp downstream of the TSS (SNAP, HNF6, HoxF, CHRF and FKHD) and one (CDXF) <50 bp upstream of the TSS. It is also possible that one of the two HNF1 sites is present in the five species but a gap in the canine sequence makes it only suggestive. The second HNF1 site is only present in horses and humans, while an E2FF site is present in three species. The TF sites are less conserved in the ACEN3 promoter because none are present simultaneously in the five species. Our criteria detected five types of transcription factor sites (GLIF, IRFF, HAML, HAND and NR2F), which were all different from those present in the ACEN1 promoter.

Discussion

The present study describes the structure of two horse genes that may be involved in pain production in horses submitted to high-intensity training. The horse ACEN1 and ACEN3 genes have the same number of exons as their vertebrate homologues, but we were unable to identify the horse counterpart of the human ACEN1 exon 1b. BLAST analysis of this human exon sequence finds a hit on the horse genome sequence in the 5’ region of the ACEN1 gene. Nevertheless, this human 1581 bp sequence shows 85% identity with the first 548 bp and more than 95% with the 357 last bp in the horse DNA, leaving a gap in the middle. This gap could not be detected in the equine traces database by directly blasting the corresponding human sequence or through blasting its two flanking sequences. It thus seems that this sequence is presently unresolved in the equine genome, appearing as an undetermined sequence and a 545 bp non-homologous sequence.
The ACCN1 gene showed a very high conservation between mammalian species at both the nucleotide and amino acid levels, with the same number of base pairs in the exons and the same protein length. In contrast, the ACCN3 equine protein is 1 (human) and 3 (murine) amino acids longer and 2 amino acids shorter than in canines. A careful analysis of the sequences of the ACCN1 and ACCN3 genes in different species shows that the length of introns and exons of the ACCN1 gene is very similar. The ACCN3 gene is more variable, especially in exon 4 (Table 2).

The three SNPs identified in the coding region of ACCN3 modify the amino acid sequence (p.Ala439Thr, p.Asn569Ile and p.Asn582Lys). All these amino acid substitutions change the polarity of the variants, as well as the charge being modified by the third SNP. It is not known whether these modifications affect the protein activity, but it should be noted that these mutations are not located in the two transmembrane domains nor in the glycosylation or phosphorylation sites when compared to the human protein (Human Protein Reference Database: http://www.hprd.org/).

Concerning alternative transcripts, the deletion of ACCN3 exon 9 removes a region between amino acids 465 and 488 that contains a potential phosphorylation site in position 478. Retention of intron 8 introduces a frameshift, modifying the C-terminal sequence from position 464 and deleting two potential phosphorylation sites in positions 478 and 493. A new stop codon appears at position 2199 followed by a polyadenylation site 1179 nt downstream. However, it is possible that this transcript represents an immature mRNA of this gene.

The existence of several alternative ACCN3 transcripts (exon skipping of exon 9 and retention of intron 8) is not surprising. Four types of alternative splicing (exon skipping, partial 3’ or 5’ exon alternative splicing and intron retention) have been described (Kim et al. 2007, 2008). Rearrangements such as alternative splicing may affect 40% of the genes in humans and chickens and 30% in rats and mice. Their distribution is around 35–45% for exon skipping, 20–25% for each 3’ or 5’ exon partial splicing and <4–10% for intron retention.

Amiloride-sensitive Cation Channel, Neuronal gene expression is well documented in the central and peripheral nervous system but knowledge is rather scarce for other tissues. In one such, it was shown that ACCN3 is expressed in 14 different tissues, while ACCN1 var 2 is only expressed in the nervous system (Babinski et al. 2000). In the horse, the expression studies have yielded similar results, although ACCN1 mRNA is more highly expressed in the spinal cord; it was also detected at low levels in other tissues. There is no possible interference with the expression of ACCN1 var 1, because this variant could not be detected in the horse.

The promoter sequence of the ACCN1 gene is well conserved in the 2–300 bp upstream of the TSS in the five mammalian species studied, highlighting the importance of this sequence in the regulation of the expression of this gene (Xia et al. 2003). However, it should be noted that the transcription factors identified in silico by MatInspector in this and other studies are different, which may be because of the ever growing knowledge of sequence–function relationship data, and thus the evolution of the transcription factor concept. The promoter sequences of the ACCN3 gene in the five species are less conserved, which explains the low level of conservation of transcription factors and precludes a simple common regulation of both genes.

This study brings experimental data to the predicted genomic structure of the ACCN1 and ACCN3 genes in the horse, as well as to their tissue expression, and opens new areas for further investigation of the role of ACCN polymorphisms in electrophysiological properties of cation channels and possibly also in horse pain.
Acknowledgements

This work was partly supported by a grant from the Haras Nationaux. Breeders and Haras Nationaux staff are greatly acknowledged for their help and constant interest in this work. We thank Dr Laurent Schibler for his advice during the promoter sequence analysis with MatInspector and W. Brand-Williams for reviewing the manuscript.

Conflicts of interest

The authors have declared no potential conflicts.

References


Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 ASIC2 multiple sequence alignment.

Appendix S2 ASIC3 multiple sequence alignment.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Estimated prevalence of the Type 1 Polysaccharide Storage Myopathy mutation in selected North American and European breeds

M. E. McCue*, S. M. Anderson†, S. J. Valberg*, R. J. Piercy‡, S. Z. Barakzai§, M. M. Binns¶, O. Distl**, M. C. T. Penedo††, M. L. Wagner† and J. R. Mickelson†

*Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA. †Department of Veterinary and Biomedical Sciences, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA. ‡Comparative Neuromuscular Diseases Laboratory, Department of Veterinary Clinical Sciences, The Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire AL9 7TA, UK. §Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian EH25 9RG, UK. ¶Equine Analysis Systems, Inc, 668 Lansing Lane, Midway, KY 40347, USA. **Institute of Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Bünneweg 17p, 30559 Hannover, Germany. ††Veterinary Genetics Laboratory, University of California, Davis, CA 95616, USA.

Summary

The GYS1 gene mutation that is causative of Type 1 Polysaccharide Storage Myopathy (PSSM) has been identified in more than 20 breeds of horses. However, the GYS1 mutation frequency or Type 1 PSSM prevalence within any given breed is unknown. The purpose of this study was to determine the frequency of the GYS1 mutation and prevalence of genetic susceptibility to Type 1 PSSM in selected breeds from Europe and North America. The GYS1 mutation was detected in 11 breeds, including, in order of increasing allele frequency, Shires, Morgans, Appaloosas, Quarter Horses, Paints, Exmoor Ponies, Saxon-Thuringian Coldbloods, South German Coldbloods, Belgians, Rhenish German Coldbloods and Percherons. The prevalence of genetic susceptibility to Type 1 PSSM in these breeds varied from 0.5% to 62.4%. The GYS1 mutation was not found in the sampled Thoroughbreds, Akhal-Tekes, Connemaras, Clydesdales, Norwegian Fjords, Welsh Ponies, Icelandics, Schleswig Coldbloods or Hanoverians, but failure to detect the mutation does not guarantee its absence. This knowledge will help breed associations determine whether they should screen for the GYS1 mutation and will alert veterinarians to a possible differential diagnosis for muscle pain, rhabdomyolysis or gait abnormalities.

Keywords glycogen storage disease, glycogen synthase, PSSM.

Introduction

Equine Polysaccharide Storage Myopathy (PSSM) is characterized by increased skeletal muscle glycogen concentration, abnormal polysaccharide accumulation in myofibers (Valberg et al. 1992; Firshman et al. 2006; McCue et al. 2006), and signs of painful muscle cramping, exertional rhabdomyolysis, and/or progressive muscle atrophy (McCue et al. 2006). Recently, a mutation in the GYS1 gene, and the resulting p.Arg309His substitution in the skeletal muscle isoform of glycogen synthase, was identified in 77% of Quarter Horses diagnosed with PSSM (McCue et al. 2008b). Heterozygosity for the mutation is sufficient to cause PSSM, however, the penetrance is clearly affected by environmental factors, including diet and exercise, and possibly by breed (McCue et al. 2008a,b). The disease resulting from this mutation has been termed Type 1 PSSM, and is associated with accumulation of amylase-resistant polysaccharide and the development of exertional rhabdomyolysis in horses fed diets high in non-structural carbohydrate.

Retrospective analysis demonstrated that the GYS1 mutation and Type 1 PSSM occur in more than 20 breeds from North America and Europe (McCue et al. 2008a; Stanley et al. 2009). Type 1 PSSM is a worldwide cause of...
neuromuscular disease in horses, although it is likely that other forms of PSSM exist (McCue et al. 2008a,b; Stanley et al. 2009). All horses with the GYS1 mutation share a minimal 350-kb haplotype, and we hypothesize that the mutation arose many hundreds of years ago in heavy horses and has spread from modern draft breeds to lighter breeds (McCue et al. 2008b). However, the frequency of the GYS1 mutation, and the prevalence of PSSM within individual breeds, is as yet unknown. The purpose of this study was to determine the frequency of the GYS1 mutation and estimate the prevalence of genetic susceptibility to Type 1 PSSM in a sampling of breeds from Europe and North America, with particular attention directed to those breeds where the GYS1 mutation has previously been identified.

Materials and methods

Sample collection

Whole blood or hair roots were obtained from horses of selected breeds in North America and Europe. Shires and Clydesdales were sampled in both Europe and North America.

North American horse samples were obtained from the Veterinary Genetics Laboratory (VGL) at the University of California at Davis (Table 1) by systematic random sampling, where every 10th hair root submission for the purpose of breed registration was obtained to ensure even geographical distribution.

DNA samples from European horses were obtained by three different sampling schemes. Simple random sampling was performed to obtain a subset of all foals born to registered Hanoverian breeding mares or stallions in 2001. These studs were distributed throughout all breeding districts in Germany.

Stratified random sampling was used to collect samples from Clydesdales, Saxon-Thuringian Coldblood, Schleswig Coldblood and South German Coldblood horses (Table 1), as well as the Rhenish German Coldblood (n = 12). Clydesdale samples were collected from 10 Scottish breeding yards. Fifty-eight of the horses were derived from six different stallion bloodlines; neither dam nor sire of the remaining 40 horses was significantly over-represented in Scottish

<table>
<thead>
<tr>
<th>Breed/origin</th>
<th>Number horses tested</th>
<th>PSSM (A) allele frequency</th>
<th>Confidence interval A allele</th>
<th>Wild-type (G) allele frequency</th>
<th>Confidence interval G allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percheron</td>
<td>149</td>
<td>0.346</td>
<td>0.293 0.401</td>
<td>0.654 0.599 0.707</td>
<td></td>
</tr>
<tr>
<td>Belgian</td>
<td>149</td>
<td>0.242</td>
<td>0.195 0.292</td>
<td>0.758 0.708 0.805</td>
<td></td>
</tr>
<tr>
<td>Paint</td>
<td>195</td>
<td>0.041</td>
<td>0.024 0.064</td>
<td>0.959 0.936 0.976</td>
<td></td>
</tr>
<tr>
<td>Quarter Horse</td>
<td>335</td>
<td>0.034</td>
<td>0.022 0.050</td>
<td>0.966 0.950 0.978</td>
<td></td>
</tr>
<tr>
<td>Appaloosa</td>
<td>152</td>
<td>0.030</td>
<td>0.014 0.053</td>
<td>0.970 0.947 0.986</td>
<td></td>
</tr>
<tr>
<td>Morgan</td>
<td>214</td>
<td>0.005</td>
<td>0.001 0.014</td>
<td>0.995 0.986 0.999</td>
<td></td>
</tr>
<tr>
<td>Shire</td>
<td>195</td>
<td>0.003</td>
<td>0.000 0.011</td>
<td>0.997 0.989 1.000</td>
<td></td>
</tr>
<tr>
<td>Thoroughbred</td>
<td>96</td>
<td>0.000</td>
<td>0.000 0.010</td>
<td>1.000 0.990 1.000</td>
<td></td>
</tr>
<tr>
<td>Akhal-Teke</td>
<td>50</td>
<td>0.000</td>
<td>0.000 0.019</td>
<td>1.000 0.981 1.000</td>
<td></td>
</tr>
<tr>
<td>Connemara</td>
<td>49</td>
<td>0.000</td>
<td>0.000 0.019</td>
<td>1.000 0.981 1.000</td>
<td></td>
</tr>
<tr>
<td>Clydesdale</td>
<td>48</td>
<td>0.000</td>
<td>0.000 0.020</td>
<td>1.000 0.980 1.000</td>
<td></td>
</tr>
<tr>
<td>Norwegian Fjord</td>
<td>46</td>
<td>0.000</td>
<td>0.000 0.021</td>
<td>1.000 0.979 1.000</td>
<td></td>
</tr>
<tr>
<td>Welsh Pony</td>
<td>45</td>
<td>0.000</td>
<td>0.000 0.021</td>
<td>1.000 0.979 1.000</td>
<td></td>
</tr>
<tr>
<td>Icelandic</td>
<td>36</td>
<td>0.000</td>
<td>0.000 0.026</td>
<td>1.000 0.974 1.000</td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South German CB</td>
<td>265</td>
<td>0.117</td>
<td>0.091 0.146</td>
<td>0.883 0.854 0.909</td>
<td></td>
</tr>
<tr>
<td>Saxon-Thuringian CB</td>
<td>44</td>
<td>0.068</td>
<td>0.028 0.133</td>
<td>0.932 0.867 0.972</td>
<td></td>
</tr>
<tr>
<td>Shire</td>
<td>33</td>
<td>0.000</td>
<td>0.000 0.029</td>
<td>1.000 0.971 1.000</td>
<td></td>
</tr>
<tr>
<td>Schleswig CB</td>
<td>33</td>
<td>0.000</td>
<td>0.000 0.029</td>
<td>1.000 0.971 1.000</td>
<td></td>
</tr>
<tr>
<td>Clydesdale</td>
<td>98</td>
<td>0.000</td>
<td>0.000 0.007</td>
<td>1.000 0.993 1.000</td>
<td></td>
</tr>
</tbody>
</table>

PSSM, Polysaccharide Storage Myopathy.

1Samples collected by systematic random sampling.
2Samples collected by stratified random sampling.
3Samples collected by non-random sampling.
4Samples collected by simple random sampling.
Clydesdale pedigrees, Saxon-Thuringian Coldblood, Schleswig Coldblood and South German Coldblood horses were sampled from April 2001 to June 2002. Studs were distributed over the entire breeding district to obtain a representative sample. Additional samples from the progeny of registered South German Coldblood horses were obtained in 2001, 2003 and 2004.

Lastly, non-random sampling was employed to obtain samples from Shires and Exmoor Ponies. All European Shire samples were taken at the Shire Horse Society Annual Show at Peterborough, UK (Table 1). Exmoor pony (n = 12) samples were taken from available horses in Germany.

**GYSI genotyping**

Genomic DNA was isolated with commercially available kits (Gentra; Qiagen Ltd, UK; Qiagen, USA) according to the manufacturers’ protocols. **GYSI** genotypes for the c.926G>A mutation were obtained by the established PCRRFLP technique (McCue et al. 2008b). Genotypes were recorded as homozygous normal (G/G), heterozygous (G/A) or homozygous affected (A/A).

**Statistical analysis**

G and A allele frequencies, G/G, G/A and A/A genotype frequencies, and predicted genotypic prevalence of Type 1 PSSM (based on both G/A and A/A genotypes causing susceptibility to PSSM) were calculated within each breed that had >30 samples. Ninety-five per cent confidence intervals for true allele frequency and disease prevalence were calculated using a likelihood-ratio confidence interval method (http://www.r-project.org).

**Results**

The **GYSI** mutation was detected and mutant allele frequency estimated in samples from nine different breeds across North American and Europe (Table 1). In addition, the mutation was detected in 6 of 12 Rhenish German Coldbloods (1 A/A, 5 G/A) and 1 of 12 Exmoor Ponies (G/A). The mutant allele frequency was the highest in the North American Percheron and Belgian Draft breeds followed by the South German Coldbloods. The mutation was also found at relatively high frequency in Paints, Appalosas and Quarter Horses, and at low frequency in Shires and Morgans. The **GYSI** mutation was not detected in the Thoroughbred, Akhal-Teke, Connemara, Clydesdale, Norwegian Fjord, Welsh Pony, Icelandic, Schleswig Coldblood and Hannoverian (Table 1).

**GYSI** genotypic frequencies, and their resulting estimated Type 1 PSSM prevalences with 95% confidence intervals, are listed in Table 2. The prevalence of genetic susceptibility to Type 1 PSSM (the sum of the A/A and G/A genotype frequencies) ranged from 0.005 and 0.009 in the Shires and Morgans respectively to 0.114, 0.389 and 0.624 in Saxon-Thuringian Coldblood, Belgian and Percheron respectively.

**Discussion**

The results of this study provide an indication of the frequency of the **GYSI** mutation in a variety of European and North American breeds. This study does not address how frequently horses that possess the **GYSI** mutation develop clinical signs of PSSM. The development of clinical myopathy in horses with the **GYSI** mutation has previously been shown to be affected by diet, management, other genes and perhaps other factors (Ribeiro et al. 2004, McCue et al. 2009), and thus the true penetrance of the G/A and A/A genotypes is not known. Under certain conditions, however, disease susceptibility appears to have a dominant mode of inheritance and complete penetrance (McCue et al. 2008). Therefore, the upper limit of Type 1 PSSM prevalence in a breed can be determined from the sum of the G/A and A/A genotype frequencies, assuming a dominant mode of inheritance and complete penetrance. By this estimation, the prevalence of genetic susceptibility to Type 1 PSSM varied from 0.00 to 0.624 among the breeds in this study.

Our results clearly show that the prevalence of the **GYSI** mutation and homozygosity for the mutation was high in the North American draft and some of the German Coldblood breeds, similar to previous results for Cob Normands (Herszberg et al. 2009). This may be because of the dramatic reduction in population sizes after modernization of the agricultural industry (Druml et al. 2008; Kavar & Dovc 2008), accompanied by founder effects and subsequent inbreeding. The low or zero prevalence of the **GYSI** mutation in Shires, Clydesdales and Schleswig Coldbloods may be reflective of breed origin and/or genetic divergence. Shires and Clydesdales are of British and Scottish origin and may be more genetically distant to the other draft breeds of continental European origin. Furthermore, the Schleswig Coldblood is the most genetically divergent of the German Coldblood breeds (Aberle et al. 2004). It is also possible that the European draft horse sampling schemes may have introduced bias because of relatedness of individuals, population stratification or selection of healthy individuals at a performance event.

The prevalence of the **GYSI** mutation in North American Quarter Horses and Paint horses of 0.066 and 0.077 respectively is similar to those found previously by Tryon et al. In that study, the prevalence of type 1 PSSM was estimated to be 0.113 in a stratified population of elite competitive Quarter Horses and 0.045 in a random population of Paint horses (Tryon et al. 2009).

The **GYSI** mutation was not identified in the North American Thoroughbreds, Akhal-Teke, Connemaras, Clydesdales, Norwegian Fjords, Icelandic horses, Welsh Ponies, Schleswig Coldbloods, or the European Clydesdales, Shires, or Hanoverians. However, sample sizes were low for
some of these breeds, and genotyping a minimum of 300 individuals from each breed would be necessary to have a 95% probability of detecting the GYS1 mutation if the true frequency is less than 0.01 (Gregorius 1980). Even with less than optimum sample sizes, the upper limit of the confidence interval for mutation frequency was less than 0.03 in all breeds where the mutation was not identified, suggesting that if the mutation segregates in these breeds it is at a low frequency. The mutation was not detected in Hanoverian samples in this study, but it has previously been detected in Hanoverian horses (McCue et al. 2008a; Stanley et al. 2009), suggesting that it does segregate at low frequency in this breed.

In conclusion, the highest GYS1 mutation frequencies were found in several draft breeds originating from continental Europe. It is also prevalent in breeds with Quarter Horse influence. We report for the first time the presence of the GYS1 mutation in Exmoor Ponies, and Rhenish German Coldbloods. The GYS1 mutation is rare to non-existent in some draft breeds, as well as some light horse breeds, including Thoroughbreds, among others. In these breeds, the most applicable first diagnostic procedure to determine the cause of a myopathy remains the traditional histopathological evaluation via muscle biopsy (McCue et al. 2008a). Despite some study design limitations, the GYS1 mutation frequency estimates provided in this paper provide clinically useful guidelines for veterinarians, breeders and breed associations when making genetic testing decisions.

**Acknowledgements**

This work was funded by Morris Animal Foundation Grants: D07EQ-041, ‘Frequency of a Polysaccharide Storage Myopathy Gene in Diverse Horse Breeds’ and D07EQ-402, ‘The genetic basis of Polysaccharide Storage Myopathy in diverse horse breeds’ (ME McCue, salary support), and American Quarter Horse Foundation Grant ‘Genetic Analysis of Glycogen Storage Disorders in Quarter Horses’. Sequence data from this article was originally published in McCue et al. Genomics 91 (2008) 458–466. Sequences were deposited with the EMBL/GenBank Data Libraries under Accession Nos. EU373800 (equine_GYS1_5’_exon1), EU373801 (equine_GYS1_cDNA), EU373802 (equine_GYS1_cDNA_PSSM_variant), EU373803 (equine_GYS1_exon16_3’).

---

**Table 2** Estimated Polysaccharide Storage Myopathy (PSSM) prevalence in North American and European horse breeds.

<table>
<thead>
<tr>
<th>Breed/origin</th>
<th>Number horses with A alleles</th>
<th>A/A frequency</th>
<th>G/A frequency</th>
<th>PSSM prevalence</th>
<th>Confidence interval PSSM prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Upper</td>
</tr>
<tr>
<td>North America¹</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percheron</td>
<td>93/149</td>
<td>0.067</td>
<td>0.557</td>
<td>0.624</td>
<td>0.545</td>
</tr>
<tr>
<td>Belgian</td>
<td>58/149</td>
<td>0.094</td>
<td>0.295</td>
<td>0.389</td>
<td>0.313</td>
</tr>
<tr>
<td>Paint</td>
<td>15/195</td>
<td>0.005</td>
<td>0.072</td>
<td>0.077</td>
<td>0.045</td>
</tr>
<tr>
<td>Quarter Horse</td>
<td>22/335</td>
<td>0.003</td>
<td>0.063</td>
<td>0.066</td>
<td>0.042</td>
</tr>
<tr>
<td>Appaloosa</td>
<td>9/152</td>
<td>0</td>
<td>0.059</td>
<td>0.059</td>
<td>0.042</td>
</tr>
<tr>
<td>Morgan</td>
<td>2/214</td>
<td>0</td>
<td>0.009</td>
<td>0.009</td>
<td>0.002</td>
</tr>
<tr>
<td>Shire</td>
<td>1/195</td>
<td>0</td>
<td>0.005</td>
<td>0.005</td>
<td>0.000</td>
</tr>
<tr>
<td>Thoroughbred</td>
<td>0/96</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Akhal-Teke</td>
<td>0/50</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Connemara</td>
<td>0/49</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Clydesdale</td>
<td>0/48</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Norwegian Fjord</td>
<td>0/46</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Welsh Pony</td>
<td>0/45</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Icelandic</td>
<td>0/36</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

| Europe       |                              |               |               |                 |                                 |
| South German CB² | 54/265                 | 0.030         | 0.174         | 0.204           | 0.158                           |
| Saxon-Thuringian CB² | 5/44                | 0.023         | 0.099         | 0.114           | 0.042                           |
| Shire³       | 0/33                         | 0             | 0             | 0.000           | 0.000                           |
| Schleswig CB² | 0/33                         | 0             | 0             | 0.000           | 0.000                           |
| Clydesdale²  | 0/98                         | 0             | 0             | 0.000           | 0.000                           |
| Hanoverian⁴  | 0/214                        | 0             | 0             | 0.000           | 0.000                           |

¹Samples collected by systematic random sampling.
²Samples collected by stratified random sampling.
³Samples collected by non-random sampling.
⁴Samples collected by simple random sampling.
Conflicts of interest

Drs Valberg, Mickelson and McCue own the license for PSSM testing and receive sales income from its use. Their financial and business interests have been reviewed and managed by the University of Minnesota in accordance with its conflict of interest policies.

References


Identification of equine major histocompatibility complex haplotypes using polymorphic microsatellites

C. T. Tseng*, D. Miller*, J. Cassano*, E. Bailey† and D. F. Antczak*

*Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA. †Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, KY 40546 USA

Summary

A system for identifying equine major histocompatibility complex (MHC) haplotypes was developed based on five polymorphic microsatellites located within the MHC region on ECA 20. Molecular signatures for 50 microsatellite haplotypes were recognized from typing 353 horses. Of these, 23 microsatellite haplotypes were associated with 12 established equine leucocyte antigen (ELA) haplotypes in Thoroughbreds and Standardbreds. Five ELA serotypes were associated with multiple microsatellite subhaplotypes, expanding the estimates of diversity in the equine MHC. The strong correlations between serological and microsatellite typing demonstrated a linkage to known MHC class I protein polymorphisms and validated this assay as a useful supplement to ELA serotyping, and in some applications, a feasible alternative method for MHC genotyping in horse families and in population studies.

Keywords equine, haplotype, major histocompatibility complex (MHC), microsatellite, polymorphism.

Diversification of major histocompatibility complex (MHC) class I and class II genes in the vertebrate genome is a key feature of their role in antigen presentation in the adaptive immune system (Hughes & Nei 1992). However, assigning MHC haplotypes to individuals remains challenging even in the age of whole-genome sequencing. International workshops convened in the 1980s identified 19 serological specificities as products of the equine leucocyte antigen (ELA) system (Lazary et al. 1988). The serological assay is limited by the amount and variety of antibody reagents available, and the complex alloantisera are directed primarily against MHC class I antigens. To address these limitations, alleles at five polymorphic microsatellite loci within the equine MHC (Fig. S1, supporting online information) were used to identify distinct haplotypes based upon the well-known linkage disequilibrium within the MHC (see Appendix S1 for Materials and Methods). In addition, we evaluated these molecular haplotypes as proxies for serologically defined markers by testing their level of correspondence to the serological haplotypes of the ELA system (Lazary et al. 1988) within and between horse breeds (Table 1).

Address for correspondence
D. F. Antczak, Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA.
E-mail: dfa1@cornell.edu
Accepted for publication 17 August 2010

Microsatellite haplotypes were defined in three ways: (i) horses homozygous for microsatellite allele constellations that allowed unambiguous identification of MHC microsatellite haplotypes with supporting serological data; (ii) horses belonging to families that allowed definition of microsatellite haplotypes by familial transmission; (iii) association of haplotypes with known serotypes that allowed definition of haplotypes in cis configuration. Haplotypes were defined when they could be identified in at least two individuals who were not related within two generations, or when observed segregating in families. When possible, microsatellite haplotype nomenclature was derived from nomenclature of the ELA complex. When a serotype was associated with multiple microsatellite haplotypes, indicating a genetic complexity not reflected by serotyping, the ELA type was followed by a lower-case letter to denote the subtype. Finally, haplotypes not associated with serotypes were given the prefix ‘COR’ (for Cornell) followed by a number as a working title.

A total of 50 microsatellite haplotypes were identified from typing 353 horses using the five intra-MHC microsatellite loci (Table 1; Fig. S1). Variable numbers of alleles were detected for each microsatellite locus (range 8–13, Table S1), but most were not uniquely associated with any single haplotype (Table 1). MHC haplotype definitions were most accurate when based on alleles at all five loci (Fig. S2). Of the 50 haplotypes, 23 were linked to 12 known ELA serotypes based on previous serotyping results (Tables 1 & S2).
Table 1  Assigned equine major histocompatibility complex (MHC) microsatellite constellations with corresponding serological equine leucocyte antigen (ELA) haplotypes.

<table>
<thead>
<tr>
<th>Microsatellite haplotype</th>
<th>Serological ELA haplotype</th>
<th>$r^2$</th>
<th>Class I</th>
<th>Class II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>UMN-JH34-2</td>
<td>COR112</td>
</tr>
<tr>
<td>A1</td>
<td>A1</td>
<td>0.93</td>
<td>211</td>
<td>236</td>
</tr>
<tr>
<td>A2</td>
<td>A2</td>
<td>0.86</td>
<td>211</td>
<td>262</td>
</tr>
<tr>
<td>A3a</td>
<td>A3</td>
<td>0.93</td>
<td>207</td>
<td>254</td>
</tr>
<tr>
<td>A3b</td>
<td>A3</td>
<td></td>
<td>207</td>
<td>262</td>
</tr>
<tr>
<td>A3c</td>
<td>A3</td>
<td></td>
<td>207</td>
<td>262</td>
</tr>
<tr>
<td>A3d</td>
<td>A3</td>
<td></td>
<td>211</td>
<td>252</td>
</tr>
<tr>
<td>A4a</td>
<td>A4</td>
<td>0.92</td>
<td>205</td>
<td>236</td>
</tr>
<tr>
<td>A4b</td>
<td>A4</td>
<td></td>
<td>205</td>
<td>252</td>
</tr>
<tr>
<td>A4c</td>
<td>A4</td>
<td></td>
<td>217</td>
<td>262</td>
</tr>
<tr>
<td>A5a</td>
<td>A5</td>
<td>0.88</td>
<td>221</td>
<td>254</td>
</tr>
<tr>
<td>A5b</td>
<td>A5</td>
<td></td>
<td>221</td>
<td>254</td>
</tr>
<tr>
<td>A6</td>
<td>A6</td>
<td>0.76</td>
<td>197</td>
<td>243</td>
</tr>
<tr>
<td>A7</td>
<td>A7</td>
<td>0.83</td>
<td>205</td>
<td>256</td>
</tr>
<tr>
<td>A8</td>
<td>A8</td>
<td>0.95</td>
<td>219</td>
<td>268</td>
</tr>
<tr>
<td>A9a</td>
<td>A9</td>
<td>0.93</td>
<td>217</td>
<td>264</td>
</tr>
<tr>
<td>A9b</td>
<td>A9</td>
<td></td>
<td>217</td>
<td>262</td>
</tr>
<tr>
<td>A10a</td>
<td>A10</td>
<td>0.86</td>
<td>221</td>
<td>236</td>
</tr>
<tr>
<td>A10b</td>
<td>A10</td>
<td></td>
<td>221</td>
<td>236</td>
</tr>
<tr>
<td>A10c</td>
<td>A10</td>
<td></td>
<td>221</td>
<td>236</td>
</tr>
<tr>
<td>A10d</td>
<td>A10</td>
<td></td>
<td>221</td>
<td>252</td>
</tr>
<tr>
<td>A15</td>
<td>A15</td>
<td>0.75</td>
<td>219</td>
<td>254</td>
</tr>
<tr>
<td>A19</td>
<td>A19</td>
<td>0.75</td>
<td>211</td>
<td>262</td>
</tr>
<tr>
<td>COR1</td>
<td></td>
<td></td>
<td>221</td>
<td>262</td>
</tr>
<tr>
<td>COR2</td>
<td></td>
<td></td>
<td>217</td>
<td>262</td>
</tr>
<tr>
<td>COR3</td>
<td></td>
<td></td>
<td>211</td>
<td>254</td>
</tr>
<tr>
<td>COR4</td>
<td></td>
<td></td>
<td>221</td>
<td>264</td>
</tr>
<tr>
<td>COR5</td>
<td></td>
<td></td>
<td>197</td>
<td>236</td>
</tr>
<tr>
<td>COR6</td>
<td></td>
<td></td>
<td>221</td>
<td>254</td>
</tr>
<tr>
<td>COR7</td>
<td></td>
<td></td>
<td>219</td>
<td>262</td>
</tr>
<tr>
<td>COR8</td>
<td></td>
<td></td>
<td>211</td>
<td>254</td>
</tr>
<tr>
<td>COR9</td>
<td></td>
<td></td>
<td>211</td>
<td>260</td>
</tr>
<tr>
<td>COR10</td>
<td></td>
<td></td>
<td>207</td>
<td>262</td>
</tr>
<tr>
<td>COR11</td>
<td></td>
<td></td>
<td>219</td>
<td>264</td>
</tr>
<tr>
<td>COR12</td>
<td></td>
<td></td>
<td>215</td>
<td>254</td>
</tr>
<tr>
<td>COR13</td>
<td></td>
<td></td>
<td>215</td>
<td>252</td>
</tr>
<tr>
<td>COR14</td>
<td></td>
<td></td>
<td>209</td>
<td>260</td>
</tr>
<tr>
<td>COR15</td>
<td></td>
<td></td>
<td>211</td>
<td>252</td>
</tr>
<tr>
<td>COR16</td>
<td></td>
<td></td>
<td>209</td>
<td>258</td>
</tr>
<tr>
<td>COR17</td>
<td></td>
<td></td>
<td>207</td>
<td>248</td>
</tr>
<tr>
<td>COR18</td>
<td></td>
<td></td>
<td>209</td>
<td>252</td>
</tr>
<tr>
<td>COR19</td>
<td></td>
<td></td>
<td>209</td>
<td>268</td>
</tr>
<tr>
<td>COR20</td>
<td></td>
<td></td>
<td>207</td>
<td>244</td>
</tr>
<tr>
<td>COR21</td>
<td></td>
<td></td>
<td>217</td>
<td>252</td>
</tr>
<tr>
<td>COR22</td>
<td></td>
<td></td>
<td>221</td>
<td>256</td>
</tr>
<tr>
<td>COR23</td>
<td></td>
<td></td>
<td>197</td>
<td>254</td>
</tr>
<tr>
<td>COR24</td>
<td></td>
<td></td>
<td>211</td>
<td>243</td>
</tr>
<tr>
<td>COR25</td>
<td></td>
<td></td>
<td>221</td>
<td>262</td>
</tr>
<tr>
<td>COR26</td>
<td></td>
<td></td>
<td>203</td>
<td>252</td>
</tr>
<tr>
<td>COR27</td>
<td></td>
<td></td>
<td>209</td>
<td>262</td>
</tr>
</tbody>
</table>

1 Data collected from a cohort of 353 horses (2n = 706 chromosomes) with 600 named haplotypes and 106 blank haplotypes. Microsatellite subtypes of common ELA serotypes are indicated by lowercase letters. Local Cornell (COR) haplotypes did not have an associated ELA serotype. For subhaplotypes, r values were calculated individually and the combined statistics are reported here. For detailed data and calculations, see Table S2 and Fig. S3. 
2 Breeds: APP, Appaloosa; AR, Arabian; Gyp, Gypsy Vanner; Han, Hanoverian; MFT, Missouri Fox Trotter; Old, Oldenburg; P, pony of unknown breeding, presumably mixed breed; QH, Quarter horse; SB, Standardbred; TB, Thoroughbred; WB, warmblood; U, breed of horse unknown, presumed mixed breed.
3 Number of times each named microsatellite haplotype was identified and counted.

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 150–153
The remaining 27 haplotypes, designated as ‘COR’ haplotypes, were discovered in horses that carried unidentified haplotypes (negative for known ELA serotypes) or that had not been serotyped.

The equine MHC haplotypes described here represent a large increase in the number of recognized haplotypes segregating within families and detected among unrelated individuals. In addition to defining new levels of diversity in the equine MHC, this study also confirms the distribution of associated broad ELA serotypes in Thoroughbreds and Standardbreds, where a small number of ELA types are carried by the majority of horses in these breeds (Antczak et al. 1986). Overall, the named microsatellite haplotypes defined 90% and 97% of the haplotypes in Thoroughbreds and Standardbreds, respectively, and 63% in Arabian horses (Table S3).

Five ELA serotypes were associated with multiple microsatellite haplotypes. The origin and extent of the diversity within subtypes remain unclear. A strong association was observed between ELA serotypes and alleles at microsatellite locus UMN-JH34-2, located in a cluster of expressed MHC class I genes (Tallmadge et al. 2005) (Fig. S2). Microsatellite variation within serologically defined ELA haplotypes was restricted largely to the MHC class II region. For some subtypes, there was little or no sharing of MHC class II microsatellite alleles (Table 1). In the ELA-A3 haplotype carried by the donor horses of the equine bacterial artificial chromosome library (Gustafson et al. 2003) and the genome sequence (Wade et al. 2009), there is little evidence of sequence variation in MHC class I or II genes among the subhaplotypes (Tallmadge et al. 2005, 2010; Miller and Antczak, personal communication). The subhaplotypes could reflect microsatellite evolution in ancient haplotypes that is independent of alterations in MHC class II structural genes or the recombination between the MHC class I and class II regions that is manifest in new haplotypes. Haplotypes that were serology positive and microsatellite negative could be as yet unrecognized subhaplotypes that were not captured by a single microsatellite haplotype. In contrast, MHC haplotypes that were serology negative and microsatellite positive may have been defined as a result of inaccurate serotyping (Table S2).

The reported high linkage disequilibrium within the equine genome between breeds presumably reflects the recent breed divergence and sharing of founders (Wade et al. 2009). While some ELA haplotypes are shared among breeds, Antczak et al. (1986) previously described differences between breeds in the frequency of detection of various ELA haplotypes, including the apparent absence of some haplotypes in certain breeds. Similar restriction was seen in the microsatellite haplotypes defined here (Tables 1 & S3). The high correlation of serotypes with microsatellite haplotypes within certain breeds indicates that intra-MHC microsatellite typing with the described five-member panel can be used for MHC genotyping within breeds for population and family studies, particularly in Thoroughbreds and Standardbreds.

The microsatellite typing reported here allows rapid and accurate identification of equine MHC haplotypes in most instances, requires no specialized alloantibody reagents, and as such offers advantages over ELA serotyping. This study used ELA serotyping, known MHC homozygous horses, and MHC gene sequencing (Tallmadge et al. 2010) to link key microsatellite constellations to common MHC haplotypes and to expand the estimates of diversity in the equine MHC. Microsatellite typing is thus a powerful complement to classical ELA serotyping and sequencing for identifying MHC haplotypes of the horse.

Acknowledgements
We thank Professor Sidney W. Ricketts of the Beaufort Cottage Stables, Newmarket, Suffolk, England, and Dr. Samantha Brooks of the Animal Science Department at Cornell for generously supplying samples for testing. This work was supported by grants from the Harry M. Zweig Memorial Fund for Equine Research, the Morris Animal Foundation, and the US National Institutes of Health (HD-049545). DFA is an investigator of the Dorothy Russell Havemeyer Foundation, Inc.

Conflicts of interest
The authors declare no conflicts of interest.

References
Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Additional material.

Figure S1 Map positions of selected intra-major histocompatibility complex microsatellites in the horse.

Figure S2 The 2 × 2 contingency tables for microsatellite allele association using alleles of the A2 and A3a microsatellite haplotypes as examples.

Figure S3 The 2 × 2 contingency tables for correlation coefficient (r) analysis in 218 serotyped and microsatellite (msat)-typed horses.

Table S1 Microsatellite allele size and number of animals positive for each allele in a sample population of 320 horses and 33 unknown breed ponies.

Table S2 Correlation between major histocompatibility complex microsatellite (msat) haplotypes and serological (sero) equine leucocyte antigen haplotypes in 218 serotyped and msat-typed horses.

Table S3 Equine major histocompatibility complex microsatellite haplotype frequencies across breeds.

Table S4 Equine major histocompatibility complex microsatellite primer sequences.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Identification of the myostatin locus (MSTN) as having a major
effect on optimum racing distance in the Thoroughbred horse in
the USA

M. M. Binns, D. A. Boehler and D. H. Lambert
Equine Analysis, PO Box 4570, Midway, KY 40347, USA

Summary

One hundred and eighty-nine Thoroughbred horses that had won Graded Stakes races in North America were genotyped with the Illumina Equine SNP50 bead chip. Association tests using PLINK to determine whether any SNPs were associated with optimum racing distance (7 furlongs and under compared to 8–10 furlongs) identified a locus on ECA18 that was statistically significant \((-\log 10 \text{EMP2} = 1.63)\) at the genome-wide level following permutation analysis (10 000 permutations). Bioinformatic analysis revealed that the two ECA18 SNPs with the highest statistical significance spanned the MSTN (myostatin) locus. Mutations in myostatin in several mammalian species have been associated with increased muscling, with a preferential increase in fast glycolytic type IIB fibres, which would increase power potential. Thoroughbred horses that race over sprint distances, which are 5–7 furlongs, are often characterized by impressive hind quarter musculature, strongly suggesting that the association observed between the ECA18 SNPs and optimum race distance is mediated through MSTN.

Keywords association test, horse, myostatin, racing distance, SNP array, Thoroughbred.

Introduction

The Thoroughbred horse breed was established in England in the early 1700s based on crosses between stallions of Arabian origin and a poorly defined, possibly indigenous, group of mares. The founder population was small; all contemporary males trace back to one of three stallions, the Godolphin Arabian, the Byerley Turk and the Darley Arabian, whilst on the female side about seventy foundation mares have been identified (Willett 1970). A stud book for Thoroughbred horses was initiated in 1791, and pedigree records for the breed, which now numbers about 500 000 horses, are maintained by Thoroughbred registries worldwide.

Initially, the new breed raced over extended distances; match races in which two horses competed against each other over 4–6 miles were common, and horses frequently had to run multiple heats during a day to determine the winner. Today’s English Classic races, the St. Leger (run over 1¾ miles), the Derby and Oaks (run over one and a half miles) and the 1000 and 2000 Guineas (run over 1 mile) were established in 1776, 1780, 1779, 1809 and 1814, respectively. The decreasing distances over time suggest that there was already a trend towards racing over shorter distances by 1800, and the fashion has continued to the point that, in North America, races at one and a half miles and over are now frequently referred to as ‘marathons’ and are relatively uncommon. In North America, whilst the classic Triple Crown races, The Kentucky Derby, The Preakness and the Belmont are run over one and a quarter 1½ miles and one and a half miles, respectively, there are important races at 7 furlongs and under, and there has been an emphasis on selecting for speed through the fast pace at which most US races are run when compared to European races. In general, contemporary Thoroughbred horses are classified by optimum distance into sprint, middle distance and stayers, and distinct pedigrees specific for each distance have been developed by breeders.

The biochemistry of this distance specialization has remained obscure. An examination of the muscle fibre types present in sprinters and stayers and the response of muscle to different training regimes has been hampered by the difficulty of reliably sampling the same muscle location in different individuals (Snow & Guy 1980; Sewell et al. 1994).

Mysostatin (MSTN) is a member of the transforming growth factor \(\beta\) family that is expressed in skeletal muscle
and acts as a negative regulator of muscle mass. Several natural mutations in MSTN have been identified in cattle (Grobet et al. 1997, 1998; Kambadur et al. 1997, McPherron & Lee 1997; Marchitelli et al. 2003) sheep (Clop et al. 2006), mice (Nishi et al. 2002) and recently human (Schuelke et al. 2004), and these result in a double-muscling phenotype. In addition, a myostatin mutation in whippet dogs was shown to be associated with increased muscling and athletic ability, although no equivalent mutation was found in the closely related greyhound breed (Lee 2007; Mosher et al. 2007).

This study seeks to investigate whether there are any specific loci in the equine genome associated with distance specialization in the Thoroughbred horse using a whole genome association approach.

Methods

DNA and genotyping

DNA was prepared from blood and hair samples using standard protocols from 189 Thoroughbred horses that had won Graded Stakes races in North America. The samples represent a random selection of horses that won Graded Stakes races, and no full siblings were included in the sample group. DNA was quantitated with a Quant-iT PicoGreen dsDNA assay kit (Invitrogen) using a Nanodrop ND3300 fluorospectrometer following the Nanodrop protocol ‘PicoGreen assay for dsDNA’.

SNP genotyping and data analysis

SNP genotyping was carried out with the Equine SNP50 bead chip (Illumina) using standard protocols at the Genotyping Shared Resource facility, Mayo Clinic, Rochester, MN.

Quality control of the genotyping data was assessed in several ways. Duplicate samples (n = 5) revealed a concordance in genotype calls >99.9%. Using PLINK (Purcell et al. 2007), Mendelian inheritance errors in a trio set of sire/dam/offspring were <0.01%. Data were subjected to the SEXCHECK option to ensure that sample fidelity had been maintained, and the MC1R mutation that produces chestnut coat colour in horses was mapped as a control to its expected location on ECA3 using the SNP genotype data and the coat colour phenotypes of the samples tested.

Horses were stratified into three distance classes: 7 furlongs and under, 8–10 furlongs and over 10 furlongs, based on their individual optimum race performances obtained from Equibase. That is, the distance of the horse’s best performance in Graded Stakes races was recorded, and if an individual won the same class of Graded Stakes race at more than one distance, the longer distance was recorded. Association test statistics and permutations to assess statistical significance were determined using the ASSOC and MPERM functions of PLINK.

Results

One hundred and eighty-nine horses that had won Graded Stakes races were classified into those that had won races at seven or less furlongs (n = 125), those that had won at eight to ten furlongs (n = 44) and those that had won races at more than 10 furlongs (n = 20).

Whole genome SNP genotypes were determined using the Illumina Equine SNP50 bead chip system, which comprises 54 602 SNPs uniformly distributed across the equine genome. The ASSOC option was run to compare the genotypes of the 7 furlongs and under group to the 8–10 furlong group. The results are presented in Fig. 1a, where it can be seen that the SNPs with the most significant P values (smallest P value = 2.77E-07) are located on ECA18. To confirm the significance of this association, a permutation analysis with 10 000 permutations was carried out using the MPERM option in PLINK. The results of the permutation analysis, reported as max (T) empirical P values (EMP2) are presented in Fig. 1b. Genome-wide significance for the horse is estimated to be −log 10 EMP2 = 1.3. Two markers on ECA18, BIEC2-417274 (located at 65 868 604 bp) and BIEC2-417495 (located at 67 186 093 bp) reach genome-wide significance with −log 10 EMP2 values of 1.3829 and 1.6345, respectively. The location of these markers on ECA18, together with other SNPs in the immediate region that show association but do not reach genome-wide significance, is shown in Fig. 1c.

It should be noted from Fig. 1b that a second locus on ECA17, whilst not reaching genome-wide significance (−log 10 EMP2 = 0.9179), is high enough to suggest that a second locus may make a contribution to optimum racing distance.

A second analysis compared the 7 furlongs and under group of horses (n = 125) to the group that won at over 10 furlongs (n = 20). Whilst the same region on ECA18 was again identified, there was insufficient power in this analysis to reach genome-wide significance with the highest −log 10 EMP2 = 0.261, because of the limited number of samples in the longer-distance category.

The genotype and allele frequencies for BIEC2-417495 (the SNP with the highest statistical significance) were estimated from a set of 464 Thoroughbred horses that had not been selected on the basis of distance. The A allele that is associated with a longer optimum distance was present at a frequency of 0.473, whilst the G allele associated with sprinting was present at a frequency of 0.527. The genotype frequencies for the AA:AG:GG genotypes in this population were 0.189:0.567:0.244, respectively.

Discussion

The association analysis comparing Thoroughbred horses that won Graded Stakes races at a distance of 7 furlongs or less with those that won Graded Stakes races over a distance of 8–10 furlongs identified a single locus on ECA18 that...
achieved genome-wide significance following permutation analysis.

Whilst a clear association peak involving a large number of SNPs towards the distal end of ECA18 is evident in Fig. 1c, only two SNPs, BIEC2-417274 (located at 65,868,604 bp) and BIEC2-417495 (located at 67,186,093 bp), reach genome-wide significance with $-\log 10$ EMP2 values of 1.3829 and 1.6345, respectively. When the genes in this

Figure 1 The myostatin (MSTN) locus shows a significant association with optimum winning distance. (a) Results of genome-wide association test for optimum distance, 7 furlongs vs. 8–10 furlongs. The x axis shows each individual equine chromosome, where 1–31 are ECA1–ECA31 and 32 is the X chromosome. The y axis shows the $-\log 10$ P value. (b) Results of permutation analysis (10,000 permutations) showing a genome-wide significant signal on ECA18. The x axis shows the individual chromosomes as above, and the y axis shows the $-\log 10$ EMP2 value. (c) Results of the association test for each ECA18 SNP plotted along the length of the chromosome to locate the locus associated with optimum distance. The x axis shows the base pairs along ECA18, and the y axis shows the $-\log 10$ P value. A strong signal is observed at approximately 67,000,000 base pairs.
region of ECA18 were examined using the Ensembl genome browser, it was noted that the MSTN (myostatin) gene is located between the two significant SNPs at 66 490 208–66 495 180 bp.

Traditionally, the Thoroughbred horse industry has regarded horses that race over less than seven furlongs, between seven and twelve furlongs, and longer than twelve furlongs as different, referring to them as sprinters, middle-distance and stayers, respectively. The pedigrees of horses racing over these different distances tend to be different and breeders have frequently tried to produce middle-distance horses by breeding sprinters to stayers. The horses in the different distance groups also often possess distinct body morphologies, with the sprinters being heavily muscled, whilst the middle-distance and staying horses have a lighter musculature. This phenotypic difference suggests that myostatin might play a role in the distance preference of Thoroughbred horses.

The increased muscling seen with myostatin mutations does not affect all muscle fibre types evenly, and there is a preferential increase in fast glycolytic type IIB fibres (Deveaux et al. 2001; Hennebry et al. 2009), which is consistent with improved sprinting ability.

The identification of SNP markers linked to the myostatin loci that are associated with athletic ability in the Thoroughbred horse represents the first gene influencing athletic ability identified in horses.

Whilst this paper was in preparation, Hill et al. (2010) published research showing that a polymorphism in the myostatin gene is strongly associated with best race distance in elite Thoroughbred horses. In their study, myostatin was selected as a candidate gene, and whilst the intronic mutation used is in the myostatin gene, they state that the results do not preclude the functional variant being in a neighbouring gene. They state that there are no other plausible candidates within 2 Mb upstream or downstream of the myostatin gene, although it is unclear whether the functional mutation could be further away than this. In addition, they do not appear to have controlled for the possibility of population stratification between the sprinting and staying populations by testing association with other genes unlikely to be involved in muscle physiology.

In this study, following a whole genome-wide association analysis strategy, we identify a single locus on ECA18 that is significant at the genome-wide level following permutation analysis. In our analysis, population stratification between sprinters and middle-distance horses is internally controlled for, and excluded, by the remaining 54 600 SNPs uniformly distributed across the genome. No other region of the genome showed a genome-wide statistically significant difference, indicating that no population stratification exists.

The results clearly have implications for the Thoroughbred racing and breeding industries, although it is important to stress that the markers identified do not in any way define the likely racing ‘class’ of the individual, merely over what distance the horse is likely to be most effective. It would be wrong to directly equate the preference for sprint distances with speed, and myostatin does not represent ‘the speed gene’. There will be many horses that, whilst homozygous for the sprint genotype, are poor racehorses and would be beaten over sprint distances by better racehorses with one of the other two genotypes.

**Acknowledgements**

We thank Michael and Diane Blowen, Old Friends, Kentucky (http://www.oldfriendsequine.org/) for help with samples.

**Conflicts of interest**

The authors have not declared any conflicts of interest.

**References**


Morphological variation in the horse: defining complex traits of body size and shape

S. A. Brooks*, S. Makvandi-Nejad†, E. Chu†, J. J. Allen†, C. Streeter*, E. Gu†, B. McCleery†, B. A. Murphy‡, R. Bellone§ and N. B. Sutter†

*Department of Animal Science, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY 14853, USA. †Department of Clinical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA. ‡School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland. §Department of Biology, University of Tampa, Tampa, FL 33606, USA

Summary

Horses, like many domesticated species, have been selected for broad variation in skeletal size. This variation is not only an interesting model of rapid evolutionary change during domestication, but is also directly applicable to the horse industry. Breeders select for complex traits like body size and skeletal conformation to improve marketability, function, soundness and performance in the show ring. Using a well-defined set of 35 measurements, we have identified and quantified skeletal variation in the horse species. We collected measurements from 1215 horses representing 65 breeds of diverse conformation such as the American Miniature, Shetland Pony, Arabian Horse, Thoroughbred, Shire and Clydesdale. Principal components analysis has identified two key dimensions of skeletal variation in the horse. Principal component 1 is positively correlated with every measurement and quantifies overall body size. Principal component 2 captures a pattern of bone widths vs. lengths and thus quantifies variation in overall bone thickness. By defining these complex skeletal traits, we have created a framework for whole genome association studies to identify quantitative trait loci that contribute to this variation.

Keywords complex trait, horse, morphology, principal components analysis, skeleton.

Introduction

Selective modification of domesticated plants and animals began millennia ago and is still ongoing today. In the present era of continually decreasing costs for whole genome genotyping and sequencing, it is now feasible to work in any number of species. Domesticated mammals are particularly powerful systems for gene discovery, as well as valuable models for identifying traits under selection. To fully leverage the intrinsic genetic value of these highly structured populations, however, it is critical to obtain high-quality phenotypes for study. The horse, like other domesticated species, has been moulded through selection within breeds into diverse skeletal morphologic types, often to fit specific functions. As a result of this selection, there are now breeds of horse that greatly differ from one another in body size and shape.

Quantitative measurement of the horse’s morphology has been previously carried out by a number of groups (Sadek et al. 2006; Weller et al. 2006; Komosa & Purzyc 2009). In each case, only one or two breeds were studied, and therefore the identified patterns of variation were primarily within breeds and not between them. High estimates of heritability for skeletal measures have been calculated in several breeds of horse. In the Murgese, for example, heritability of cannon bone circumference was estimated at 0.44 (Dario et al. 2006). However, most of the previous studies documenting the heritability of skeletal traits have focused on qualitatively described traits like ‘over at the knee,’ ‘turned-in feet,’ ‘weak pasterns’ and ‘tied below the knee,’ instead of quantitative traits as described in this study (Love et al. 2006, Dario et al. 2006). Here, we aim to define skeletal variation for the horse species as a whole by collecting quantitative body measurements from a broad set of horse breeds with diverse body types. We have used a principal components analysis (PCA) to identify patterns of
variation across 35 measurements of the head, neck, trunk and limbs. Unlike individual measurements, PCA can identify whole-body patterns like bone thickness. Furthermore, the distillation of complex skeletal morphometric data sets through PCA has proven successful in other domestic mammals such as the dog (Sutter et al. 2007). These data will enable mapping of genes contributing to these complex traits.

Materials and methods

Collection of measurements

We developed a protocol for body measurement of live horses using a provided tape measure and a uniform set of illustrated instructions (Fig. S1). A laboratory staff member, collaborator, or individual horse owner used the protocol to guide measurement of each horse and to subjectively score a set of traits coded as factors for body condition score (Henneke et al. 1983), muzzle profile (‘dish-faced’ to neutral to Roman nosed), degree of feathering (none to copious), bite evenness (underbite to neutral to overbite) and bone thickness (Fig. S2a–f). We noted if the horse was shod on the fore or hind hooves. Vital statistics and basic history (date or year of birth, registered name, registry, barn name, brief notes on any injury or disease, a photo and a pedigree for each horse) was also collected from the horse owner or agent. The coefficient of variation for each measure is shown in Fig. S3. All of the horses measured were at least four years of age at the time of trait collection and therefore relatively skeletally mature (range 4–34.5 years, median age 11.5; Fig. S2a) (Sadek et al. 2006).

For each horse, we collected the following quantitative measurements: (1) Eye to eye width: standing in front of the horse, measure across the forehead between the inside corners of the eyes. Keep the tape taut and straight. (2) Jaw width: across the underside of the head, measure straight across from the outsides of the cheekbones. Keep the tape taut and straight. (3) Head length: starting between the top corners of the two nostrils, measure straight to the front of the poll. (4) Muzzle circumference: take the circumference of the muzzle underneath the halter, settling the tape directly in front of the cheekbones. (5) Left eye to mouth length: measure from the corner of the mouth to the back corner of the eye. (6) Left eye to jaw length: measure from the back corner of the eye to the deepest point of the cheek curve. (7) Left ear length: on the side of the ear closest to the poll, measure from the base to the tip of the ear. (8) Neck length, head level with withers: measure from the poll to the withers, with your horse’s head level with the withers. (9) Neck length, head down to the ground: measure as in #8, but with your horse’s head stretched as close to the ground as possible. (10) Neck circumference at throat latch: settle the tape where the throat latch of the bridle goes. Pull it snugly but not uncomfortably for your horse. (11) Neck circumference at base: settle the tape just in front of the withers. Let it rest on the chest and curve comfortably around the base of the neck. (12) Height at withers: measure from the ground to the highest point of the withers. Keep the tape taut and straight. (13) Height at croup: measure from the ground straight up to the highest point of the rump. (14) Height at dock: measure from the ground straight up to the base of the tail. (15) Tail length: start at the end of the bony portion of the tail and measure to the dock. (16) Withers to croup, straight tape: measure across the back from the withers to the point of the croup. The tape will not touch the back except at the ends. (17) Withers to croup, contoured: measure the back from the withers to the croup, allowing the tape to relax and touch the entire length of the spine. (18) Length from croup to dock: measure from the point of the croup to the base of the tail. (19) Chest width: feel for the humeral bones that project forward out of the chest and measure the distance between the outside edges. Keep the tape taut and straight. (20) Barrel girth at heart: settle the measuring tape where the girth of the saddle fits, directly behind the forelegs. (21) Barrel girth, maximum: measure around the barrel at greatest circumference. (22) Left forearm length: measure from the point of the elbow to the back of the kneecap. (23) Left fore cannon length: measure from the back of the kneecap to the ergot. (24) Left fore cannon mid-point circumference: measure around the cannon bone halfway between the knee and the ergot. (25) Left fore pastern length: measure from the bottom of the fetlock joint to the top of the coronet. (26) Left fore pastern circumference: measure around the pastern. (27) Left fore coronet circumference: measure around the coronet. (28) Left hoof length: measure from the coronet to the bottom of the hoof. (29) Left gaskin length: measure from the stifle joint to the point of the hock. Keep the tape taut and straight. (30) Left hind cannon length: measure from the point of the hock to the ergot (bottom of the fetlock). For measurements 31–35 of the hind limb (left hind cannon mid-point circumference, left hind pastern length, left hind pastern circumference, left hind coronet circumference and left hind hoof length), the same directions were followed as for measurements 24–28.

Body measures were recorded in units of inches, as the majority of owners were from within the United States and most familiar with this system of units. Results reported here have been converted to SI units. Horses were measured while standing on level, solid ground and restrained by either cross-ties or a halter and rope. Fewer than ~2% of horses (<20) were found to be non-compliant with the measurement protocol upon attempted measurement. These horses were not included in the study. Our sample collection protocol is approved by the Cornell University Institutional Animal Care and Use Committee as protocol number 2007-0155.
Breed effects are not independent of each other and of body condition. Therefore, we were interested in the patterns of skeletal size and shape in the horse species, we subjected our set of 33 quantitative measures to a PCA. We examined the first and second PCs in detail. Although subsequent PCs may contain biologically relevant patterns, we focused on just the first two PCs for two reasons. First, inspection of the scree plot of the proportion of variance explained by each PC shows a levelling off for principal component 3 (PC3) and onward (Fig. S4). Second, we also applied the commonly used ‘stopping rule’ stating that PCs retained for analysis should explain a greater proportion of variance than 1/n.

**Results**

We aim to identify the patterns of skeletal size and shape variation that exist in the domestic horse species. Our goal here is to derive robust and quantifiable morphologic traits that can subsequently be subjected to genetic analysis for discovery of genes contributing to these traits. We therefore collected 35 body measurements from the head, neck, trunk and limbs using the most readily identifiable body landmarks (Table S1).

To examine body size variation in the horse, we first looked at variation in the height at the withers across the horse species and then within specific breed populations. The extremes for height in our dataset are an American Miniature mare and stallion, with recorded withers heights of 74 and 76 cm, respectively, and two 2-m-tall Shire and Clydesdale stallions. We therefore observe over a 2.5-fold range in withers height in horses, irrespective of sex. The median withers height for the 1215 horses is 1.6 m, and half of all measured horses are between 1.5 and 1.7 m at the withers (Fig. 1). Despite the large variation in the species as a whole, each breed is characterized by a much narrower range of heights (Fig. 1). For example, our largest sample collection of 35 horses that met our inclusion criteria. We used the correlation matrix rather than covariance matrix because of the large variation in scale between our different measures (e.g. median heart girth is 1.9 m but median eye width is only 18 cm).

**Data handling and statistics**

Sample and trait data were initially captured onto paper forms. Data were entered to a database and double-checked by a second lab member. After collecting 35 measurements from each horse, we excluded two measures from further analysis. These measures were tail length, which is inconsistent in some breeds because of docking, and barrel girth, which is confounded by a horse’s body condition and in mares, pregnancy. Extreme outlier measurements were identified by calculating the median and interquartile range for each measurement in each breed with at least three samples. We then removed from the dataset 15 of the 22 breeds that had at least one measurement value greater than ten times the interquartile range from the median for that measurement in that breed (two Am. Miniature, one Appaloosa, three Connemara Pony, two Quarter Horse, one Standardbred, one Suffolk Punch, four Tenn. Walking Horse and one Thoroughbred). The remaining seven of the 22 horses (one each of Akhal-Teke, Fjord, Hanoverian, Irish Sport Horse, N.A. Spanish Colonial, Oldenburg and Paso Fino) with measures beyond this range were subjectively retained, because their breeds had extremely small interquartile ranges (often because of highly similar measurement values in relatively small sample sizes). We removed an additional six horses with measures that were between five and ten times the interquartile range from the median for their breed (one each of Lipizzan, Oldenburg, Standardbred, Trakehner and two Falabella). This was based on subjective assessment that the measurements were unrealistic outliers (gross errors in measurement or recording). We ignored a horse’s breed or sex when making a determination that it would be excluded from the dataset because of outlier measurements. Statistical analysis was conducted using R v2.10.1, PERL v5.10.1 and JMP v8.0 software (SAS Institute Inc.).
where \( n \) is the number of body measurements we collected (Peres-Neto et al. 2005). With 33 measurements, this means our retained PCs should explain greater than \( \sim 3\% \) of the variance. Principal component 1 (PC1) explains 65.9\% of the variance and principal component 2 (PC2) explains 6.4\%. Every one of the 33 body measurements has a positive factor-loading onto PC1, indicating a positive correlation (Fig. 2a). PC1 is therefore quantifying a proportional scaling of overall body size, where the head, neck, trunk and limbs increase or decrease in size coordinately.

Principal component 1 values for horses by breed are plotted in Fig. 3. The Falabella and American Miniature breeds have the smallest median PC1 values (Fig. 3, left side), while the Shire, Percheron, Clydesdale and American Belgian have the largest (Fig. 3, right side). PC1, by nature of its derivation from all 33 body measurements, quantifies a comprehensive assessment of body size information that no single measurement (like withers height, back length, heart girth or cannon length) is able to provide. Nevertheless, to the extent that single measurements are correlated with PC1 (Fig. 2a), they do provide some approximation of a horse’s PC1 size score. However, there are key differences in the relative ranks of some breeds by withers height vs. by PC1. For example, while the Holsteiner is the 4th tallest breed at the withers, it ranks 8th in terms of overall size scored by PC1. Moreover, the Ardennais is the 6th largest breed by PC1 but only the 26th tallest breed by withers height.

The second PC explains a much smaller proportion of the variance than does PC1 (6.4\% vs. 65.9\%; see Fig. S4), indicating that it identifies a more subtle pattern of variation in body shape. PC2 nevertheless has a clear biological interpretation: it quantifies variation across the entire body in bone thickness and body shape. This is apparent from the pattern of factor loadings (Fig. 2b). A horse with a large positive PC2 score has a short and thick neck, short and thick limb bones, a relatively short back, a broad eye and muzzle width, and long hooves. In contrast, a horse with a negative PC2 score demonstrates the opposite pattern.

In addition to the set of quantitative body measures, we also collected factor scores for several traits, including the degree of feathering on the lower limbs and head profile (Fig. S2). Both of these traits are anecdotally correlated with...
body type in the horse: draft breeds often have ‘Roman’ nose profiles and copious feathering. Factor scores for both of these traits were associated with PC2 score in our dataset ($P < 0.0001$ for each, ANOVA).

We collected a subjective measure of bone thickness for each horse we measured and asked whether this could serve as a proxy for PC2. Observers could discern the difference between the very thickest horses and all others; the subjective factor score of $5$, the thickest level, is predictive of a high PC2 score ($P < 0.0001$, ANOVA). However, observer scores for average or thin horses did not predict PC2 score well ($P = 0.6$, ANOVA), which illustrates the need for objective quantitative measurements to accurately capture this phenotype.

**Discussion**

Size and body conformation are critically important traits in nearly all horse breeds and are presumably under strong selection. Many breed registries select horses on functional criteria and encourage the breeding of horses with body types most suitable for those particular functions. The correct skeletal conformation is a key determinant of body type.

We find that body ‘size’ (PC1) and ‘thickness’ (PC2) have only limited variation within particular breeds compared to the total variation seen across our large panel of diverse breeds, consistent with selection within breed lines for preferred body types. The remaining variation within breeds is presumably because of ongoing segregation of causal alleles at contributory loci, measurement error and environmental factors.

We speculate that these data from 65 breeds may not have captured all skeletal variation in the horse species. Expansion of our sample sets from lightly sampled breeds and the addition of breeds that are geographically isolated, rare or under heavy selective pressure may provide additional patterns of skeletal variation.

PC1 is a single value for each horse that integrates measurement data from all over that horse’s body; it can easily be applied to QTL studies. Furthermore, PC1 cleanly separates horse breeds based on the more traditional description of body type. Small ponies are grouped together with low PC1 scores (Fig. 3), light horses have mid-values and the large draft breeds all have high median PC1 scores. Unsurprisingly, the fine-boned Thoroughbred and Akhal-Teke breeds scored the lowest for PC2, while the heavy draft Ardennais and Clydesdale breeds have very high PC2 scores (Figs 4 & 5). Notably, however, many of the small pony breeds also have high PC2 scores, reflecting their relatively thick build for their small stature. Two of the smallest breeds sampled, the Falabella and the American Miniature, rank very close to the large draft breeds on the PC2 axis, despite their vast differences in PC1 size. With the aspect of ‘body scale’ removed by the first principal component, unexpected relationships like this can be clearly quantified so they can be put to use in genetic studies.
With genome-scale sequencing power increasing very rapidly, the ‘limiting factor’ in many genetic studies of the near future may no longer be genotyping but rather the acquisition of high-quality phenotypes. Further, the likelihood of mapping success will still be largely determined by the quality and characterization of samples and phenotypes (Carlson et al. 2004). Here, we have described skeletal variation in the horse species as two key axes of variation, PC1 (size) and PC2 (thickness), thus providing a solid foundation for future mapping studies that will define the genetic control of body size and shape in the horse.

Acknowledgements
Thanks to Bethany Austin, Debbie Cook, John Eberth, Caroline Salter, Kate McDermott, Karla Brown and Elizabeth Kowalski, as well as other members of the Brooks and Sutter Laboratories who volunteered their time and effort to measure horses. Our appreciation to the Equine Research Team at The Pennsylvania State University for their efforts in measuring horses. We thank JoAnn Bellone for providing photographs. We greatly appreciate funding for this project from Cornell University’s Center for Vertebrate genomics, College of Agriculture and Life Sciences and the College of Veterinary Medicine. Finally, this project would not be possible without the support of horse owners who graciously provided their time and access to their animals.

Conflicts of interest
The authors have not declared any conflicts of interest.

References

Supporting information
Additional supporting information may be found in the online version of this article.
Figure S1 Body measurements collected for each horse. Figure S2 Histograms of counts of horses by (a) age; (b) body condition score where 1 = extremely underweight, 5 = ideal condition and 9 = overweight; (c) muzzle profile where 1 = deeply dish-faced (muzzle seen in profile curves in), 3 = straight and 5 = strongly Roman nosed (muzzle seen in profile curves out); (d) degree of feathering on lower

Figure 5 Two horses of similar height and body condition that exemplify the extremes of principal component 2 (PC2). Horse ‘a’, an Akhal-Teke, scored –1.73 for PC2, while horse ‘b’, an Ardennais, scored 6.04.
limbs where 1 = no feathering and 5 = copious feathering; (e) bite conformation where 1 = extreme overbite and 5 = extreme underbite; and (f) subjective assessment of skeletal thickness without consideration of the horse’s size, where 1 = a light, thin skeleton and 5 = a heavy, thick skeleton.

**Figure S3** Coefficient of variation calculated for each measurement by breed. The box plot central bar indicates the median breed coefficient of variation.

**Figure S4** Scree plot of the proportion of variance explained by each of the first ten principal components from the PCA performed on the 1215 horses measured.

**Table S1** Count of measured horses by breed and sex.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Microarray analysis after strenuous exercise in peripheral blood mononuclear cells of endurance horses

S. Capomaccio*•1, K. Cappelli*•1, E. Barrey†,‡, M. Felicetti*, M. Silvestrelli* and A. Verini-Supplizi*

*Centro di Studio del Cavallo Sportivo-DPDCV, University of Perugia, Via San Costanzo 4, 06126 Perugia, Italy. †Unité de Biologie Intégrative des Adaptations à l’Exercice – INSERM 902, Genopole Evry, France. ‡Génétique Animale et Biologie Intégrative, INRA, UMR1313, Jouy-en-Josas, France

Summary

It is known that moderate physical activity may have beneficial effects on health, whereas strenuous effort induces a state resembling inflammation. The molecular mechanisms underlying the cellular response to exercise remain unclear, although it is clear that the immune system plays a key role. It has been hypothesized that the physio-pathological condition that develops in athletes subjected to heavy training is caused by derangement of cellular immune regulation. The purpose of the present study was to obtain information on endurance horse gene transcription under strenuous conditions and to identify candidate genes causing immune system derangement. We performed a wide gene expression scan, using microarray technology, on peripheral blood mononuclear cells of ten horses chosen from high-level participants in national and international endurance races. The use of three different timepoints revealed changes in gene expression when post-effort samples (T1, taken immediately after the race; and T2, taken 24 h after the race) were compared with basal sample (T0, at rest). Statistical analysis showed no differences in gene expression between T0 and T2 samples, indicating complete restoration of homeostasis by 24 h after racing, whereas T1 showed strong modulation of expression, affecting 132 genes (97 upregulated, 35 downregulated). Ingenuity pathway analysis revealed that the main mechanisms and biofunctions involved were significantly associated with immunological and inflammatory responses. Real-time PCR was performed on 26 gene products to validate the array data.

Keywords gene expression, inflammation, peripheral blood mononuclear cells (PBMCs), qRT-PCR, stress.

Introduction

The horse is a natural athlete and good athletic performance is economically valuable. For historical and cultural reasons, selection of the best racehorses is not always based on objective parameters. However, all breeders and researchers agree that a winner is healthy and of high genetic value.

Knowledge of the molecular mechanism of exercise-induced stress is of fundamental importance in planning an appropriate training schedule to obtain better performance and to preserve the health of the animal. Exercise induces marked transient physiological changes in the body, including increased blood flow, increased body temperature, changes in hormone levels, synthesis of heat-shock proteins, oxidative stress, and the production of interleukins and receptors together with many other substances related to inflammatory processes. Some of these physiological effectors have been discussed with reference to their effects on the blood immune cell system, the study of which is termed immunocompetence (Gleeson 2007).

It is known that leucocytes are involved not only in eradication of pathogens but also play roles in wound repair, muscle growth, and other key developmental processes (Sabroe & Whyte 2007; Tidball & Wehling-Henricks 2007). In addition, it is now known that substantial changes in the gene expression profile pattern of circulating leucocytes occur rapidly during exercise in both humans
and horses (Cappelli et al. 2007; Barrey et al. 2006; Cappelli et al. 2009; Fehrenbach 2007). The impact of exercise on the immune response is increasingly seen as a novel mechanism by which physical activity modulates health, growth, and disease risk, owing to the immunomodulatory effects of exercise, training, and fatigue.

Regular moderate exercise is associated with a reduced incidence of infection compared with a completely sedentary state. However, prolonged bouts of strenuous exercise cause a temporary depression of various aspects of immune function, inducing an inflammation-like condition, depending on the intensity and duration of the exercise bout. Owing to the importance of this topic in medical and sports sciences, many studies have focused on the effects of exercise on white blood cells in peripheral blood, and the recent introduction of microarrays has promoted extensive research activity, particularly in this tissue, including studies on cytokine and receptor expression, apoptosis, and control of cell proliferation (Connolly et al. 2004; Zieker et al. 2005; Fehrenbach 2007; Buttner et al. 2007). Although strenuous exercise in horses has also been associated with increased susceptibility to infection (Folsom et al. 2001), little is known regarding the early immune responses involved in this phenomenon. It has been hypothesized that the physio-pathological condition that develops in horses subjected to heavy training (i.e. the overtraining syndrome) is caused by derangement of cellular immune regulation (Lakier Smith 2003).

Thus, the purpose of the present study was to analyse transcriptome modulation of the acute response during horse endurance races to identify candidate genes involved in immune system derangement. Similar to a human marathon runner, an endurance horse invests an extreme effort in racing, and peripheral blood mononuclear cells (PBMCs) may be the most appropriate cell type for investigation of immune and physiological changes connected to exhausting exercise.

Methods

Blood collection and RNA extraction

Animals were treated according to standard procedures governing animal welfare, with the permission of the owners, team veterinarians, and official racing veterinary commissions.

Blood samples were taken from the jugular veins of horses chosen from among high-level performers in national and international endurance races (90–120 km). Samples were obtained at three different time-points: before (T0), at the end of the race (T1) and 24 h after the race (T2). Only horses that passed FISE (Italian Equestrian Sports Federation) compulsory medical checks (pre-, during, and post-race) were included in this study. Ten endurance horses, comparable in terms of type and intensity of training, were recruited. Horse age varied from 5 to 13 years, with a mean of 6.8 ± 2.44. Veterinary examination included analysis of cardio-respiratory function, physical integrity, and metabolic condition, together with collection of a blood sample for a complete count.

Immediately after collection of blood, PBMCs were isolated by the Ficoll-Hypaque method (GE Healthcare). Total RNA was extracted using the Aurum Total RNA Fatty and Fibrous Tissue kit (Bio-Rad) according to the manufacturer’s instructions. Genomic DNA was eliminated using DNase supplied with the kit. Extracted RNA was quantified using the Quant-It RNA assay (Invitrogen) in a VersaFluor fluorometer (Bio-Rad), and RNA quality was verified by microchannel electrophoresis (RNA 6000 Nano LabChip; Bioanalyzer). Successful removal of DNA contaminants was demonstrated by absence of PCR amplification of the MC1R gene [GenBank accession number X98012; primers as described by Rieder et al. (2001)].

Production of the equine oligonucleotide microarray

Gene expression analysis was performed using an equine/murine oligonucleotide microarray including 384 equine transcripts, of which 50 were probes for the mitochondrial and 334 were probes for the nuclear genome (GEO Platform a.n.; GPL8349; http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL8349). The 50 probes of the equine mitochondrial genome were designed to measure expression of the 13 genes (two different probes per gene) coding for protein units of the respiratory chain, as well as 22 tRNAs and 2 rRNAs. This oligonucleotide microarray was synthesized by the method employed to construct open-access human and mouse long oligonucleotide microarrays (Barrey et al. 2009). Microarrays were spotted by a microarray service (LEFG; CEA, Evry, France) by printing equine probes suspended in a spotting buffer of 50% (v/v) dimethyl sulfoxide (DMSO) in TE onto hydrogel slides, using a Microgrid-II robot (Schott Nexterion).

Gene expression analysis using the oligonucleotide microarray

The race (T1) and the 24-h (T2) samples were hybridized against the basal sample (T0). The hybridization protocol has been described previously (Barrey et al. 2009). Fluorescence intensities of Cy3 (532 nm) and Cy5 (635 nm) were separately scanned using a GenePix 4000B laser scanner.

TIFF images (10 μm/pixel) were analysed using specific software (GenePix Pro 6.0 software) that automatically locates spots, with a manual control, and quantifies fluorescence intensity. For each microarray, fluorescence intensities of all spots corresponding to any particular gene constituted the data output. This information was imported into genomic data analysis software (Gene Spring) for
normalization prior to filtering and statistical analysis. For each sample, duplicate analysis (dye-swap) was performed and average fluorescence results were used. Next, all results were normalized by individual spot, and by microarray, using the Lowess regression method, to obtain a linear relationship between each sample and control. Finally, a gene expression ratio was obtained for each sample and gene:

\[
\text{Ratio} = \frac{\text{Expression of T1 or T2}}{\text{Expression at T0}}
\]

For statistical analysis, a log transformation ratio was used to normalize distribution. A list of significant genes was obtained by filtering genes with a cut-off t-test \(P\) value of <0.05, with Benjamini and Hochberg multiple testing correction, to reduce or eliminate false discovery (Benjamini & Hochberg 1995).

Reference gene selection and primer design

Housekeeping genes encoding succinate dehydrogenase complex subunit 1 (SDHA), and hypoxanthine phosphoribosyl transferase (HPRT), selected from a group of nine potential genes, were utilized as controls to accurately estimate gene expression during exercise-induced stress (Cappelli et al. 2008). Primers were designed based on available sequences using Primer3 software (Rozen & Skaltsky 2000). Mfold (Zuker 2003) was employed to check chosen sequences for absence of template secondary structure. Whenever possible, primers were located in different exons or at an exon–exon junction. Primer sequences, amplicon lengths, and amplification efficiencies are shown in Table S1.

Reverse transcription qPCR

Total RNA (500 ng) was retro-transcribed using a Superscript VILO cDNA Synthesis Kit (Invitrogen) according to the manufacturer’s instructions. PCR amplifying the horse \(\beta\)-actin gene (forward primer 5’-GAGCAAGAGGGCAT-

\[\text{CCTGA-3'}\]; reverse primer 5’-GGTCATCTTCTCGCGGTTGG-

\[\text{3'}\]; Gene ID: 100033878) was performed on each cDNA sample to confirm successful retro-transcription. RT-qPCR, with SYBR Green chemistry, was performed as described elsewhere (Cappelli et al. 2009). Raw \(C_q\) (quantification cycle) data from the MX3000P instrument were exported to a common exchange data file and analysed using GeneEx Pro (Kubista et al. 2006) to normalize data to those of the two housekeeping genes. During pre-processing, data were corrected for PCR efficiency and averaged over three repeats. The selected reference genes HPRT and SDHA were subsequently used to normalize \(C_q\) values, the relative levels of which were calculated using the maximum \(C_q\) numbers. Software R (R Development Core Team 2007) was used to perform analysis of variance (ANOVA) employing a non-linear mixed-effects model.

Data mining

Biological interpretation of the data was performed using Ingenuity pathway analysis (IPA) (http://www.ingenuity.com). This identified the most significant biological functions affected by changes in gene expression. Genes described in prior reports that were significantly associated with known biological functions were considered for analysis. For each gene, Fischer’s exact test was used to calculate a \(P\) value determining the probability that any biological function assigned to a particular dataset resulted from chance alone.

Results

After filtering the microarray data, we found 132 (97 upregulated, 35 downregulated) genes that showed significantly different (fold change from T0 of at least \(\pm 1.5\)) expression levels when T0 and T1 samples were compared. In the comparison between T0 and T2, only 19 (four upregulated, 15 downregulated) such genes were found. All array data are plotted and shown in supplementary material (Fig. S1). We focused our bioinformatics comparisons on the T0 vs T1 samples. We used IPA to identify the biological functions that were most significantly affected by changes in gene expression. As input, we chose all genes that were significantly modulated (\(P < 0.05\)), without filtering by fold change to avoid loss of gene interconnection information (Table S2). This resulted in 262 eligible network molecules that were used in functional analysis and to generate network clusters. The modulated gene list was classified by cellular functional categories and canonical and signalling systems, revealing that the main pathways and biofunctions affected were significantly related to immunological and inflammatory responses (Table 1), with a relative abundance of molecules involved in cell-to-cell cross-talk (i.e. cytokines). Furthermore, within the category ‘Inflammatory Response’, 129 molecules fell into the following subcategories: inflammatory response (59); modulation of levels of phagocytes and neutrophils (76); chemotaxis, migration, and activation of leucocytes (73); immune response (36); and immune cell trafficking (99). These results are supported by examination of the genes most affected in expression. Table 2 lists the genes with the greatest up- or downregulation.

Among individual IPA networks, the three with the highest number of involved genes are shown in Figs. 1–3. These are Network #1 Cell Death, Cellular Development, and Hematological System Development and Function (score 20, 35 molecules); Network #2 Inflammatory Response, Inflammatory Disease, and Antigen Presentation (score 8, 18 molecules); and Network #3 Inflammatory Response, Cell-To-Cell Signalling and Interaction, and Free Radical Scavenging (score 3, 20 molecules). These networks have not been established in the horse, but instead are inferred in the horse.
To validate array data, 30 samples (from ten horses at three timepoints) were analysed by RT-qPCR for the expression of 26 genes. The selected list included genes identified as being maximally either up- or downregulated in array analysis, or critically involved in the networks described (Table 3).

**Discussion**

We chose for our experimental design an equine/murine oligonucleotide microarray validated according to the methods previously described (Muller et al. 2004; E Mucher et al. 2006) and performed a wide gene expression scan on PBMCs from ten endurance horses.

The analysis of three different time-points revealed changes in gene expression when post-effort samples T1 and T2 were compared with resting samples, T0. Twenty-six of the maximally up- or downregulated genes were tested by RT-qPCR, and modulation in both intensity and direction for almost all genes (20 of 23 amplified) revealed good concordance of the two techniques.

The present work provides a snapshot of the acute response to strenuous exercise in horses. Many of the genes affected by exercise clearly belong to pathways with crucial roles in both the immune response and inflammation. Acute exercise is considered a stress factor inducing severe inflammation (Gleeson 2007) that involves many tissues, including the liver, muscles, and joints, but after chronic exercise training, the acute-phase reaction induced by a

---

**Table 1** Main functions identified by Ingenuity pathway analysis ($P > 0.0001$).

<table>
<thead>
<tr>
<th>Top biological functions</th>
<th>#Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases and disorders</td>
<td></td>
</tr>
<tr>
<td>Inflammatory response</td>
<td>129</td>
</tr>
<tr>
<td>Immunological disease</td>
<td>36</td>
</tr>
<tr>
<td>Connective tissue disorders</td>
<td>35</td>
</tr>
<tr>
<td>Inflammatory disease</td>
<td>37</td>
</tr>
<tr>
<td>Skeletal and muscular disorders</td>
<td>35</td>
</tr>
<tr>
<td>Molecular and cellular functions</td>
<td></td>
</tr>
<tr>
<td>Cell death</td>
<td>49</td>
</tr>
<tr>
<td>Cellular compromise</td>
<td>19</td>
</tr>
<tr>
<td>Cellular function and maintenance</td>
<td>36</td>
</tr>
<tr>
<td>Antigen presentation</td>
<td>72</td>
</tr>
<tr>
<td>Cellular movement</td>
<td>63</td>
</tr>
<tr>
<td>Physiological system development and function</td>
<td></td>
</tr>
<tr>
<td>Haematological system development and function</td>
<td>160</td>
</tr>
<tr>
<td>Hematopoiesis</td>
<td>86</td>
</tr>
<tr>
<td>Immune cell trafficking</td>
<td>99</td>
</tr>
<tr>
<td>Tissue development</td>
<td>41</td>
</tr>
<tr>
<td>Tissue morphology</td>
<td>57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top canonical pathways</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR5 signalling in macrophages</td>
<td>27/63 (0.429)</td>
</tr>
<tr>
<td>Hepatic fibrosis/hepatic stellate cell activation</td>
<td>40/103 (0.388)</td>
</tr>
<tr>
<td>T-cell receptor signalling</td>
<td>36/94 (0.383)</td>
</tr>
<tr>
<td>MIF regulation of innate immunity</td>
<td>14/31 (0.452)</td>
</tr>
<tr>
<td>April mediated signalling</td>
<td>16/37 (0.432)</td>
</tr>
</tbody>
</table>

**Table 2** Top 20 array genes ranked by fold change up and downregulated from T1.

<table>
<thead>
<tr>
<th>Fold change</th>
<th>P-value</th>
<th>Entrez Gene ID (Homo sapiens)</th>
<th>Symbol</th>
<th>Gene name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.924</td>
<td>2.20E–10</td>
<td>7850</td>
<td>IL1R2</td>
<td>Interleukin 1 receptor. type II</td>
<td>Transmembrane receptor</td>
</tr>
<tr>
<td>6.435</td>
<td>8.41E–12</td>
<td>4312</td>
<td>MMP1</td>
<td>Matrix metallopeptidase 1 (interstitial collagenase)</td>
<td>Peptidase</td>
</tr>
<tr>
<td>2.911</td>
<td>6.41E–08</td>
<td>3606</td>
<td>IL18</td>
<td>Interleukin 18 (interferon-gamma-inducing factor)</td>
<td>Cytokine</td>
</tr>
<tr>
<td>2.646</td>
<td>1.11E–09</td>
<td>85439</td>
<td>STON2</td>
<td>Stokin 2</td>
<td>Other</td>
</tr>
<tr>
<td>2.64</td>
<td>1.13E–07</td>
<td>1051</td>
<td>CEBPB</td>
<td>CCAAT/enhancer binding protein (C/EBP). beta</td>
<td>Transcription regulator</td>
</tr>
<tr>
<td>2.63</td>
<td>1.58E–06</td>
<td>2920</td>
<td>CXL2</td>
<td>Chemokine (C-X-C motif) ligand 2</td>
<td>Cytokine</td>
</tr>
<tr>
<td>2.446</td>
<td>4.04E–08</td>
<td>10468</td>
<td>FST</td>
<td>Follistatin</td>
<td>Other</td>
</tr>
<tr>
<td>2.259</td>
<td>6.34E–06</td>
<td>3576</td>
<td>IL8</td>
<td>Interleukin 8</td>
<td>Cytokine</td>
</tr>
<tr>
<td>1.981</td>
<td>1.51E–06</td>
<td>9021</td>
<td>SOCS3</td>
<td>Suppressor of cytokine signalling 3</td>
<td>Other</td>
</tr>
<tr>
<td>1.927</td>
<td>3.50E–12</td>
<td>4311</td>
<td>MME</td>
<td>Membrane metallo-endopeptidase</td>
<td>Peptidase</td>
</tr>
<tr>
<td>-1.489</td>
<td>1.64E–07</td>
<td>140499</td>
<td>UB2J2</td>
<td>Ubiquitin-conjugating enzyme E2. J2 (UBC6 homolog. yeast)</td>
<td>Enzyme</td>
</tr>
<tr>
<td>-1.494</td>
<td>1.33E–06</td>
<td>1231</td>
<td>CCR2</td>
<td>Chemokine (C-C motif) receptor 2</td>
<td>G-protein coupled receptor</td>
</tr>
<tr>
<td>-1.529</td>
<td>3.65E–06</td>
<td>29810</td>
<td>BAG3</td>
<td>BCL2-associated athanogene 3</td>
<td>Other</td>
</tr>
<tr>
<td>-1.537</td>
<td>6.49E–07</td>
<td>2205</td>
<td>FCER1A</td>
<td>Fc fragment of IgE. high affinity I. receptor for; alpha polypeptide</td>
<td>Transmembrane receptor</td>
</tr>
<tr>
<td>-1.544</td>
<td>3.22E–04</td>
<td>26401</td>
<td>MAP3K1</td>
<td>Mitogen-activated protein kinase kinase kinase 1</td>
<td>Kinase</td>
</tr>
<tr>
<td>-1.556</td>
<td>3.36E–04</td>
<td>6775</td>
<td>STAT4</td>
<td>Signal transducer and activator of transcription 4</td>
<td>Transcription regulator</td>
</tr>
<tr>
<td>-1.567</td>
<td>5.05E–06</td>
<td>6352</td>
<td>CCL5</td>
<td>Chemokine (C-C motif) ligand 5</td>
<td>Cytokine</td>
</tr>
<tr>
<td>-1.684</td>
<td>6.51E–05</td>
<td>13714</td>
<td>ELK4</td>
<td>ELK4. ETS-domain protein (SRF accessory protein 1)</td>
<td>Transcription regulator</td>
</tr>
<tr>
<td>-1.745</td>
<td>1.19E–06</td>
<td>15975</td>
<td>IFNAR1</td>
<td>Interferon (alpha, beta and omega) receptor 1</td>
<td>Transmembrane receptor</td>
</tr>
<tr>
<td>-20.803</td>
<td>4.64E–02</td>
<td>20463</td>
<td>COX7A2L</td>
<td>Cytochrome c oxidase subunit Vila polypeptide 2 like</td>
<td>Enzyme</td>
</tr>
</tbody>
</table>
single exercise session is modified, and the acute inflammatory response at the systemic level is attenuated, inducing changes that favour an anti-inflammatory milieu (Lira et al. 2009).

Blood cells are envisaged to be the ‘server’ of this network, playing decisional and causal roles leading to a pathological state or establishment of a new physiological homeostasis.

Thus, in the first IPA-generated network (Fig. 1), containing most of the mis-regulated molecules in the experiment, transient inflammation induction, which is immediately blocked by activation of feedback signals, is highlighted. Tumour necrosis factor (TNF), an acute-phase response molecule belonging to the pro-inflammatory cytokine family, has a central role in this network, and indeed it is involved in a wide spectrum of cellular processes during inflammation, together with interleukins like IL18, IL8, IL1β and CCL3 (chemokine C-C motif ligand 3), the genes encoding which were in the list of maximally upregulated genes by IPA analyses. Conversely, genes encoding other chemokines of the network, such as CCL5 and IFNGR1 (interferon gamma receptor 1), were downregulated in the array and also in RT-qPCR, together with STAT4 (signal transducer and activator of transcription), encoding a signal transducer through which IFN proteins stimulate cellular responses to cytokines (Trenerry et al. 2008). The general repression of inflammatory pathways

Figure 1 Ingenuity pathway analysis Network 1. Differentially regulated genes in PBMC involved in Cell Death, Cellular Development, and Hematological System Development and Function. The network is displayed graphically as nodes (gene/gene products) and edges (the biological relationships between nodes). Node colour intensity indicates the level of expression. Red, upregulated; green, downregulated, with respect to T1 vs T0. The fold change and P values are indicated under each node. The shape of a node indicates the functional class of the gene product, as shown in the legend.
induced by cytokine production was confirmed by the fact that a significant negative regulator of STAT and cytokine signalling, SOCS3, was one of the most upregulated molecules, which demonstrates the trend towards restoration.

Moreover, the poor induction of TNF (1.46 AFC at T1) may be caused by the action of muscle IL6, a potent suppressor of TNFα (Gleeson 2007). IL6 is abundant in circulation during exercise. We did not assess plasma or muscle IL6 expression levels, but IL6-dependent DNA-binding protein (IL6DBP), directly induced by IL6, was one of the most upregulated molecules in T1 samples, and it is known that exercise induces its transient increase in liver, contributing to signalling for PEPCK (phosphoenolpyruvate carboxykinase 1) gene transcription and increasing gluconeogenesis during exercise (Nizielski et al. 1996).

The second network (Fig. 2) shows upregulation of genes encoding receptors and proteins that play roles during activation of innate immunity (TRI1, MEFV), signal

Figure 2 Ingenuity pathway analysis Network 2. Differentially regulated genes in PBMC involved in inflammatory response, inflammatory disease and antigen presentation. The network is displayed graphically as nodes (gene/gene products) and edges (the biological relationships between nodes). Node colour intensity indicates the level of expression of genes: Red, upregulated; green, downregulated, with respect to T1 vs T0. The fold change and P values are indicated under each node. The shape of a node indicates the functional class of the gene product, as shown in Fig. 2.
Figure 3 Ingenuity pathway analysis Network 3. Differentially regulated genes in PBMC involved in inflammatory response, cell-to-cell signalling and interaction, and free radical scavenging. The network is displayed graphically as nodes (gene/gene products) and edges (the biological relationships between nodes). Node colour intensity indicates the level of expression of genes: Red, upregulated; green, downregulated, with respect to T1 vs T0. The fold change and P values are indicated under each node. The shape of a node indicates the functional class of the gene product, as shown in Fig. 2.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Entrez Gene ID (Homo sapiens)</th>
<th>Race vs Basal Array FC</th>
<th>P Value</th>
<th>Real time FC</th>
<th>P Value</th>
<th>24 h vs Basal Array FC</th>
<th>P Value</th>
<th>Real time FC</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL1R2</td>
<td>Interleukin 1 receptor type II</td>
<td>7850</td>
<td>7.924</td>
<td>2.20E-10</td>
<td>4.978</td>
<td>0.00E+00</td>
<td>1.485</td>
<td>2.16E-02</td>
<td>2.282</td>
<td>3.00E-04</td>
</tr>
<tr>
<td>MMP1</td>
<td>Matrix metallopeptidase 1 (interstitial collagenase)</td>
<td>4312</td>
<td>6.435</td>
<td>8.41E-12</td>
<td>2.722</td>
<td>0.00E+00</td>
<td>1.301</td>
<td>7.98E-03</td>
<td>1.152</td>
<td>3.69E-01</td>
</tr>
<tr>
<td>MMP13</td>
<td>Matrix metallopeptidase 13 (collagenase 3)</td>
<td>4322</td>
<td>4.585</td>
<td>2.12E-10</td>
<td>N/A</td>
<td>N/A</td>
<td>1.226</td>
<td>1.14E-03</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>IL18</td>
<td>Interleukin 18 (interferon-gamma-inducing factor)</td>
<td>3606</td>
<td>2.911</td>
<td>6.41E-08</td>
<td>2.102</td>
<td>0.00E+00</td>
<td>0.802</td>
<td>1.14E-01</td>
<td>-1.440</td>
<td>6.58E-02</td>
</tr>
<tr>
<td>STON2</td>
<td>Stonin 2</td>
<td>85439</td>
<td>2.646</td>
<td>1.11E-07</td>
<td>1.222</td>
<td>6.27E-02</td>
<td>0.802</td>
<td>7.69E-03</td>
<td>-1.056</td>
<td>6.39E-01</td>
</tr>
<tr>
<td>CEPB</td>
<td>CCAAT/enhancer binding protein (C/EBP) beta</td>
<td>1051</td>
<td>2.646</td>
<td>1.11E-07</td>
<td>3.350</td>
<td>0.00E+00</td>
<td>1.048</td>
<td>6.78E-01</td>
<td>-1.018</td>
<td>9.39E-01</td>
</tr>
<tr>
<td>CXCL2</td>
<td>Chemokine (C-X-C motif) ligand 2</td>
<td>2920</td>
<td>2.63</td>
<td>1.58E-06</td>
<td>1.988</td>
<td>0.00E+00</td>
<td>1.289</td>
<td>4.28E-02</td>
<td>1.244</td>
<td>1.67E-01</td>
</tr>
<tr>
<td>FST</td>
<td>Follistatin</td>
<td>10468</td>
<td>2.446</td>
<td>4.04E-08</td>
<td>-1.234</td>
<td>3.66E-01</td>
<td>0.969</td>
<td>5.08E-01</td>
<td>1.241</td>
<td>2.53E-01</td>
</tr>
<tr>
<td>IL8</td>
<td>Interleukin 8</td>
<td>3576</td>
<td>2.259</td>
<td>6.34E-06</td>
<td>2.118</td>
<td>0.00E+00</td>
<td>1.188</td>
<td>2.61E-01</td>
<td>1.312</td>
<td>1.46E-01</td>
</tr>
<tr>
<td>SOCS3</td>
<td>Suppressor of cytokine signalling 3</td>
<td>9021</td>
<td>1.981</td>
<td>1.51E-06</td>
<td>2.596</td>
<td>0.00E+00</td>
<td>1.100</td>
<td>1.37E-01</td>
<td>1.170</td>
<td>4.76E-01</td>
</tr>
<tr>
<td>MME</td>
<td>Membrane metallopeptidase</td>
<td>4311</td>
<td>1.927</td>
<td>3.50E-12</td>
<td>N/A</td>
<td>N/A</td>
<td>1.022</td>
<td>3.81E-01</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>IL1β</td>
<td>Interleukin 1, beta</td>
<td>3553</td>
<td>1.596</td>
<td>4.96E-04</td>
<td>2.099</td>
<td>1.00E-04</td>
<td>1.150</td>
<td>1.61E-01</td>
<td>1.497</td>
<td>3.26E-02</td>
</tr>
<tr>
<td>CCL3</td>
<td>Chemokine (C-C motif) ligand 3</td>
<td>6348</td>
<td>1.586</td>
<td>1.85E-03</td>
<td>1.436</td>
<td>5.30E-03</td>
<td>0.914</td>
<td>4.24E-01</td>
<td>-1.032</td>
<td>8.26E-01</td>
</tr>
<tr>
<td>TLR4</td>
<td>Toll-like receptor 4</td>
<td>7099</td>
<td>1.58</td>
<td>6.22E-05</td>
<td>1.528</td>
<td>0.00E+00</td>
<td>0.935</td>
<td>3.48E-01</td>
<td>-1.154</td>
<td>1.45E-01</td>
</tr>
<tr>
<td>GLUL</td>
<td>Glutamate-ammonia ligase (glutamine synthetase)</td>
<td>14645</td>
<td>1.577</td>
<td>2.15E-05</td>
<td>2.197</td>
<td>0.00E+00</td>
<td>1.023</td>
<td>6.79E-01</td>
<td>1.030</td>
<td>7.78E-01</td>
</tr>
<tr>
<td>UBE2J2</td>
<td>Ubiquitin-conjugating enzyme E2, J2 (UBG6 homolog, yeast)</td>
<td>140499</td>
<td>-1.489</td>
<td>1.64E-07</td>
<td>N/A</td>
<td>N/A</td>
<td>1.070</td>
<td>5.13E-02</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>CCR2</td>
<td>Chemokine (C-C motif) receptor 2</td>
<td>1231</td>
<td>-1.494</td>
<td>1.33E-06</td>
<td>2.293</td>
<td>0.00E+00</td>
<td>0.788</td>
<td>4.54E-04</td>
<td>1.158</td>
<td>4.64E-01</td>
</tr>
<tr>
<td>LCK</td>
<td>Lymphocyte-specific protein-tyrosine kinase</td>
<td>18618</td>
<td>-1.518</td>
<td>3.55E-03</td>
<td>-1.775</td>
<td>2.00E-04</td>
<td>0.890</td>
<td>1.32E-01</td>
<td>1.076</td>
<td>4.38E-01</td>
</tr>
<tr>
<td>BAG3</td>
<td>BCL2-associated athanogene 3</td>
<td>29810</td>
<td>-1.529</td>
<td>3.65E-06</td>
<td>-2.072</td>
<td>3.00E-04</td>
<td>1.020</td>
<td>7.68E-01</td>
<td>1.044</td>
<td>7.05E-01</td>
</tr>
<tr>
<td>FCER1A</td>
<td>Fc fragment of IgE high affinity I receptor for, alpha polypeptide</td>
<td>2205</td>
<td>-1.537</td>
<td>6.49E-07</td>
<td>-1.731</td>
<td>1.00E-04</td>
<td>0.747</td>
<td>9.87E-04</td>
<td>-1.284</td>
<td>1.43E-02</td>
</tr>
<tr>
<td>MAP3K1</td>
<td>Mitogen-activated protein kinase kinase kinase 1</td>
<td>26401</td>
<td>-1.544</td>
<td>3.22E-04</td>
<td>-1.149</td>
<td>3.71E-01</td>
<td>0.888</td>
<td>5.66E-02</td>
<td>1.113</td>
<td>4.33E-01</td>
</tr>
<tr>
<td>STAT4</td>
<td>Signal transducer and activator of transcription 4</td>
<td>6775</td>
<td>-1.556</td>
<td>3.36E-04</td>
<td>-1.458</td>
<td>3.22E-02</td>
<td>1.020</td>
<td>7.79E-01</td>
<td>1.402</td>
<td>8.20E-03</td>
</tr>
<tr>
<td>CCL5</td>
<td>Chemokine (C-C motif) ligand 5</td>
<td>6352</td>
<td>-1.567</td>
<td>5.05E-08</td>
<td>-1.286</td>
<td>1.24E-01</td>
<td>0.790</td>
<td>3.19E-04</td>
<td>-1.053</td>
<td>7.21E-01</td>
</tr>
<tr>
<td>ELK4</td>
<td>ELK4, ETS-domain protein (SRF accessory protein 1)</td>
<td>13714</td>
<td>-1.684</td>
<td>6.51E-05</td>
<td>-1.150</td>
<td>2.43E-01</td>
<td>0.913</td>
<td>1.44E-01</td>
<td>1.040</td>
<td>7.16E-01</td>
</tr>
<tr>
<td>IFNAR1</td>
<td>Interferon (alpha, beta and omega) receptor 1</td>
<td>15675</td>
<td>-1.745</td>
<td>1.19E-06</td>
<td>1.911</td>
<td>2.00E-04</td>
<td>0.812</td>
<td>1.62E-03</td>
<td>1.217</td>
<td>2.82E-01</td>
</tr>
<tr>
<td>COX7A2L</td>
<td>Cytochrome c oxidase subunit VIIa polypeptide 2 like</td>
<td>20463</td>
<td>-20.803</td>
<td>4.64E-02</td>
<td>-1.471</td>
<td>7.41E-02</td>
<td>0.183</td>
<td>N/A</td>
<td>1.181</td>
<td>2.98E-01</td>
</tr>
</tbody>
</table>

FC, fold change expressed in \log_2 base.
transduction (CXCR4, TRL1 and FKBPIA) and transcriptional regulation (MAF), together with the major pro-inflammatory cytokine IL1β, up-regulated in microarray and RT-qPCR. CXCR4 and Toll-like receptors (TLRs) were top molecules upregulated in IPA analyses and evident in Networks #1/#2 (Fig. 1: TRL2, TRL4 Fig. 2: TRL1, CXCR4) and, together with IL1β play an important role in mediating whole-body inflammation in antigen-presenting cells and in production of inflammatory cytokines and proteins (Gleeson et al. 2006). In addition, this second network highlights the pathway of "controlled" inflammation.

The final network (Fig. 3) shows IL8 to be the main gene of the cluster. IL8 is involved in immune cell communication and glucocorticoid receptor signalling, together with other cytokines (CCL5, CCL3); all three genes are strongly up-regulated in the microarray and RT-qPCR. IL8, together with MMP1 and MMP13 (matrix metalloproteinase 1 and 13), is in the list of maximally upregulated array molecules, indicating that lymphocyte trafficking and recruitment of progenitor stem cells from bone marrow (Cappelli et al. 2009) is probably activated during the response to acute exercise.

Of the maximally downregulated genes, CCR2 (encoding chemokine C-C motif receptor 2), the monocyte chemoattractant protein-1, involved in inflammatory processes, BCL2-associated athanogenes (BAG3) encoding cytoprotective proteins that bind to and negatively regulate Hsp70 family molecular chaperones (Homma et al. 2006), and the EST-domain protein (encoded by ELK4) transcription factor involved in regulation of gene expression were prominent. Moreover, we detected downregulation of the kinase signalling pathway through suppression of mitogen-activated kinase kinase 1 (encoded by MAP3K1) and lymphocyte-specific protein-tyrosine kinase (encoded by LCK).

The negative modulation of these genes and the suppression of chemokine receptor gene expression (CCL5 and CCR2), together with upregulation of the interleukin 1 receptor type II gene (IL1R2, the most highly upregulated gene in T1) that inhibits IL1 activity by acting as a decoy target for IL1 (Colotta et al. 1993), indicate again that after an increase in inflammation signalling during racing, a rapid restoration of homeostasis occurs over a few hours.

In our view, "basal homeostasis", which was deeply disrupted by major exercise, was restored in <24 h in the well-trained horses included in the present study, probably because the repetition of various exercises during training may increase adaptive responses. However, limited recovery from exercise, reduced substrate availability, reduction in protein synthesis, and the influence of other stressors would cause fatigue, decrease adaptation and may result in loss of function in cases of overreaching or overtraining (Steinacker et al. 2004).

In conclusion, this work presents a transcription snapshot of acute response to exercise in top endurance horses, and identifies the genes and the pathways involved in the relationship between acute inflammation, prolonged exercise and homeostasis restoration.

Acknowledgements

The authors thank Mr Gianluca Alunni for his valuable technical support. We are also grateful to Dr Alessandro Centinario, Dr Marcello Conte, and Mr Carlo Formica for allowing us to collect blood samples from animals they own or care for. MIPAF Sel Mol supported this work.

Conflicts of interest

The authors have declared no potential conflicts.

References


Rieder S., Taourit S., Mariat D., Langlois B. & Guérin G. (2001) Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (*Equus caballus*). *Mammalian Genome*, 12, 450–5.


### Supporting information

Additional supporting information may be found in the online version of this article.

#### Table S1

Primer pairs used in the real-time experiments.

#### Table S2

List of probes of the array significant modulated in T1 and T2.

#### Figure S1

‘Volcano plot’ of the horse array data for the T0 vs T1 (a) and T0 vs T2 comparisons (b).

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Mitochondrial DNA insertions in the nuclear horse genome

S. G. Nergadze, M. Lupotto, P. Pellanda, M. Santagostino, V. Vitelli and E. Giulotto
Dipartimento di Genetica e Microbiologia ‘Adriano Buzzati-Traverso’, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy

Summary

The insertion of mitochondrial DNA in the nuclear genome generates numts, nuclear sequences of mitochondrial origin. In the horse reference genome, we identified 82 numts and showed that the entire horse mitochondrial DNA is represented as numts without gross bias. Numts were inserted in the horse nuclear genome at random sites and were probably generated during the repair of DNA double-strand breaks. We then analysed 12 numt loci in 20 unrelated horses and found that null alleles, lacking the mitochondrial DNA insertion, were present at six of these loci. At some loci, the null allele is prevalent in the sample analysed, suggesting that, in the horse population, the number of numt loci may be higher than 82 present in the reference genome. Contrary to humans, the insertion polymorphism of numts is extremely frequent in the horse population, supporting the hypothesis that the genome of this species is in a rapidly evolving state.

Keywords horse genome, insertion polymorphism, mitochondrial DNA insertion.

Introduction

Integration of mitochondrial DNA (mtDNA) sequences into the nuclear genome gives rise to so-called numts (nuclear sequences of mitochondrial origin) (Lopez et al. 1994). Following the pioneering work in which the presence of genomic DNA hybridizing with mitochondrial DNA probes was identified more than 40 years ago (Du Buy et al. 1966; Du Buy & Riley 1967), numts were studied in different eukaryotes, from protists to mammals (Leister 2005; Hazkani-Covo et al. 2010 and references therein). Studies aimed at identifying the mechanisms of integration and evolutionary dynamics of numts are now greatly facilitated by the availability of whole nuclear genome sequences. These studies demonstrated that numt abundance varies in different eukaryotic genomes from those with no detectable numt sites such as some protists (Cyanidioschyzon merolae, Monosiga brevicollis, Naegleria gruberi and Thalassiosira pseudonana) and animals (Anopheles gambiae, Branchiostoma floridae, Ciona savignyi and Danio rerio), up to more than 2.1 Mb of numt sequences in the opossum Monodelphis domestica (Hazkani-Covo et al. 2010). Taking into account the density of numts in the genome, the highest estimate seems to be 0.1% (1 bp/kb) in the honeybee (Pamilo et al. 2007).

Insertion of mitochondrial DNA fragments into nuclear chromosomes, together with the insertion of transposons, retroviruses and telomeric-like sequences, is a driving force in evolution. While transposons and retroviruses are integrated by well-described mechanisms relying on specific proteins encoded by the inserted element itself (Kazazian 2004), and interstitial telomeric sequences are integrated by retrotranscription of the telomerase RNA component (Nergadze et al. 2004, 2007), the mechanisms responsible for the transfer of DNA fragments from mitochondria to nuclei are still elusive (Hazkani-Covo et al. 2010). Numt integration represents the prototype of exogenous insertions in the nucleus (Ricchetti et al. 2004). Studies on a yeast experimental system showed that the integration of mitochondrial sequences can occur during the repair of DNA double-strand breaks (Ricchetti et al. 1999). Numts, similarly to transposons and interstitial telomeres (Salem et al. 2003; Nergadze et al. 2004; Hazkani-Covo et al. 2010), have been successfully used in several studies as evolutionary markers, because the insertion events can be easily dated by comparing the sequence of orthologous loci in related species: the presence of a locus containing the insertion in one species and of the corresponding ‘empty’ locus, lacking the insertion, in another species demonstrates that the radiation of the two species occurred before the integration event. Besides phylogenetic studies, insertion polymorphisms are particularly informative for population studies, because the probability that two insertion events occur independently in
the same position is essentially equal to zero and, conversely, the precise removal of an inserted sequence is extremely rare.

The analysis of numts has been extensively carried out in humans and other primates. Of the 452 estimated numts in the human genome (Hazkani-Covo et al. 2010), at least 40 were inserted in the human lineage after the split from the chimpanzee lineage. Of these human-specific numts, 12 display insertion polymorphism in the human population, reflecting their recent integration; thus, the colonization of nuclear genomes by mitochondrial DNA sequences is an ongoing process (Zischler et al. 1995; Thomas et al. 1996; Hazkani-Covo et al. 2010). It was also shown that 23 human-specific numts are inserted into known or predicted genes, mainly in introns (Ricchetti et al. 2004). Although insertions of mtDNA fragments into the nuclear genome usually appear as neutral mutations, in rare cases numt insertions into genes are associated with human diseases: a 251-bp numt insertion into the gene for plasma factor VII caused a splice site junction abnormality (Borenzstajn et al. 2002); a 72-bp insertion into exon 14 of the GLI3 gene created a pre-mature stop codon resulting in a truncated protein product. This mutation was shown to be responsible for a sporadic case of Pallister-Hall syndrome detected in a patient that was exposed to Chernoby accident (Turner et al. 2003), hinting that environmental factors inducing DNA double-strand breaks may facilitate numt insertions. A 93-bp numt insertion into the MCOLN1 gene caused an inherited case of mucolipidosis IV (Goldin et al. 2004), and a 36-bp numt insertion into the USH1C gene is associated with Usher syndrome 1c (Chen et al. 2005). In one case, a 41-bp fragment from the mitochondrial 12S rRNA gene was integrated at the breakpoint junction of a familial constitutional reciprocal translocation t(9;11)(p24;q23) (Willett-Brozick et al. 2001).

The insertion of mitochondrial sequences into the nuclear genome could occur in principle both by direct DNA transfer and by cDNA-mediated transfer. It was proposed that, during the repair of a double-strand break in the nuclear DNA, either a mitochondrial DNA duplex fragment or a cDNA derived from retrotranscription of mitochondrial RNA can be recruited as filler DNA by the non-homologous end-joining repair process (Shay & Werbin 1992; Mourier et al. 2001; Ricchetti et al. 2004; Leister 2005; Hazkani-Covo et al. 2010), similarly to what we have proposed for the insertion of interstitial telomeric repeats (Nergadze et al. 2004, 2007).

The transfer of functional genes from mitochondrial to nuclear genomes has occurred extensively during evolution (Henze & Martin 2001), leading to the very streamlined mitochondrial genome of today’s mammals. It is conceivable that we are now observing the continuation of this type of event that is mainly transferring non-functional DNA.

The genomes of the species from the genus Equus (horses, asses and zebras) are in a rapidly evolving phase (Ryder et al. 1978; Wichman et al. 1991; Trifonov et al. 2008). For instance, the formation of evolutionarily novel centromeres (Carbone et al. 2006) still lacking satellite repeats has been recently described (Piras et al. 2009; Wade et al. 2009; Piras et al. 2010). This situation makes the study of numts particularly attractive to obtain molecular data on the evolutionary dynamics of these genomes.

In this work, the availability of the horse genome sequence offered us the opportunity to study numt insertions and their polymorphism in this species.

Materials and methods

Search of numts within the horse nuclear genome

The horse mitochondrial database sequence (acc. No.: NC_001640) was used as query for BLAT search (BLAST-Like Alignment Tool; Kent 2002; http://genome.ucsc.edu/cgi-bin/hgBlat) to reveal all numt loci present in the Equus caballus nuclear genome database sequence (September 2007 Broad/equCab2 assembly) (Wade et al. 2009). BLAT finds sequences with at least 95% similarity that are 40 bp or longer and may be interrupted. It may miss genomic alignments that are more divergent or shorter, although it will find perfect sequence matches of as few as 22 nucleotides; BLAT is particularly useful to reveal sequences that are interrupted relatively to the query. As numt sequences undergo post-insertional rearrangements, such as deletions or insertions of interspersed fragments and insertion of known repetitive elements, BLAT is a good tool for locating them (Lascaro et al. 2008). We found 82 numt loci containing 32 or more nucleotides homologous to horse mtDNA sequence. A complete list and description of the numt loci used for this analysis is presented in Table 1. It is important to underline here that Hazkani-Covo et al. (2010) detected 203 numt loci in the horse genome using a BLASTN search with a threshold E-value of 0.0001.

Source of DNA and polymerase chain reaction amplification

Unique primer pairs were deduced from the sequences flanking 12 numt loci (Table 1) inserted within single-copy sequence (Table 1). The sequence of the primers is external to numt insertions and is reported in Table 2. Genomic DNA was extracted from 20 unrelated horses. For DNA extraction, blood was collected from 17 animals used for show-jumping competitions; these horses derive from different European stud farms and, according to their pedigree chart, do not share common ancestors up to the third generation. DNA was also extracted from fibroblast cell
<table>
<thead>
<tr>
<th>Numt name</th>
<th>Chromosome</th>
<th>Starting nucleotide of numt insertion</th>
<th>Length of numt insertion</th>
<th>Starting nucleotide in horse mtDNA sequence</th>
<th>Identity with horse mtDNA (%)</th>
<th>Sequence analysis of numt</th>
<th>Analysis of flanking sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1&lt;sup&gt;1&lt;/sup&gt;</td>
<td>1</td>
<td>21812131</td>
<td>916</td>
<td>41</td>
<td>96.2</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-2</td>
<td>1</td>
<td>69705164</td>
<td>1684</td>
<td>4943</td>
<td>90.2</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-3</td>
<td>1</td>
<td>93457156</td>
<td>68</td>
<td>4278</td>
<td>92.7</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-4</td>
<td>1</td>
<td>108174892</td>
<td>1100</td>
<td>14027</td>
<td>84.1</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-5&lt;sup&gt;2&lt;/sup&gt;</td>
<td>1</td>
<td>111356763</td>
<td>655</td>
<td>149</td>
<td>98.4</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-6</td>
<td>1</td>
<td>115458270</td>
<td>64</td>
<td>1994</td>
<td>96.9</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-7</td>
<td>1</td>
<td>139009433</td>
<td>5525</td>
<td>14563</td>
<td>93.4</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-8</td>
<td>1</td>
<td>155181861</td>
<td>36</td>
<td>5322</td>
<td>94.5</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-9</td>
<td>1</td>
<td>155181895</td>
<td>67</td>
<td>9820</td>
<td>94.1</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>1-10</td>
<td>1</td>
<td>160432288</td>
<td>8319</td>
<td>6233</td>
<td>91.3</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>2-1</td>
<td>2</td>
<td>663306</td>
<td>113</td>
<td>242</td>
<td>90.0</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>2-2</td>
<td>2</td>
<td>65735556</td>
<td>2606</td>
<td>145</td>
<td>84.1</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>2-3</td>
<td>2</td>
<td>110796771</td>
<td>80</td>
<td>11578</td>
<td>93.8</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>3-1&lt;sup&gt;3&lt;/sup&gt;</td>
<td>3</td>
<td>3487086</td>
<td>73</td>
<td>1230</td>
<td>98.8</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>3-2</td>
<td>3</td>
<td>91988479</td>
<td>53</td>
<td>7734</td>
<td>90.0</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>3-3</td>
<td>3</td>
<td>102069272</td>
<td>83</td>
<td>8921</td>
<td>78.4</td>
<td>Del</td>
<td>Flanked by LINE/L1</td>
</tr>
<tr>
<td>3-4</td>
<td>3</td>
<td>111796087</td>
<td>71</td>
<td>10685</td>
<td>91.6</td>
<td>NR</td>
<td>Interrupts LINE/L1</td>
</tr>
<tr>
<td>4-1</td>
<td>4</td>
<td>37856882</td>
<td>265</td>
<td>13147</td>
<td>99.3</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>4-2</td>
<td>4</td>
<td>50562363</td>
<td>73</td>
<td>560</td>
<td>94.0</td>
<td>NR</td>
<td>Interrupts LINE/L1</td>
</tr>
<tr>
<td>4-3</td>
<td>4</td>
<td>54379375</td>
<td>89</td>
<td>397</td>
<td>89.3</td>
<td>NR</td>
<td>Flanked by LINE/L1</td>
</tr>
<tr>
<td>4-4</td>
<td>4</td>
<td>75685370</td>
<td>119</td>
<td>478</td>
<td>90.2</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>4-5</td>
<td>4</td>
<td>91210387</td>
<td>148</td>
<td>11596</td>
<td>87.9</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>5-1&lt;sup&gt;3&lt;/sup&gt;</td>
<td>5</td>
<td>87592742</td>
<td>492</td>
<td>16340</td>
<td>94.1</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>5-2</td>
<td>5</td>
<td>7272609</td>
<td>9456</td>
<td>1105</td>
<td>83.2</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>5-3</td>
<td>5</td>
<td>89850192</td>
<td>47</td>
<td>9863</td>
<td>93.7</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>5-4</td>
<td>5</td>
<td>91318326</td>
<td>2003</td>
<td>5778</td>
<td>85.9</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>6-1&lt;sup&gt;3&lt;/sup&gt;</td>
<td>6</td>
<td>62865624</td>
<td>488</td>
<td>16340</td>
<td>94.1</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>6-2&lt;sup&gt;3&lt;/sup&gt;</td>
<td>6</td>
<td>67014027</td>
<td>170</td>
<td>7242</td>
<td>98.9</td>
<td>NR</td>
<td>Interrupts LINE/L2</td>
</tr>
<tr>
<td>6-3</td>
<td>6</td>
<td>29621487</td>
<td>320</td>
<td>8684</td>
<td>83.7</td>
<td>NR</td>
<td>Interrupts LINE/L2</td>
</tr>
<tr>
<td>6-4</td>
<td>6</td>
<td>29622055</td>
<td>398</td>
<td>9684</td>
<td>88.1</td>
<td>NR</td>
<td>Interrupts LINE/L2</td>
</tr>
<tr>
<td>8-1</td>
<td>8</td>
<td>26539775</td>
<td>116</td>
<td>11602</td>
<td>84.3</td>
<td>Del</td>
<td>Flanked by LINE/L2</td>
</tr>
<tr>
<td>8-2</td>
<td>8</td>
<td>84205943</td>
<td>884</td>
<td>1868</td>
<td>87.7</td>
<td>NR</td>
<td>Interrupts LINE/L2</td>
</tr>
<tr>
<td>9-1</td>
<td>9</td>
<td>5390309</td>
<td>96</td>
<td>11579</td>
<td>92.6</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>9-2</td>
<td>9</td>
<td>30275929</td>
<td>324</td>
<td>1</td>
<td>83.5</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>9-3</td>
<td>9</td>
<td>30279245</td>
<td>6716</td>
<td>7354</td>
<td>79.3</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>9-4</td>
<td>9</td>
<td>38479521</td>
<td>54</td>
<td>4843</td>
<td>94.5</td>
<td>NR</td>
<td>Interrupts LINE/L1</td>
</tr>
<tr>
<td>9-5</td>
<td>9</td>
<td>43143577</td>
<td>70</td>
<td>15406</td>
<td>92.9</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>9-6</td>
<td>9</td>
<td>43794027</td>
<td>57</td>
<td>1</td>
<td>91.3</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>9-7</td>
<td>9</td>
<td>57719166</td>
<td>95</td>
<td>146</td>
<td>86.3</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>9-8</td>
<td>9</td>
<td>61532187</td>
<td>453</td>
<td>13880</td>
<td>81.8</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>10-1</td>
<td>10</td>
<td>12352453</td>
<td>463</td>
<td>10767</td>
<td>85.4</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>10-2</td>
<td>10</td>
<td>54855904</td>
<td>60</td>
<td>12488</td>
<td>100.0</td>
<td>NR</td>
<td>Interrupts SINE/ERE1</td>
</tr>
<tr>
<td>10-3</td>
<td>10</td>
<td>73429678</td>
<td>78</td>
<td>1736</td>
<td>92.4</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>10-4</td>
<td>10</td>
<td>73703521</td>
<td>43</td>
<td>2820</td>
<td>90.7</td>
<td>Del</td>
<td>Flanked by LINE/L1</td>
</tr>
<tr>
<td>10-5</td>
<td>10</td>
<td>75481662</td>
<td>537</td>
<td>2698</td>
<td>99.1</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>11-1</td>
<td>11</td>
<td>12667007</td>
<td>45</td>
<td>14936</td>
<td>91.2</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>11-2</td>
<td>11</td>
<td>28842472</td>
<td>41</td>
<td>8382</td>
<td>92.7</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>11-3</td>
<td>11</td>
<td>46019891</td>
<td>53</td>
<td>6980</td>
<td>96.3</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>12-1</td>
<td>12</td>
<td>20025244</td>
<td>259</td>
<td>14461</td>
<td>86.2</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>14-1&lt;sup&gt;3&lt;/sup&gt;</td>
<td>14</td>
<td>16213507</td>
<td>845</td>
<td>14275</td>
<td>99.6</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>14-2</td>
<td>14</td>
<td>39907520</td>
<td>32</td>
<td>4774</td>
<td>100.0</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>14-3</td>
<td>14</td>
<td>43692682</td>
<td>87</td>
<td>14277</td>
<td>92.0</td>
<td>NR</td>
<td>Single-copy sequence</td>
</tr>
<tr>
<td>14-4</td>
<td>14</td>
<td>91148501</td>
<td>4457</td>
<td>1</td>
<td>84.4</td>
<td>Del</td>
<td>Single-copy sequence</td>
</tr>
</tbody>
</table>

Nergadze et al.

Table 1 Numts identified in the horse reference genome.<sup>1</sup>
lines established from the skin of three different slaughtered animals. Genomic DNA (50–100 ng) was amplified by polymerase chain reaction as previously described (Nergadze et al. 2004).

Reaction products were analysed by electrophoresis on 1–2% agarose gels, and fragments of interest were gel-extracted and purified (Wizard SV gel and PCR clean-up system; Promega Corporation) for sequencing.

Sequence generation and analysis

Gel-purified amplification products were sent to sequencing facilities for direct sequencing with both the forward and the reverse amplification primers. To prove specifcity of amplified fragments, their sequences were compared with the corresponding loci from the horse genome database sequence (http://genome.ucsc.edu/cgi-bin/hgGateway?org=Horse&db=equCab2&hgsid = 153566333).

The REPEAT-MASKER software at EMBL (http://woody.embl-heidelberg.de/repeatmask/) was used to identify known repetitive elements (such as SINEs, LINEs, microsatellites, etc.). Sequences were aligned using the Multiple sequence Alignment software, MultAlin (http://prodes.toulouse.inra.fr/multalin/multalin.html).

Results and discussion

Search of numts in the horse genome reference sequence

We used the horse mitochondrial sequence from GenBank as query for a BLAT search aimed at revealing the nuclear genome loci containing mitochondrial DNA insertions (numts) in the horse genome reference sequence. Eighty-two such loci with 78–100% identity to mtDNA were found. Numt length varies from 32 (at nt 39907520 of horse
chromosome 14) to 9456 bp (at nt 7272609 of horse chromosome 5). In the horse nuclear reference genome, 77191 bp (0.0029%) are composed of sequences of mitochondrial origin. A complete list of the 82 numt insertions found in this search is reported in Table 1. Twenty-nine of the 82 numts (35%) are inserted into introns of known or predicted genes. The remaining 53 numts are located at intergenic sites. No numt insertions were observed within exons. Hazkani-Covo et al. (2010) performed a BLAST search and found 203 BLAST hits. The discrepancy between the number of numts found by Hazkani-Covo et al. and by us derives from the different approaches used (see Materials and methods). In addition, while we performed our search on the genome of a single individual (the reference genome of the Thoroughbred mare Twilight), the other authors carried out a BLAST search that includes sequence runs from the genomes of other individuals; because of the high frequency of numt insertion polymorphisms that we demonstrate in the present work, it is likely that the search of Hazkani-Covo et al. (2010) may have included some numts that are absent in Twilight.

Distribution and sequence analysis of numts

Contrary to the situation described in the human and other genomes, in which numt insertion was sometimes followed by duplication of the surrounding region, leading to duplication of the numt loci, in the horse reference genome we did not find duplicated numt loci. This observation matches with the relatively low abundance of duplications present in the horse genome in comparison with other genomes (Wade et al. 2009).

The overall distribution in the horse reference genome is one numt per 33 Mb, with values ranging between less that one per 124 Mb (the length of chromosome X, which seems to be devoid of numt insertions) and 1 per 10.5 Mb (chromosome 9) (Table 1). It should be noted, however, that owing to their insertion polymorphism, not all numt loci are present in the reference genome. In conclusion, although some chromosomes seem more prone to numt insertion than others, similar to human chromosome 18 and Y (Ricchetti et al. 2004), given the relatively small number of insertions, our data do not allow us to identify the presence of chromosomal regions clearly favouring or disfavouring numt insertions. In a similar manner to interstitial telomeric repeat insertions (Azzalin et al. 2001; Nergadze et al. 2004, 2007; Ruiz-Herrera et al. 2008), numt insertions (Ricchetti et al. 1999, 2004; Hazkani-Covo & Covo 2008) are the consequence of DNA double-strand break repair; therefore, we may expect that these events also occur preferentially at chromosomal regions more prone to breakage.

To detect mutations that may have occurred in the mitochondrial DNA following its insertion in the horse genome, we compared the sequence of the 82 numts with the horse mitochondrial DNA reference sequence. However, it is important to underline that mitochondrial DNA is characterized by high mutation rates and polymorphism and that the identity values reported in Table 1 derive from the comparison of one nuclear genome (the genome of Twilight) with one mitochondrial DNA sequence (the reference horse mtDNA sequence). From Table 1, it appears that four relatively short numts share an identical sequence with the corresponding mitochondrial genome fragment (numts 10-2, 14-2, 21-2 and 31-1); all other numts contain variable numbers of point mutations (transitions, transversions, nucleotide additions or deletions), with identity values ranging between 78.4% and 99.6%. Among these, 16 numts also underwent gross rearrangements (Table 1); the sequence of four numts representative of different situations is depicted in Fig. 1. In Fig. 1a, an example of a numt in which no gross rearrangement has occurred is reported. Nine numts (1-2, 1-7, 2-2, 5-4, 8-1, 9-2, 14-4, 21-1 and 27-3) contain non-continuous pieces of mitochondrial DNA in the same orientation; these rearrangements are probably attributed to deletions (23–2070 bp) of internal fragments from integrated numt sequences; an example is reported in Fig. 1b. Four numts (1-10, 14-5, 21-3 and 24-1) are interrupted by 17- to 2141-bp sequences; the insertion of 259 bp from an

<table>
<thead>
<tr>
<th>Numt loci</th>
<th>Primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>GGAAGATTCTGTGCTGGACT</td>
</tr>
<tr>
<td>1-5</td>
<td>AGTCACATCAGCAGGAAGAAC</td>
</tr>
<tr>
<td>3-1</td>
<td>TGATGGAATTAGCAGTGGAG</td>
</tr>
<tr>
<td>5-1</td>
<td>CGACTGCTTCTGCTAACCTC</td>
</tr>
<tr>
<td>6-1</td>
<td>GCAGAATTCTAGTAATAGAAGC</td>
</tr>
<tr>
<td>6-2</td>
<td>CTGATGCCCTGAAAGTGGAAAG</td>
</tr>
<tr>
<td>14-1</td>
<td>CTGATCAAGTACATCACCAG</td>
</tr>
<tr>
<td>19-3</td>
<td>TGCTACATTCCCAAACACTTC</td>
</tr>
<tr>
<td>21-2</td>
<td>GCCAGAAGGAGATGCTGAAG</td>
</tr>
<tr>
<td>23-1</td>
<td>GCCATGTCATCTACAGCAGAATC</td>
</tr>
<tr>
<td>26-1</td>
<td>GCAACTGTGCTGAGAAGCAXAT</td>
</tr>
<tr>
<td>31-2</td>
<td>TGATGGCTGGAAGAGGTA</td>
</tr>
</tbody>
</table>

Table 2 Primers used in PCR numt analysis.
(a) No rearrangement

<table>
<thead>
<tr>
<th>mtDNA</th>
<th>numt</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>13147</td>
<td>4-1</td>
<td>AGGCAAGCTCTTTAATTTGGCAACATTT...CAAAATATAAGCACTACCGAGCCCTCGCTAAAGCTATCA</td>
</tr>
</tbody>
</table>

(b) Deletion

<table>
<thead>
<tr>
<th>mtDNA</th>
<th>numt</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>11174</td>
<td>21-1</td>
<td>TCAATACATATTCTGGC...TCATACATATTCAATCGACCAAGGGGA...ACAGCATTAAACCTTACATTACGAGAAAA</td>
</tr>
</tbody>
</table>

(c) Insertion

<table>
<thead>
<tr>
<th>mtDNA</th>
<th>numt</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>6233</td>
<td>1-10</td>
<td>CATGTTTACAGTAGGG...AAAATAAGAAGAAGAAGAGG...ACGGAAACCCCTCT</td>
</tr>
</tbody>
</table>

(d) Deletion and insertion

<table>
<thead>
<tr>
<th>mtDNA</th>
<th>numt</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>15701</td>
<td>20-1</td>
<td>AGTCAA...GGACATA...GGACAAACGATGCAAC...GTGCACCTCTGACAC...GTGAC...GTGATGTA</td>
</tr>
</tbody>
</table>

![Figure 1](image)

**Figure 1** Post-integration modifications of numt loci. The sequence of four numt insertions (red nucleotides) is compared with the sequence of the corresponding region of the horse mitochondrial DNA (blue nucleotides); flanking and inserted nucleotides are in black. (a) At the numt locus on chromosome 4 (nt 37856882), the 265 bp from mtDNA (see Table 1) did not undergo gross rearrangements. (b) At the numt locus on chromosome 21 (nt 8672162), 64 nt were deleted following insertion. (c) At the numt locus on chromosome 1 (nt 160432288), a 259-nt fragment from the repetitive element ERE2_SINE/tRNA was inserted. (d) At the numt locus from chromosome 20 (nt 3493253), a 207-bp fragment was deleted and an 85-bp random sequence was inserted.

![Figure 2](image)

**Figure 2** Horse mtDNA coverage of numt fragments. The mitochondrial protein coding genes and the control region are depicted in different colours on the horse mitochondrial DNA map (purple: rRNA genes, red: tRNA genes, grey: control region, green: protein genes). Each black line represents a numt insertion in the horse genome reference sequence.

ERE2_SINE/tRNA element is reported in Fig. 1c. Three numts (5-2, 9-3 and 20-1) contain more complex rearrangements, including both insertions and deletions (one example is shown in Fig. 1d).

In Fig. 2, a map of the horse mtDNA with the distribution of its genes is depicted. Each external black line represents the sequence of a numt. The map shows that the entire horse mitochondrial genome is represented as...
numts, without gross bias. In particular, the ribosomal RNA genes do not seem to be conspicuously more represented, contrary to what is suggested for other mammals (Qu et al. 2008).

It has been suggested that numt fragments could derive from the insertion of either retrotranscribed RNAs or mtDNA fragments. The available data do not allow us to distinguish between these two possibilities. The mechanism of mtDNA transcription involves the processing of two long complementary transcripts covering the entire genome, and the transcript of the light strand is fragmented into a number of short primers; therefore, the possibility of retrotranscription must certainly be considered, even if this is made less likely by the fact that polyA tracts appear to be absent in the inserts, similarly to what has already been described in humans (Leister 2005).

All of these results suggest that mtDNA sequences have appeared in nuclear intrachromosomal locations by transfer of mtDNA fragments into DNA double-strand break sites as filler DNA derived either directly from mtDNA fragmentation or by retrotranscription of mitochondrial RNA.

Analysis of the numt insertion sites

Analysis of the sequences flanking the insertions showed that the GC content at numt integration sites did not differ from that of the entire Equus caballus genome: 39.62% using 1-kb intervals spanning the insertion sites, and 39.26% using 10-kb intervals, when compared with 40.27% (Leeb et al. 2006) (data not shown).

REPEAT-MAKER analyses of the sequences flanking the numt insertions have shown that in 64 cases of 82 (78%) mtDNA sequences are inserted within single-copy sequence (Table 1). Because 26% of the horse genome is made up of LINEs and SINEs (Wade et al. 2009), and 18 numts of 82 (22%) are inserted within or are flanked by one of these repetitive elements, numt insertions seem to occur essentially at random sites.

Polymorphism of numt insertion in the horse

As mentioned previously, a large body of evidence suggests that the horse genome is in a rapidly evolving phase; therefore, we might expect that several numt insertions may have occurred in the horse lineage in relatively recent evolutionary times and may not have been fixed yet in the horse species, giving rise to polymorphism. To test this possibility, 12 numt loci (marked in Table 1) were analysed by PCR in 20 unrelated horse individuals, using the primers listed in Table 2. The numt inserts at these loci comprised between 73 and 1286 bp are >87% identical to the horse mtDNA and are flanked by single-copy sequences on which the primer pairs were designed.

The results (Table 3) showed that, for six of the 12 numts (1-1, 1-5, 6-1, 21-2, 23-1 and 26-1), all individuals analysed were homozygous for the presence of the numt (numt+/+), suggesting that these insertions are either very frequent or even fixed in the horse species. The remaining six numts (50% of numts analysed) are characterized by insertion polymorphism, as in the studied individuals null alleles (i.e. amplified sequences not containing the numt insertion) are also present. The observation of these ‘empty’ alleles at 50% (six out of 12) of numt loci strongly supports the hypothesis that the horse genome is evolving rapidly, at least in comparison with the human genome, in which the fraction of polymorphic numt loci is lower (Hazkani-Covo et al. 2010). The frequency of null alleles in the 20 individuals is variable, ranging between 7.5% (numt 5-1) and 100% (numt 14-1). In the case of numt 14-1, only null alleles were found in the 20 horses, and this numt locus was

<table>
<thead>
<tr>
<th>Numt name</th>
<th>Length of PCR fragments (bp)</th>
<th>Number of alleles (%)</th>
<th>Homozygous individuals (%)</th>
<th>Heterozygous individuals (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>numt+</td>
<td>numt−</td>
<td>numt</td>
<td>null</td>
</tr>
<tr>
<td>1-1</td>
<td>1230</td>
<td>NF</td>
<td>40 (100)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>1-5</td>
<td>978</td>
<td>NF</td>
<td>40 (100)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>3-1</td>
<td>585</td>
<td>507</td>
<td>36 (90.0)</td>
<td>4 (10.0)</td>
</tr>
<tr>
<td>5-1</td>
<td>782</td>
<td>290</td>
<td>37 (92.5)</td>
<td>3 (7.5)</td>
</tr>
<tr>
<td>6-1</td>
<td>745</td>
<td>NF</td>
<td>40 (100)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>6-2</td>
<td>1522</td>
<td>1346</td>
<td>5 (12.5)</td>
<td>35 (87.5)</td>
</tr>
<tr>
<td>14-1</td>
<td>694</td>
<td>NF</td>
<td>0 (0)</td>
<td>40 (100)</td>
</tr>
<tr>
<td>19-3</td>
<td>323</td>
<td>239</td>
<td>30 (75.0)</td>
<td>10 (25.0)</td>
</tr>
<tr>
<td>21-2</td>
<td>439</td>
<td>355</td>
<td>40 (100)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>23-1</td>
<td>413</td>
<td>NF</td>
<td>40 (100)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>26-1</td>
<td>845</td>
<td>NF</td>
<td>40 (100)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>31-2</td>
<td>1669</td>
<td>383</td>
<td>9 (22.5)</td>
<td>31 (77.5)</td>
</tr>
</tbody>
</table>

Table 3 Numt loci in 20 unrelated horses.

1NF, not found in the 20 individuals.
only present in Twilight, suggesting that this insertion was a recent event. Figure 3 reports the results of PCR amplification of the DNA of three representative polymorphic loci in one heterozygous individual, one individual homozygous for the presence of the numt and one homozygous for the null allele. The molecular weight of the fragments is shown to the left of the gel.

In Fig. 4, the sequence of the numt and the null allele of three loci is compared with the corresponding horse mtDNA portion. In all three cases, the inserted mitochondrial sequence interrupts the ancestral empty sequence, but the addition of 1–6 random nucleotides occurred at the 3′ end of the insert. This sequence organization suggests that numts were probably inserted during the repair of DNA double-strand breaks via the non-homologous end-joining pathway.

Concluding remarks

In our survey of mitochondrial DNA insertions in the horse reference genome sequence, we identified 82 numts. Unexpectedly, at some loci, the null allele is prevalent in the analysed sample, suggesting that the number of numt loci in the global horse population is higher than 82 present in Twilight. The analysis of the insertion sites is consistent with the argument that numts arise from the repair of DNA double-strand breaks by non-homologous end-joining during evolution, similarly to interstitial telomeric sequences (Nergadze et al. 2004, 2007; Ruiz-Herrera et al. 2008).

It is worth noting that in the human genome (Giampieri et al. 2004; Ricchetti et al. 2004; Hazkani-Covo et al. 2010), the proportion of loci characterized by insertion polymorphism is much lower than that in the horse. This observation is in agreement with the notion that the horse genome is in a rapidly evolving state (Ryder et al. 1978; Wichman et al. 1991; Carbone et al. 2006; Trifonov et al. 2008; Wade et al. 2009; Piras et al. 2010).

Acknowledgements

This work was funded by Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale (PRIN 2008).
Conflicts of interest

The authors have declared no potential conflicts.

References


A conserved segmental duplication within ELA

C. L. Brinkmeyer-Langford*, W. J. Murphy†, C. P. Childers‡ and L. C. Skow§

*Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843-4458, USA. †Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843-4458, USA. ‡Department of Biology, Georgetown University, Washington, DC, USA. §Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843-4458, USA

Summary

The assembled genomic sequence of the horse major histocompatibility complex (MHC) (equine lymphocyte antigen, ELA) is very similar to the homologous human HLA, with the notable exception of a large segmental duplication at the boundary of ELA class I and class III that is absent in HLA. The segmental duplication consists of a ~710 kb region of at least 11 repeated blocks: 10 blocks each contain an MHC class I-like sequence and the helicase domain portion of a BAT1-like sequence, and the remaining unit contains the full-length BAT1 gene. Similar genomic features were found in other Perissodactyls, indicating an ancient origin, which is consistent with phylogenetic analyses. Reverse-transcriptase PCR (RT-PCR) of mRNA from peripheral white blood cells of healthy and chronically or acutely infected horses detected transcription from predicted open reading frames in several of the duplicated blocks. This duplication is not present in the sequenced MHCs of most other mammals, although a similar feature at the same relative position is present in the feline MHC (FLA). Striking sequence conservation throughout Perissodactyl evolution is consistent with a functional role for at least some of the genes included within this segmental duplication.

Keywords equine lymphocyte antigen, horse, major histocompatibility complex, Segmental duplication.

Introduction

Many proteins of the innate and adaptive immune systems are encoded within the major histocompatibility complex (MHC) (Trowsdale 1995), and genetic variation among these genes is associated with predispositions to autoimmune diseases and susceptibility to various pathogens and parasites (Escayg et al. 1997; Thorsby 1997; Shurif et al. 1998, 1999; Bailey et al. 2000; Kaufman 2000; Park et al. 2004; Dukkipati et al. 2006).

The gene content and organization of the MHC have been described for human (MHC Sequencing Consortium 1999), mouse (Mouse Genome Sequencing Consortium 2002), dog (Wagner 2003), cat (Yuhki et al. 2008), pig (Renard et al. 2006), cow (Brinkmeyer-Langford et al. 2009), horse (Gustafson et al. 2003; Tallmadge et al. 2005) and opossum (Gouin et al. 2006).

Overall, the structure of the MHC appears to be remarkably conserved among mammalian species. The genetic map of the horse equine lymphocyte antigen (ELA) shares the general structure of MHCs of other mammalian species (for review, see Bailey et al. 2000), but assembly of the horse genome sequence (EquCab2.0) predicted an unexpected segmental duplication of approximately 710 kb at the boundary of the class III and class I regions (Wade et al. 2009).

Here, we describe a detailed analysis of the ELA segmental duplication, which revealed at least 11 duplicated units within the region, including ten that contain both class I-like and BAT1-like helicase domain sequences, flanked by a single unit with the full-length equine BAT1 gene. Evidence for each of the duplicated blocks was found in the genomes of other perissodactyls, and reverse-transcriptase PCR (RT-PCR) studies of mRNA from horse WBCs revealed several transcribed open reading frames (ORFs) in the duplication. The persistence of the segmental duplication in perissodactyl species separated by more than 55 million years of evolution suggests a functional role for genes contained within this region.

Accepted for publication 21 September 2010
Materials and methods

Sequence analysis and identification of duplicated BAT1-like regions

Equine homologues of HLA genes were identified by performing a local BLAST alignment (Altschul et al. 1990) for 229 confirmed or predicted gene loci from HLA (positions 29.57–33.50 Mb on HSAtt; NCBI build 36.2) against the EquCab2.0 assembly of ECA20. Dotplot analysis was used to visualize architecture of the segmental duplication after identification of the expansion between the ELA class I and class III regions. The dotplot figure was generated by aligning the horse sequence from 30.60 to 31.35 Mb on ECA20 to itself using NCBI’s blast program, bl2seq option (http://blast.ncbi.nlm.nih.gov/) (Fig. 1).

Primer design

Conservation of BAT1-like sequences among the duplicated blocks in ELA precluded the design of copy-specific PCR primers; consequently, two non-overlapping generic primer pairs were designed (Bat1 and Bat1-10) to amplify target DNAs from domestic horse (positive control) and other perissodactyl species. Both were designed using PRIMER3 software (http://frodo.wi.mit.edu/) from highly conserved sequence to amplify each of the identified BAT1 copies. In combination, the primer pairs amplified ten of the eleven copies identified on ECA20. A third primer pair, designated Bat1.1, was designed to specifically amplify the remaining BAT1 copy. Sequences of the three primer pairs are

- Bat1 forward 5'-AATTGGCTGGGAAGTTCTCCTGCT-3';
- Bat1 reverse 5'-TCTCCTATAAGAGGCTGTAG-3';
- Bat1-10 forward 5'-AGGTTGATGATGTTCAACACTGAA-3';
- Bat1-10 reverse 5'-CTGCTTCTGTTCTGTTGAGCTA-3';
- Bat1.1 forward 5'-AGGGGGATGTATGTTCATGG-3';
- Bat1.1 reverse 5'-AGGGGGATGTATGTTCATGG-3'.

Bat1 PCR products were 198 bp in size, while those for Bat1-10 and Bat1.1 primers were 187 and 119 bp, respectively. Bat1 and Bat1-10 products contained a ~51 bp overlap, while Bat1.1 amplicons were from a different region.

Amplification and purification of duplicated sequences in perissodactyl species

PCR was performed in a GeneAmp 9700 thermal cycler (Applied Biosystems) in 25-μl reaction volumes containing 50 ng DNA, 10× buffer, 10 μM of each primer, 20 mM dNTPs, 50 mM MgCl₂ and 0.25 U Platinum™ Taq polymerase (Invitrogen). Amplification parameters included an initial 5-min denaturation at 95 °C; two cycles of 95 °C for 30 s, 60 °C for 30 s and 72 °C for 30 s; followed by 33 cycles of 30 s at 95 °C, 30 s at the optimized annealing temperature (58 °C for Bat1-10 and Bat1.1 primers; 60 °C for Bat1 primers) and 30 s at 72 °C; ending with a final extension for 5 min at 72 °C. To verify the presence of the duplication and number of duplicated blocks, we screened DNA samples from two domestic horses (Equus caballus): Twilight, a Thoroughbred (TB) mare used for the Horse GenomeProject, and Bravo, a TB gelding and donor for the CHORI-241 BAC library. DNAs from non-horse perissodactyls were also analysed, including Przewalski horse (Equus caballus przewalskii), donkey and Poitou (Equus asinus), Damara zebra (Equus burchelli antiquorum), Grants zebra (Equus quagga boehmi), Grey's zebra (Equus grevyi), Hartmann mountain zebra (Equus zebra hartmannae), kiang (Equus kiang), onager (Equus hemionus), Baird's tapir (Tapirus bairdii) and white rhinoceroses (Ceratotherium simum). Negative controls included DNA of human, cow and hamster. Amplification products were resolved by electrophoresis in 2.0% agarose gels containing 0.25 μg/ml ethidium bromide. Dog and cat DNAs were also tested, but no amplification was obtained with any of the primers.

Cloning and sequencing of amplicons

PCR products were purified using the QiAquick PCR Purification Kit (Qiagen). Multiple bands were observed for rhino and tapir, so PCR products were run on 1.5% low-melting-point agarose gels containing 0.25 μg/ml ethidium bromide, and bands of the expected size were excised and purified from the gel slices with Amicon Ultra centrifugal filter devices (Millipore). Quality and concentration of all PCR products were determined using a NanoDrop™ ND-1000 Spectrophotometer (Thermo Scientific). Products from Bat1.1 primers were sequenced directly for each species.

Bat1 and Bat1-10 amplicons from donkey, Przewalski horse, Hartmann mountain zebra, tapir and rhinoceroses were cloned into the pCR 4-TOPO vector (Invitrogen) and transformed into One Shot™ chemically competent Escherichia coli cells per the manufacturer’s instructions. Ninety-six clones were picked and sequenced for each amplification reaction in each species, except the Bat1 amplification from...
rhino, from which only forty clones were obtained for sequencing. DNA was prepared and sequenced from cell pellets by SeqWright.

Phylogenetic analyses of duplicated sequences

Repetitive sequences were masked (http://www.repeatmasker.org/), trimmed, manually edited and assembled into contigs using Sequencher software (GeneCodes). Bat1 and Bat1-10 amplicon sequences were analysed separately for the construction of species-specific contigs in Sequencher, using the parameters of 100 bp minimum overlap and 99% minimum similarity between the sequences. Consensus sequences from each contig were determined by plurality, and putative orthologous sequences, as well as novel/taxon-specific copies, were organized into contigs to confirm the presence of duplicated blocks and provide an estimate of the number of different copies amplified from each of the non-caballus species. Results were confirmed via BLAST analysis of all amplicon sequences, with top alignments (>95% identity) to horse genomic sequence used to infer orthology.

Construction of phylogenetic trees

Phylogenetic analyses were conducted in RAxML Version 7.0.4 (Stamatakis et al. 2008) separately on sequences amplified with Bat1 or Bat1-10 primers and rooted with the human BAT1 sequence. Multiple alignments for tree building were performed using the edited sequences from Sequencher analysis. Trees were initially constructed using only horse sequences (Fig. 2) and then separately for each species (Fig. S1), using the maximum likelihood estimation (ML) method of phylogenetic reconstruction. Simplified ML trees were generated for each species other than horse using the Sequencher consensus sequences; horse sequences and the human BAT1 root were included for reference. Lastly, a ML tree was built on full-length sequences of each horse species.

**Figure 2** Maximum likelihood estimation phylogenetic trees for horse BAT1 copies, with bootstrap values indicated. Trees A and B are rooted with the corresponding human sequence, while tree C is rooted with the human BAT1 genomic sequence. (a and b) show relationships between the Bat1 and Bat1-10 primer amplicons of each of the BAT1-like copies in the horse, respectively. (c) depicts the relationships between the full-length horse copy sequences, trimmed for uniform size, including chromosomally unassigned sequences. Bars at the bottom of each tree illustrate nucleotide substitutions per site.
BAT1-like sequence (based on EquCab2) and the chromosomally unassigned sequences WGA351, WGA477 and WGA898 (Table 1). Node support was established using the rapid bootstrapping algorithm with replicate number determined by RAxML. Phylogenetic trees were similarly generated for horse class I loci using protein sequences, trimmed for uniform length, with no outgroup sequence (Fig. 3).

Table 1  Sequences identified as BAT1-like based on alignments with genomic human (ENSG00000198563) and horse (ENSECAT00000007750) BAT1 genes.

<table>
<thead>
<tr>
<th>BAT1 copy</th>
<th>Chromosome</th>
<th>Position (bp)</th>
<th>Average identity (%) to human</th>
<th>Average identity (%) to horse</th>
<th>Exons identified in copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAT1.1</td>
<td>ECA20</td>
<td>30625450–30624815</td>
<td>80.74</td>
<td>84.07</td>
<td>9</td>
</tr>
<tr>
<td>BAT1.2</td>
<td>ECA20</td>
<td>30761350–30758939</td>
<td>82.70</td>
<td>85.58</td>
<td>6 7 8 9 10</td>
</tr>
<tr>
<td>BAT1.3</td>
<td>ECA20</td>
<td>30795763–30793621</td>
<td>81.00</td>
<td>83.41</td>
<td>6 7 8 9</td>
</tr>
<tr>
<td>BAT1.4</td>
<td>ECA20</td>
<td>30864193–30861796</td>
<td>83.09</td>
<td>84.44</td>
<td>6 7 8 9</td>
</tr>
<tr>
<td>BAT1.5</td>
<td>ECA20</td>
<td>30924827–30922427</td>
<td>82.79</td>
<td>84.58</td>
<td>6 7 8 9</td>
</tr>
<tr>
<td>BAT1.6</td>
<td>ECA20</td>
<td>30974620–30972559</td>
<td>82.55</td>
<td>84.62</td>
<td>6 7 8 10</td>
</tr>
<tr>
<td>BAT1.7</td>
<td>ECA20</td>
<td>31027523–31025109</td>
<td>82.96</td>
<td>84.59</td>
<td>6 7 8 9</td>
</tr>
<tr>
<td>BAT1.8</td>
<td>ECA20</td>
<td>31085154–31083048</td>
<td>81.66</td>
<td>83.99</td>
<td>6 7 8 9 10</td>
</tr>
<tr>
<td>BAT1.9</td>
<td>ECA20</td>
<td>31125503–31124352</td>
<td>79.96</td>
<td>87.47</td>
<td>6 7 8 10</td>
</tr>
<tr>
<td>BAT1.10</td>
<td>ECA20</td>
<td>31232990–31230674</td>
<td>82.59</td>
<td>86.77</td>
<td>6 7 8 9 10</td>
</tr>
<tr>
<td>BAT1.0</td>
<td>ECA20</td>
<td>31330528–31321132</td>
<td>85.85</td>
<td>100.00</td>
<td>1 2 3 4 5 6 7 8 9 10 16</td>
</tr>
</tbody>
</table>

Average identities were obtained by averaging the “%ID” values for all alignments found within a particular region by BLAST. Alignments to the horse BAT1 mRNA sequence (ENSECAT00000008265) facilitated identification of sequences similar to each of the 10 BAT1 exons within each BAT1-like copy. Numbers in black blocks identify exons that aligned with >90% identity to the reference horse mRNA sequence. Numbers in gray blocks identify exons that aligned with >82 but <90% identity to the reference horse mRNA sequence.

Figure 3 Maximum likelihood estimation phylogenetic tree for protein sequences of class I loci within the segmental duplication region, with bootstrap values indicated. Branch labels represent shortened transcript names: e.g. ENSECAT00000002326 has been shortened to 002326. Refer to Tables S1 & S2 for the complete list of class I transcripts in the duplicated region. The bar at the bottom of the tree illustrates nucleotide substitutions per site.
Expression studies

Primer design

RT-PCR primers were designed using NCBI's PrimerBlast for each of 22 predicted mRNA sequences (NCBI and Ensembl databases) in the 30.17–31.33 Mb sequence region of ECA20. When possible, primers were designed across exon boundaries to eliminate amplification of contaminating DNA. Six primer pairs were from BAT1-like ORFs, one from the known full-length BAT1 gene and fifteen from MHC class I-like sequences. Details regarding primer sequences and amplicon sizes may be found in Table S2.

RNA extraction

Approximately 10 ml of whole blood was collected in EDTA from each of seven healthy horses and five with acute or chronic infections. Healthy horses included four TB stallions, four mares and one colt. Horses with acute or chronic infections included a Quarterhorse (QH) filly with an infected knee, a Warmblood (WB) gelding with an infected foot and a 14-day-old male QH foal with contracted foal syndrome and possible sepsis. Horses with chronic infections included a QH gelding with inguinal abscesses and a QH mare with coffin infection. WBCs were collected from each of the above animals and RNA was extracted using the LeukoLOCK system (Invitrogen) according to manufacturer’s instructions. RNA purity and concentration were determined using a NanoDrop™ ND-1000 Spectrophotometer.

Reverse-transcriptase PCR

RT-PCR was performed in 15-μl reaction volumes containing ~100 ng total RNA and 1 μmol of each primer, with nuclease-free water, 2× buffer and SuperScript™ III RT/Platinum® Taq DNA polymerase from the SuperScript™ III One-Step RT-PCR System (Invitrogen). All reactions were performed in a GeneAmp 9700 thermal cycler (Applied Biosystems) beginning with a 30 s incubation at 55 °C and a 2-min denaturation at 94 °C, followed by 30 cycles of 15 s at 94 °C (denaturing), 30 s at 58 °C and 1 min at 68 °C (extension), with a final extension for 5 min at 68 °C. Reaction products were resolved on 2% agarose gels containing 0.25 μg/ml ethidium bromide. Dideoxy-sequencing was performed by SeqWright to confirm gene identity. In cases where multiple products were produced, gel purification was used to purify a band of the proper size for sequencing. Each sequence was aligned to the horse genomic assembly using BLAST to verify alignment at the expected location on ECA20.

Results

Sequence analysis and identification of duplicated BAT1-like regions

Identification of ELA sequences with homology to HLA in the equine genomic database (version 2.0) revealed a 710-kb segmental duplication in ELA at sequence positions 30.62–31.33 Mb on ECA20. Dotplot analysis of the predicted segmental duplication confirmed a series of ten duplicated blocks, each containing copies of BAT1-like sequence fragments designated BAT1.1–1.10, plus an eleventh block with the single full-length BAT1 gene designated BAT1.0 (Fig. 1). Each BAT1-like sequence showed ≥79% sequence identity to the corresponding region of human BAT1 genomic sequence (Table 1). Exons present within these copies encode a DNA/RNA helicase C-terminal domain (Ensembl transcript ID ENSECAT00000008265); the full-length BAT1 gene product additionally encodes a DEAD box helicase and functions in spliceosome assembly and mRNA export (http://www.uniprot.org/uniprot/Q13838; Fleckner et al. 1997; Shi et al. 2004; Zhao et al. 2004). Interspersed among the BAT1 sequences in the segmental duplication were seven ELA class I sequences with an average of 81% identity to the human class I gene HLA-A. Additionally, three chromosomally unassigned contigs were identified, each containing a BAT1-like sequence and an ELA class I sequence. Positions, sizes and exon content of the BAT1-like sequences are presented in Table 1. Predicted genes, their functions and their coordinates within the segmental duplication region on ECA20 are described in Table S1.

Phylogenetic analyses of BAT1 copy sequences

Identification of orthologous sequences in other perissodactyls

Initial BLAST analysis of ampliconic sequences generated by the Bat1 and Bat1-10 primer pairs confirmed the presence of full-length amplicon sequences with ≥95% identity to 12 of the 14 horse BAT1 copies in at least one other perissodactyl species. Orthologs for all copies of the equine sequences except BAT1.5 and BAT1.7 were identified in at least one other perissodactyl species, as defined by BLAST alignments for each clone (Table 2). The divergent BAT1.1 orthologues were detected only in zebra and donkey.

Unique primers could not be designed for BAT1-like sequences in the three chromosomally unassigned contigs (WGA351, WGA477, and WGA898, Table 1), but two lines of evidence suggest that the three contigs represent additional copies not incorporated into the EquCab2.0 assembly. First, alignment of the BAT1-like sequences in each unassigned contig to the assembled ECA20 sequence returned no full-length matches, and the sequences that did align were <96% identical to any BAT1 sequence on ECA20. Secondly, BLAST analysis of the sequenced amplicons from the all of the other perissodactyl species against the three chromosomally unassigned horse contigs returned full-length alignments with >90% identity to at least one of the horse unassigned contigs. Except for rhino, the cloned sequences from all non-horse species aligned sufficiently well to BAT1-like sequences in the three unassigned contigs of horse to meet...
and results described earlier (and in alignment. 2010 Stichting International Foundation for Animal Genetics, (Fig. 2a,b).‡ and were not as divergent as 191 ‡ 18 4 9 35 38 4 12 21 3 1 2 1 72 6 0 BAT1.3 and were detected in all species, ‡ sequences of all horse copies trimmed to uniform size (based and mountain zebra, we detected BAT1-like sequences orthologous to six of the 11 blocks present on ECA20, plus all three chromosomally unassigned sequences, although the complement of these sequences varied between the two species. In addition, sequences from some duplicated blocks were detected among more species of perissodactyls than sequences from other blocks. For example, sequences that aligned ≥95% to BAT1.2 were detected in all species, while orthologues to BAT1.5, BAT1.6 and BAT1.7 were detected infrequently or not at all. Finally, the amplification product from rhinoceros contained a single orthologous sequence that aligned with ≥95% identity to horse BAT1.2.

Phylogenetic trees Horse

Maximum likelihood estimation trees constructed for horse sequences amplified by Bat1 and Bat1-10 primers grouped most of the BAT1-like copies into one (Bat1-10) or two (Bat1) sets of related sequence clusters, while three other copies (BAT1.9, BAT1.10 and BAT1.0) were more divergent. These trees exhibited similar topologies, differing only in their placements of BAT1.10 and BAT1.0 (Fig. 2a,b). BAT1.10 and BAT1.0 were not as divergent as BAT1.9 but still were clearly more divergent than the BAT1.2-BAT1.8 sequences.

An additional ML tree (Fig. 2c) generated from full-length sequences of all horse copies trimmed to uniform size (based on the size of the smallest duplicated BAT1 segment, BAT1.1) largely agrees with the trees constructed using only horse Bat1 and Bat1-10 amplicons (Fig. 2a,b). Not surprisingly, BAT1.1 proved to be the most distantly related segment: it was the shortest BAT1-like segment and was not amplifiable by the Bat1 and Bat1-10 primers, consistent with increased sequence divergence. As observed in the other trees, the full-length sequences of BAT1.9, BAT1.10 and BAT1.0 formed a distinct cluster, suggesting that these sequences may have been derived from a separate ancestral BAT1-like sequence.

Unassembled contigs

The chromosomally unassigned BAT1-like sequences from WGA351, WGA477 and WGA898 trimmed and analysed with the full-length sequences of the BAT1-like copies on ECA20 illustrate the phylogenetic relationships among the sequences. The BAT1-like sequences in WGA351, WGA477 and WGA898 appear to have descended from a common ancestral sequence, together with BAT1.3 and BAT1.6. These results and the BLAST results described earlier (and in Table S1) suggest at least two more horse BAT1-like copies (WGA351/WGA477 and WGA898) exist in the horse genome. The BAT1-like sequences within contigs WGA351/WGA477 share greater similarity to sequences isolated from non-horse species than to any of the eleven chromosomally assigned horse sequences (Fig. S1K).

Other species

Maximum likelihood estimation trees generated using sequences from the Bat1 and Bat1-10 amplicons of different perissodactyl species (Fig. S1) showed that the same BAT1-like copies that clustered in the horse also clustered in the other species. Consensus sequences from species-specific contigs built with Sequencher were used, together with the horse sequences and the human root, to provide an abridged display of phylogeny. The trees illustrate close relationships between the Bat1 and Bat1-10 amplicon sequences in horse and non-horse species, confirming observations from the horse-only phylogenetic trees and demonstrating that orthologous sequences retain their sequence similarities across Perissodactyls. It is important to note that bootstrap values are considerably lower for the upper-level branches than the endpoints, reflecting the

### Table 2 The distribution of orthologs within each of the six perissodactyl species, based on BLAST alignment.

<table>
<thead>
<tr>
<th>Species</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Real (0)</th>
<th>WGA351</th>
<th>WGA477</th>
<th>WGA898</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donkey</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>17</td>
<td>16</td>
<td>21</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>72</td>
<td>26</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Przewalski horse</td>
<td>2</td>
<td>13</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>8</td>
<td>21</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>72</td>
<td>34</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Rhino</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>35</td>
<td>38</td>
<td>4</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tapir</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>35</td>
<td>38</td>
<td>4</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zebra</td>
<td>7</td>
<td>2</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Numbers in black blocks identified as homologous to duplicons (number indicates number of clones). Numbers in gray blocks identified as homologous to duplcon when repeat-masked.

Black blocks identify orthologous loci based on ≥95% sequence identity to the corresponding horse sequence. BAT1-like loci-0, -3, -6, -8, -9 and -10 (grey) in some species contain an AC:TG microsatellite that, when masked, gave contiguous, full-length alignments ≥95% identical to the horse sequences. Numbers within each block indicate the number of subclones recovered for each duplicon for the indicated species. No subclones were obtained for BAT1.1 (see Materials and methods).
substantial similarity between sequences as well as the relatively short lengths of sequences used in the analysis.

Highly divergent BAT1-like copies were also apparent in other species when compared to sequences with clear homology to the assembled horse sequences. These sequences included orthologues to chromosomes unassigned horse sequences, sequences too divergent to align using the Sequencer parameters 100 bp minimum overlap and 99% minimum sequence identity and taxa-specific sequences with undefined orthology (see Discussion). Trees L and M (Fig. S1) show consensus sequences from all species for Bat1 and Bam1-10 and illustrate the relationships among consensus sequences across perissodactyl species. In general, the trees constructed for the non-horse perissodactyl species reflect the orthologous relationships predicted by BLAST analysis, although some relationships were not equally supported with both methods. For example, the ML tree for the Bat1 amplicon sequences indicates an orthologue to BAT1.8 in Przewalski horse, a relationship not indicated from the BLAST analysis. Similar exceptions were also seen for sequences from zebra and tapir. It is possible that these exceptions arose as a consequence of constructing consensus sequences for phylogenetic analysis, thereby losing sequence information that could normally indicate similarity to a different horse sequence using BLAST.

Expression studies

RT-PCR primer pairs designed for each of the 22 ORFs in the interval 30.17–31.33 Mb of ECA20 identified eight transcribed sequences in peripheral white blood cells. RT-PCR confirmed the expression of the full-length BAT1 transcript (BAT1.0), as expected, but failed to amplify any transcripts that could be assigned to the various BAT1-like loci. Seven transcripts were identified with sequence similarity to ELA class I genes. Samples from all twelve horses contained transcripts of Ensemble IDs ENSECAT00000016778 and ENSECAT00000017262 (outside the segmental duplication), and ENSECAT00000026764 and LOC100058098 (within the duplication). Some but not all animals expressed ENSECAT00000012758 (outside the segmental duplication), and MHCB3 and EQMHC1 (within the duplication). Differences in expression among animals were not correlated with age, gender, breed or health status. Among the differentially expressed class I sequences, ENSECAT00000012758 was found in all but two horses, MHCB3 transcription was found in eight horses and EQMHC1 was expressed in all infected horses and two healthy horses. Interestingly, for the two different ORFs within the locus LOC100057920/ENSECA2000000247 is one (ENSECAT00000026764) was expressed, but the other (designed from LOC100057920) was not, suggesting that alternative splicing may be present in this locus. Only the product amplified with LOC100053668 primers could not be validated by sequencing.

Discussion

Sequence analysis and identification of duplicated BAT1-like regions

A segmental duplication at the boundary between ELA class I and III regions consists of at least ten duplicated blocks, each containing a partial BAT1 gene and one or more full- or partial-length ELA class I genes. An eleventh block contains the full-length BAT1 gene. This genomic structure distinguishes the horse MHC from that of most other mammals and appears to be a feature of all perissodactyl species. A segmental duplication is also present at a similar location in the cat genome (Yuki et al. 2007, 2008), but comparable structures have not been identified in the genomes of other carnivores, e.g. the dog (CanFam2.0), giant panda (Ailuropoda melanoleuca, AllMel1.0) or bat species (Myotis lucifugus, myoLuc1, and Pteropus vampyrus, petVam1).

Maximum likelihood estimation trees based on the BAT1-like sequences (Fig. S1) provided little insight into the evolutionary history of this region beyond confirming the relationships among orthologues predicted by the BLAST alignments. The sequences grouped into similar clusters in all comparisons, but the low bootstrap values for branch-points between clusters indicate little support for any higher-level topology. This lack of resolution reflects, at least in part, the high degree of similarity among the duplicated sequences and the relatively short (>200 bp) lengths of the sequences being compared. At least two additional BAT1-like copies are suggested by the presence, in other perissodactyl species, of homologues to BAT1-like sequences in unassembled horse contigs WGA351, WGA477 and WGA898 (panel K). Taken together, these results suggest that the segmental duplication likely originated from one or more ancient duplication events prior to and early in Perissodactyl evolution, with subsequent sequence divergence and/or deletions producing different sets of orthologues in each species.

Other loci near or within the segmental duplication

In addition to BAT1-like sequences, the region from 30.17 to 31.33 Mb on ECA20 contains 18 ELA class I-like loci (including one sequence, LOC100054254, most similar to the MIC loci in humans), two novel loci with transposase-like domains (ENSECAG00000007223 and ENSECAG00000007227), one sequence homologous to the human gene mitochondrial coiled-coil domain 1 (MCCD1), one homologue to the human locus C13orf37 and two novel predicted loci with no similarities to any currently annotated genes. With the exception of the ELA class I and MCCD1 loci, none of these sequences share conserved synteny with HLA or Feline lymphocyte antigen (FLA).

In general, the duplicated blocks between 30.62 and 31.33 Mb show a head to tail arrangement of pairs of [ELA
class I]-[BAT1-like] loci, indicating that micro-inversions have played no role in the architecture of the extant structure. The BAT1-like sequences (BAT1.1 through BAT1.10) are all gene fragments encoding the helicase domain portion of a full-length BAT1 gene, but the seven ELA class I-like loci are full-length or nearly so. Class I-like loci present in the segmental duplication represent a mix of classical and non-classical genes and gene fragments based on the presence of the transmembrane and/or C-terminus domains (Holmes & Ellis 1999). The block containing BAT1.10 is exceptional, as the BAT1-like sequence is flanked by homologues to MCD1 and C3orf37, rather than by class I sequences. Details regarding all loci in the segmental duplication may be found in Tables S1 & S2.

A similar segmental duplication in the domestic cat

The only other mammalian genome known to contain a segmental duplication in this region of the MHC is the domestic cat (Felis catus). The cat FLA contains nine copies of a BAT1-like gene and a single full-length BAT1 gene (Yuhki et al. 2007, 2008) and twenty copies of class I-like sequences. Perissodactyla and Carnivora are in the same super-ordinal lineage Laurasiatheria, having diverged from each other approximately 85 million years ago (Tree of Life, Springer & Murphy 2007). It is tempting to conclude that the segmental duplication in these two taxa was present in a common ancestor. Like the segmental duplication in the horse, the genes interspersed between the cat BAT1 copies include MCD1 and both classical and non-classical class I loci, but other sequences in the cat, including three copies of ribosomal protein S28, are not found within the horse segmental duplication. Such divergence may reflect the accumulated changes since the origin of such an ancient genomic feature.

Conversely, the absence of any similar structure in the MHCs of dog, bat or panda is not consistent with the common ancestral duplication hypothesis, and the relative paucity of functional genes in the region is not consistent with convergent evolution. Alternatively, the horse and cat segmental duplications may have arisen independently as a result of genomic instability in this region (Bailey et al. 2004). A more complete understanding of the origin of these structural features in the MHCs of cat and horse will await analyses of the genome assemblies of additional species of felids and perissodactyls.

Phylogenetic analyses of duplicated sequences

The conservation of Bat1/Bat1-10 primer sites among perissodactyl species and the strong identity of horse BAT1-like sequences to each other and among orthologues of other species speak to the striking evolutionary conservation of BAT1-like sequences of the segmental duplication. Orthologues to the horse sequences were evident in donkey, Przewalski horse, Hartmann zebra, tapir and rhinoceroses; chromosomally unassigned sequences WGA351, WGA477 and WGA898 provide evidence for at least two additional BAT1-like copies in each Perissodactyl species. Altogether, orthologues, defined as sequences with >95% identity to one of the horse BAT1-like copies, were identified in at least one other species for all but BAT1.5 and BAT1.7. The BAT1.2 sequence was identified in each of the Perissodactyl species, and six other BAT1-like sequences from different duplicated blocks were detected in three or more species. Such conservation across taxa is usually associated with functionality, although none of these sequences were detected as transcripts in peripheral WBCs.

The phylogenetic comparison of homologous sequences and the relative positions of those sequences in the horse segmental duplication provide no evidence that the duplicated blocks originated over time in a stepwise manner. Although subsequent shuffling by unequal recombination could have obscured such evidence, the conserved nature of this region of the horse genome and the lack of any detected structural polymorphisms among horses indicate that recent expansions/contractions are infrequent. We think a more likely scenario is that the extant segmental duplication arose from one or several ancient expansions, at least one of which predated perissodactyl evolution. A more compelling question is why these sequences have remained intact and conserved among perissodactyls over 55 million years of evolution. Duplication has played a central role in MHC evolution (Yeager & Hughes 1999; Flajnik & Kasahara 2001) and regions that are inserted or deleted in one species relative to another have been implicated in species-specific susceptibilities to disease (e.g. Anzai et al. 2003). The genes included in the duplicated region may have at some point met a need that was unique to perissodactyls, perhaps some specific immunological challenge because of a changing environment. Indeed, many more perissodactyl species were in existence 20–55 million years ago than survive today; the subsequent extinction of many of these species may have been due in part to their failure to adapt to habitat changes (Radinsky 1969).

Expression studies

Although no evidence for expression of any of the BAT1-like sequences was obtained in this study, functionality of at least some genes in the segmental duplication was indicated by the presence of transcripts in peripheral WBCs for four of the 12 class I sequences. Of the seven transcripts from class I-like loci within or immediately proximal to the segmental duplication, only two (MHCB3 and EQMHC1) have been previously described (Ellis et al. 1995; Tallmadge et al. 2010). Furthermore, evidence for differential expression of class I genes was observed. Four class I transcripts were observed in all horses and three were expressed in some but
not all animals. Whether this variation in expression is owing to differences in gene regulation or to copy number polymorphisms will require more comprehensive studies with additional tissues/cell types and larger sample sizes than reported here.

Conclusions

The persistence of a conserved large segmental duplication in the MHCs of Perissodactyla species separated 20–55 mya merits further investigations. Sequencing the genomes of the non-horse species will provide insights into the architectural features of this genomic feature that are needed to assess its stability over evolutionary time; the sequences provided in this study will help to inform assembly of these new genomes. It will be particularly interesting to know whether such a genomic feature is also found occasionally among the MHCs of other species of the Laurasiatheria clade.

The identification of hitherto unreported ELA class I transcripts in peripheral WBCs is surprising, as these cells are often used for identification of class I expression. It remains to be demonstrated whether the transcripts we identified by RT-PCR are present at biologically meaningful levels, are translated, or whether the polypeptides encoded by these loci are functional. Even so, it seems apparent that the horse genome has added another element to be considered in our efforts to understand the evolution and biology of the mammalian MHC.

Acknowledgements

We acknowledge Jan Janečka for assisting with the phylogenetic analyses and for proofreading the manuscript. Terje Raudsepp provided perissodactyl DNAs and Jana Caldwell provided blood samples for RT-PCR.

Conflicts of interest

The authors have declared no potential conflicts.

References


major histocompatibility complex DRB3 (BoLA-DRB3) alleles with occurrence of disease and milk somatic cell score in Canadian dairy cattle. Animal Genetics 29, 185–93.


Supporting information

Additional supporting information may be found in the online version of this article.

Figure S1 Maximum likelihood estimation phylogenetic trees showing relationships between the Bat1 and Bat1-10 amplicons for non-horse perissodactyl species.

Table S1 Loci, their predicted functions and coordinates within the segmental duplication region on ECA20, delineated by the locations of the end of horse–human homology.

Table S2 Characterization of loci within the segmental duplication region on ECA20.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
De novo mutation of KIT discovered as a result of a non-hereditary white coat colour pattern

H. Holl*, S. Brooks† and E. Bailey‡

*Department of Animal Science, Cornell University, Ithaca, NY 14853, USA. †Department of Animal Science, Cornell University, Ithaca, NY 14853, USA. ‡Department of Veterinary Science, University of Kentucky, Lexington, KY 40546, USA

Summary

A new dominant white allele was suspected when two Thoroughbred horses with minimal white marking on the coat produced a colt with a large amount of coat depigmentation. Because of its association with similar patterns in other horses, the KIT gene was selected as a candidate gene, and all 21 exons were sequenced in the colt. A novel 5-bp deletion was discovered in exon 3 and was confirmed with allele-specific PCR. The mutation introduced a pre-mature stop codon, resulting in truncation of the protein. The deletion was not present in either parent and is suspected to be responsible for the extensive white coat colour in the colt. Additionally, a previously described missense mutation was detected in exon 14 of both the colt and sire but is not believed to be causative. Parentage testing was conducted as required by The Jockey Club for Thoroughbred registration, and the foal qualified for the stated parentage. This novel deletion in exon 3 is the 12th discovered dominant white allele in the horse.

Keywords coat colour, equine, gastric ulcer, hepatitis, horse, KIT.

The KIT gene in mammals encodes a tyrosine kinase receptor whose signalling is involved in the development of erythrocytes, melanocytes, germ cells, mast cells and interstitial cells of Cajal (ICC) (Kitamura & Hirota 2004). Homozygous loss of function mutations in mice have been shown to cause areas of depigmentation, impairment of hemopoiesis, sterility, gastrointestinal tumours and early lethality (Bult et al. 2008). Gain of function mutations identified in humans and mice have been shown to cause tumours in mast cells, germ cells and ICCs (Kitamura & Hirota 2004).

Mutations in KIT leading to unpigmented skin and hair have been identified in humans, mice, pigs and horses (Haase et al. 2009a). Many mutations in KIT can produce the dominant white phenotype that was first reported by Pulos & Hutt (1969). So far, 14 polymorphisms involving the KIT gene have been reported for horses, with phenotypes ranging from small areas of depigmentation to white over the entire body (Haase et al. 2009a). Two of these mutations, Sabino-1 and Tobiano, produce viable homozygotes and are not classified as dominant white (W) alleles (Brooks & Bailey 2005; Brooks et al. 2007). There are no defects other than depigmentation documented for horses with a single copy of a W allele, although an embryonic lethal condition is expected for homozygotes and compound heterozygotes with some, if not all, of these alleles. KIT expression also affects development of mast cells, and a study of haematological parameters in the Franches Montagnes breed found no statistically significant effects of the W1 KIT allele compared with wild-type individuals (Haase et al. 2009b).

A novel mutation in the KIT gene was suspected when the mating of two Thoroughbreds produced a colt with a large amount of white patterning. Approximately half of the proband’s body was depigmented, with four high white stockings that extended over his midline, irregular markings on his body and a white blaze that covered his muzzle (Fig. 1). Both parents had small white markings; a white stripe on the face and three white pasterns in the case of the sire, and only one white fetlock in the case of the dam. As of March 2010, the sire had produced over 400 registered foals with no recorded dominant white offspring. The dam had produced three solid colour foals prior to this colt.

Hair samples were obtained from the 1-month-old Thoroughbred colt and his dam for genetic analysis of the KIT gene as a candidate for his unusual colour. DNA was
obtained from these two samples using a hair lysis protocol previously described by Brooks et al. (2007). DNA from the sire was provided from a previously banked sample. The proband was born healthy and appeared to develop normally up to 5 weeks of age, when the samples were obtained.

Eighteen primer pairs were created with Primer 3 based on the gapped alignment of KIT mRNA (NM_001163866) to equCab2 genomic sequence and were used to amplify the 21 KIT exons by PCR (Rozen & Skaletsky 2000). Detailed primer information is given in Table S1. Sequencing direction was selected based on the distance of the exon to the primers. PCRs were prepared by scaling the manufacturer’s protocol to 20 µl reactions with 2 µl template, with the exception of using only 1.0 U of FastStart Taq polymerase (Roche Diagnostics Corp.). Reactions were carried out on an eppendorf Mastercycler gradient using the recommended procedure from Roche. The PCR products were sent to Cornell University’s Life Sciences Core Laboratories Center for Sanger sequencing in one direction. The trace files were analysed and aligned to sequencing data from an unrelated solid colour horse as well as the equCab2 genome assembly with phredPhrap (Green Group, University of Washington). The electropherograms were visually inspected for polymorphisms using consed (Green Group). The complementary strand of the exon 3 amplicon was sequenced and added to the exon 3 data to confirm the presence of a polymorphism (accession number HQ256561).

Two allele-specific primers were designed to detect a 5-bp deletion in the 408-bp region containing exon 3 (Table S1). The PCR was scaled down to 10 µl and included 1 µl of each 5 µm primer. The wild-type allele produced a 273-bp fragment, whereas the mutant allele produced a 147-bp fragment. The region was amplified for each sample and was visualized on a 3% agarose gel containing 1× SYBR Safe (Invitrogen).

KIT exon screening by sequencing detected a deletion in exon 3 (c.559_563delTCTGC), which would result in early termination of translation (p.Ser187ArgfsX10). The allele-specific primers produced the deletion fragment in the colt and only the wild-type fragment in the dam, sire and an unrelated control (Fig. 2). One copy of a previously described polymorphism in exon 14 (c.2045A>G, p.His682Arg) was the only other polymorphism detected in the colt (Haase et al. 2009a). The exon 14 polymorphism has been found in multiple breeds and is not associated with a dominant white phenotype. The sire was heterozygous at this locus, whereas the dam was homozygous for the wild type.

The deletion detected in exon 3 has not previously been documented in horses. However, there are four previously identified mutations leading to truncation of the KIT protein which have been linked to dominant white phenotypes, two of which are the result of nucleotide deletions (Haase et al. 2007). The W3 allele is the most similar to the mutation detected in the colt: it is a nonsense mutation in exon 4 that results in receptors lacking the intercellular portion, transmembrane domain, ligand binding domain and part of the extracellular domain (Haase et al. 2007). The W3 allele was only found in the heterozygous state and had no phenotype other than depigmentation reported. Similarly, the proband’s mutation is heterozygous, and thus he is still able to produce some full-length KIT transcripts, which accounts for the pigmented areas and his apparent health. The depigmentation is likely the result of the deletion in exon 3 leading to severely truncated KIT receptors.

Unfortunately, around 5 weeks of age the colt was found dead in his stall. The body was sent to the University of
Kentucky Livestock Disease Diagnostic Center. Necropsy found mild focal suppurative bacterial omphalitis, severe chronic gastric ulceration and mild multifocal hepatitis. The death was attributed to cardiovascular shock and ventricular arrhythmia of undetermined aetiology.

Gastric ulcers have been documented for some human and mouse KIT mutants (Kitamura & Hirotab 2004). No similar cardiovascular defects were documented for murine allele variants (Bult et al. 2008). It is possible that the exon 14 polymorphism only has a mild effect on phenotype that is not usually detected; there were no detrimental effects reported in the original study documenting this allele (Haase et al. 2009b). However, if the mutation was present in the non-truncated copy of KIT, thus making the colt a compound heterozygote, the receptor’s function may have been affected enough to cause detrimental effects that could have contributed to the foal’s death. Unfortunately, as only limited tissue samples were available from the colt, it cannot be concluded with certainty that effects of the exon 3 deletion in KIT contributed to his death.

Acknowledgements
The authors thank the owner of the colt for providing DNA samples and photographs of the dam and colt, Dr James MacLeod for providing samples and Dr David Bolin for his input in the case.

Conflicts of interest
The authors have declared no potential conflicts.

References

Supporting information
Additional supporting information may be found in the online version of this article.
Table S1 Primer pairs used to amplify the 21 exons of the equine KIT gene.
As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Lavender foal syndrome in Arabian horses is caused by a single-base deletion in the \textit{MYO5A} gene

A. Bierman*, A. J. Guthrie† and C. K. Harper‡

*Faculty of Veterinary Science, Veterinary Genetics Laboratory, University of Pretoria, Private Bag X04, Onderstepoort 0110, South Africa.
†Faculty of Veterinary Science, Equine Research Centre, University of Pretoria, Private Bag X04, Onderstepoort 0110, South Africa.
‡Faculty of Veterinary Science, Veterinary Genetics Laboratory and Department of Production Animal Studies, University of Pretoria, Private Bag X04, Onderstepoort 0110, South Africa

Summary

Lavender Foal Syndrome (LFS) is a rare autosomal recessive lethal disorder affecting Arabian foals which is also characterized by a dilute coat colour and severe neurological signs. Dilute mouse and rat mutants, and Griscelli syndrome type 1 in humans, which are characterized by similar clinical signs, are caused by mutations in the \textit{MYO5A} gene. \textit{MYO5A} was, therefore, identified as a possible candidate gene for LFS. Sequencing of the coding region identified a single-base deletion in a conserved region of the tail domain. The deletion produces a truncated protein product through the insertion of a premature stop codon (p.Arg1487AlafsX13). The deletion was confirmed as the causative mutation by genotyping affected, carrier and normal individuals.

Keywords Arabian, coat colour dilution lethal, dilute lethal, horse, lavender foal syndrome, \textit{MYO5A}, myosin-Va, neurological, single-base deletion.

Lavender Foal syndrome (LFS), also referred to as dilute lethal (Bowling 1996), lethal LFS (Schott & Petersen 2005) and coat colour dilution lethal (Fanelli 2005), is a condition only reported to affect Arabian foals (Bowling 1996; Fanelli 2005; Madigan 1997; Pascoe & Knottenbelt 1999) and is inherited in an autosomal recessive manner (Bowling 1996). Affected foals have an unusual dilute coat colour, demonstrate various neurological abnormalities, are not able to stand and nurse, and if the typical coat colour characteristic is overlooked may be incorrectly diagnosed as suffering from neonatal maladjustment syndrome (NMS), neonatal septicemia or neonatal encephalopathy (Bowling 1996; Page et al. 2006). Post-mortem evaluations have failed to yield any macroscopic findings, suggesting a biochemical cause for LFS. The prevalence of LFS remains unknown (Fanelli 2005; Page et al. 2006). A recent study using SNP chip technology has allowed researchers to identify a candidate region of 1 Mb containing 216 candidate genes for the disease (Gabreski et al. 2009). Dilute mouse and rat mutants have defects in melanosome transport and a failure of their release into keratinocytes (Futaki et al. 2000; Takagishi & Murata 2006). Mice homozygous for the dilute mutation have dilute coat colour, show severe ataxia and opisthotonus and die within 3 weeks (Huang et al. 1998). These symptoms are very similar to those observed in LFS. The neurological aspect of the condition arises from aberrant transport of organelles in the neurons which in turn impairs synaptic regulation (Takagishi et al. 2007). The dilute colour observed is not because of abnormal pigment production but an abnormal dispersal of melanosomes within the hair shafts (Au & Huang 2002; Strobel et al. 1990). In man, Griscelli syndrome type I is an autosomal recessive genetic disorder associated with a mutation in \textit{MYO5A}, which is characterized by pigmentary dilution with hypotonia, marked motor developmental delay and mental retardation (Pastural et al. 1997).

To identify the molecular defect underlying this disorder, we sequenced the coding region of the \textit{MYO5A} gene in normal, affected and carrier animals. DNA was extracted from tissue and blood samples of four affected foals, their carrier sires and dams as well as four unaffected, non-carrier individuals using a phenol-chloroform DNA extraction protocol with ethanol washes. PCR amplification of the \textit{MYO5A} coding region was performed using 12 sets of primers (Lavender1–12 in Table S1) designed to amplify 12 regions of coding sequence conserved between \textit{Mus musculus} and

Address for correspondence
C. K. Harper, Faculty of Veterinary Science, Veterinary Genetics Laboratory and Department of Production Animal Studies, University of Pretoria, Private Bag X04, Onderstepoort 0110, South Africa.
E-mail: cindy.harper@up.ac.za

Accepted for publication 13 April 2010
Equus caballus. PCR amplification was performed for 35 cycles of 45 s at 95 °C, 1 min at 60 °C and 2 min at 72 °C with a final extension step of 8 min at 72 °C in 20 μl reaction volumes. PCR products were purified using the Invitrek MSB Spin PCRapace Kit by Invisorb® and sequenced in 10 μl reactions using Bigdye v3.1 sequencing chemistry (Applied Biosystems) on the ABI 3130xl Genetic Analyzer (Applied Biosystems). Sequences of the affected and normal individuals were deposited in GenBank under the accession numbers HM063929 and HM063930 respectively.

Comparison of the nucleotide sequences between affected and normal individuals revealed only one sequence variation in the fragment amplified by primer set seven. A single-base deletion of cytosine at 4459 bp (c.4459delC) produced a frameshift that resulted in a premature stop codon (p.Arg1487AlafsX13). The substituted amino acid, Arginine, is conserved between human, mouse, rat and horse sequences, and the resulting truncation of almost half the protein tail (Fig. 1) is the causative mutation for the disorder. Direct sequencing of the region containing the deletion confirmed that affected and carrier individuals were homozygous and heterozygous for the deletion, respectively, while the deletion did not occur in normal individuals (Fig. 1). To confirm the specificity of the mutation, nine samples from individuals related to the four carrier foals as well as five unrelated control samples were genotyped. Fluorescently labelled primers were designed to amplify a 154 bp fragment flanking the deletion site (ACDF01 in Table S1). PCR amplification was performed for 35 cycles of 45 s at 95 °C, 45 s at 60 °C and 1 min at 72 °C with a final extension step of 8 min at 72 °C in 20 μl reaction volumes. PCR products were subjected to capillary electrophoresis using an ABI 3130xl Genetic Analyzer (Applied Biosystems). All affected individuals showed a single peak with a fragment length of 153 bp on STRAND software (Toonen & Hughes 2001; version 2.4.16), while normal individuals had a single peak at 154 bp. Heterozygous carriers had two characteristic peaks of 153 bp and 154 bp (Fig. 1).

Myosins are cargo binding proteins that move along actin filaments, amongst others, driven by ATP hydrolysis (Woolner & Bement 2009). Myosin-Va is expressed in the brain and skin (Takagishi & Murata 2006), where it functions in organellar transport and membrane trafficking (Reck-Peterson et al. 2000). Myosin-Va also plays a role in axonal and dendritic transport in neurons (Langford & Molyneaux 1998; Reck-Peterson et al. 2000). The myosin heavy chain consists of an N-terminal globular head that is conserved across the class V myosins, a neck region with an alpha-helical structure and a tail domain consisting of a helical coiled-coil interspersed with globular domains and ending in a C-terminal globular tail. The head of the protein contains sites for ATP hydrolysis and actin binding and is approximately 765 amino acids in length. The neck region of approximately 147 amino acids contains the calmodulin binding sites in the form of six IQ motifs (Sellers 2000). The alpha-helical tail is the site of dimerization, while its distal globular segment is responsible for cargo binding and protein localization (Langford & Molyneaux 1998). The globular tail of myosin-Va contains at least two separate binding sites with a high propensity for interacting with a wide variety of cargos.

**Figure 1** Single-base deletion of MYO5A causing the p.Arg1487AlafsX13 mutation. (a) Schematic representation of myosin-Va protein sequence. The deletion results in a frame shift with a premature stop codon 12 residues from the deletion. (b) Electropherograms of sequencing data (left) show the deletion of C in affected and carrier individuals, and fragment analysis (right) highlights the heterozygous nature of the carrier animals, as can be seen from the two peaks.
range of different cargo molecules (Li & Nebenführ 2008). The c.4459delC mutation described here lies within the globular tail domain of the myosin-Va protein. The region where the c.4459delC mutation lies is within a well known deletion of a dilute mouse mutant, D20J, which is known to occur in all splice variants (Strobel et al. 1990).

Griscelli syndrome type 1 in man is associated with pigment dilution and neurological symptoms (Pastural et al. 1997), while the dilute lethal mouse and dilute-opisthotonus rat mutants exhibit dilute coat colours and intermittent opisthotonus (Futaki et al. 2000; Huang et al. 1998). These conditions, like LFS, are all associated with mutations in the MYO5A gene. Our present study shows that LFS is an autosomal recessive condition caused by a single-base deletion in the MYO5A gene on chromosome 1 of the horse.

Acknowledgements

We thank the owners of the horses used in this study for submitting material and the Veterinary Genetics Laboratory for financial support.

Conflicts of interest

The authors have declared no potential conflicts.

References


Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 List of primers used for sequencing and genotyping.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Time of day influences cytokine and clock gene response to immune stimulation in equine whole blood

O. F. McGlynn, J. A. Browne, C. M. Blake and B. A. Murphy
Veterinary Sciences Centre, School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland

Summary

The immune system demonstrates regularly recurring rhythmic variation that has important medical and veterinary clinical implications. Emerging evidence supports the existence of a bidirectional immune–circadian relationship during homeostatic challenge. In peripheral blood, circadian rhythmicity manifests in terms of clock gene expression, circulating levels and function of leucocyte cell populations and cytokine production. The involvement of interleukin 6 (IL6) in the transition from innate to acquired immunity is known, and inflammatory mediators can, in turn, phase shift peripheral molecular clocks. This study provides an initial investigation of diurnal variation in circadian clock and immune mediator response to antigenic challenge in the horse. Blood samples were collected at 4-h intervals at circadian times (CT) 0, 4, 8, 12, 16 and 20 from healthy Thoroughbred fillies maintained under winter photoperiod. Heparinized blood was cultured at 37 °C for 6 h with or without lipopolysaccharide (LPS) (1 μg/ml). Cells were harvested by the addition of PAXgene™ RNA stabilizer solution to immediately preserve the RNA profile. Analyses of Taqman qPCR expression data reveal a significant CT X LPS interaction for the canonical clock genes PER2, CRY1, ARNTL, NR1D2 (P < 0.001, P < 0.05, P < 0.05, P < 0.01 respectively) and the immunomodulatory cytokine IL6 (P < 0.0001). LPS up-regulated IL1B levels (P < 0.01), but there was no effect of CT. These results confirm that peripheral blood differentially responds to antigenic challenge over the 24-h cycle, impacting on our understanding of the pathophysiology of inflammatory responses. Further research in this area should provide valuable information regarding the optimum time for drug delivery and vaccination.

Keywords blood, circadian, clock, cytokine, interleukin-6, LPS.

Equine chronobiology investigates how natural and artificially imposed environmental variation influences health and welfare in the horse (Murphy 2009). Cytokines play a critical role in the induction and regulation of immune effector functions, and these in turn are regulated by the circadian clock (Keller et al. 2009), which functions at the molecular level via autoregulatory transcription–translation feedback cycles of the core clock genes (Ripperger & Schibler 2001). It has been suggested that the direction of an immune response may depend on the cytokine environment at time of day of antigen presentation (Petrovsky & Harrison 1997). Light is the primary synchronizer of circadian (approximately 24 h) rhythms, but accumulating evidence highlights a bidirectional regulatory relationship between the immune and circadian systems (Coogan & Wyse 2008). Previously, equine core clock genes were identified (Murphy et al. 2007a), and it was shown that acute inflammation synchronized clock gene expression in equine peripheral blood in vivo (Murphy et al. 2007b). However, it was hitherto undetermined whether the observed coordinated upregulation of clock genes was time of day sensitive or simply an acute response to antigenic stimulus. Equine viral infections are responsible for disease epidemics around the world, resulting in major economic loss to the equine industry. This study aims to make initial inroads into unravelling the complex relationships between the immune and circadian systems in the horse, such that vaccine efficacy can be improved by choosing the optimum time of day for administration.

Address for correspondence
B. A. Murphy, Veterinary Sciences Centre, University College Dublin, Belfield, Dublin 4, Ireland.
E-mail: Barbara.murphy@ucd.ie

Accepted for publication 08 June 2010
Four healthy Thoroughbred fillies (Equus caballus) (3–4 years of age) were housed under a light/dark (LD) cycle that mimicked the natural external photoperiod of early December (8 h L and 16 h D) at longitude W6.8, latitude N53.2 (County Kildare, Ireland). Blood samples were collected via jugular venipuncture at 4-h intervals beginning at lights on or circadian time (CT) 0. All procedures involving animals were approved by the Animal Research Ethics Committee of University College Dublin. Two hundred microlitres of heparinized blood was aliquotted into 24-well plates at each time point and treated with lipopolysaccharide (LPS) (1 μg/ml) or left untreated. Plates were incubated at 37 °C with 5% CO₂ for 6 h prior to harvesting by the addition of PAXgene™ RNA stabilizer solution (Qiagen, Hoddesdon, Hertfordshire, UK).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession number</th>
<th>Forward primer (5’-3’)</th>
<th>Reverse primer (5’-3’)</th>
<th>Probe (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL6</td>
<td>NM_001082496.1</td>
<td>GGATGCTTCCAATCTGGGTTCAAT</td>
<td>TCCGAAGGACACGGTGATTTTT</td>
<td>ATCAGGCGCTCTCTCCTG</td>
</tr>
<tr>
<td>IL1B</td>
<td>NM_001082526</td>
<td>CCGACACCACTGACATGAGA</td>
<td>ATCTTCCCTAAGAACAGTCGAC</td>
<td>ATGCGCCGTCAGAAG</td>
</tr>
<tr>
<td>CRY1</td>
<td>DQ988039.1</td>
<td>CGGTTTGGGTCTGCTGCTG</td>
<td>CGGAGATGGGGTTCTCCTCATTT</td>
<td>TGTGCAACTTTATAGGCTGATTTT</td>
</tr>
<tr>
<td>PER2</td>
<td>EF015879.1</td>
<td>CCAGCAATAATTGGAGCAGTCA</td>
<td>GGCATCAGAGCAGCAAGCAGG</td>
<td>TGTCTACAGCGGCACAC</td>
</tr>
<tr>
<td>ARNTL</td>
<td>NM_001081921</td>
<td>CCACCAATCCACACAGAAGCAA</td>
<td>TCTTCCCTCCGTCACATCTCT</td>
<td>CACCTCAGCTCAGGAGCAG</td>
</tr>
<tr>
<td>ACTB</td>
<td>NM_001081838</td>
<td>GCCGTCTCCCCTTCCCAT</td>
<td>GCCCCACGTATGTCCTCTCTG</td>
<td>TGGATTTGGTTGTCG</td>
</tr>
<tr>
<td>NR1D2</td>
<td>XM_001492478</td>
<td>ACGGGTCTTCCAAGACGAGCA</td>
<td>AGATGCTGCTGGCTCTCCTACA</td>
<td>ACCATTGCGAGTGGCTTAC</td>
</tr>
</tbody>
</table>

Table 1 qPCR primer and probe sequences for equine genes.

Figure 1 Diurnal variation in gene expression from equine whole blood collected at 4-h intervals over the 24-h LD cycle and cultured for 6 h with (solid line) or without (dotted line) lipopolysaccharide. Data are presented as means ± SE (n = 4 per timepoint). Within-time point significant Bonferroni statistical post hoc differences are indicated by superscript lettering; a, b = P < 0.05; c, d = P < 0.01; e, f = P < 0.001.

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 202–204
gen), to immediately stabilize transcription and preserve the RNA profile. The success of this protocol was previously reported in human clinical studies (Carrol et al. 2007). Total RNA was isolated using the PAXgene™ Blood RNA kit, quality-checked using an Agilent 2100 Bioanalyzer (Agilent Technologies), and then 70 ng of RNA from each sample was converted to cDNA using the High-Capacity cDNA Reverse Transcription kit (Applied Biosystems) and stored at −20 °C. Taqman qPCR assays were designed using Genbank (NCBI) equine sequences and previously published assays to detect expression of the core clock genes PER2, CRY1, ARNTL (Murphy et al. 2006) and NR1D2 and the cytokines IL6 and IL1B (Quinlivan et al. 2007) (Table 1). Oligonucleotide primers were commercially synthesized by Eurofins MWG Operon and dual-labelled fluorescent probes by Biosearch Technologies. Each 20-µl qPCR mixture contained 5 µl cDNA (1.5 ng RNA equivalents), 300 nM of each forward and reverse primer, 250 nM of probe, 10 µl Taqman Universal Master Mix (2× Applied Biosystems) and nuclease-free H2O (Sigma-Aldrich) and was run on the ABI 7500 Real Time PCR System (Applied Biosystems). For each PCR run, a standard curve was generated using twofold serial dilutions of pooled cDNA. The PCR programme consisted of one cycle of 50 °C for 2 min and 95 °C for 10 min, followed by 45 cycles of 95 °C for 15 s and 60 °C for 1 min. All samples were run in duplicate. ACTB was selected for its stability as a reference gene using GeNorm.

Two-way repeated measures ANOVA were conducted using GRAPHPAD Prism Version 4.0. The values of the relative expression of mRNA are presented as means ± SE, and a P-value of <0.05 was considered significant. A significant CT × LPS interaction for the canonical clock genes PER2, CRY1, ARNTL, NR1D2 (P < 0.001, P < 0.05, P < 0.05, P < 0.01 respectively) and the immunomodulatory cytokine IL6 (P < 0.0001) was observed (Fig. 1). There was a significant main effect of CT for PER2, CRY1 and ARNTL (P < 0.0001, P < 0.001, P < 0.01 respectively), supporting previous findings of a ‘phase shifting’ effect of inflammatory mediators on the circadian clock (Tsuchiya et al. 2005). The main effects of CT and LPS were both significant for NR1D2 (P < 0.01). LPS treatment significantly upregulated IL1B expression but was unaffected by time of day. These results support the view that the molecular clockwork differentially responds to an inflammatory stimulus over the 24-h cycle, which in turn regulates the transcriptional response of immune cells to an antigen (Keller et al. 2009).

Interestingly, IL6 is differentially upregulated at CT 20 in equine circulation, in opposing phase to the temporal pattern observed in mice (Marpegan et al. 2009). This likely reflects a contrasting temporal immune surveillance regulation between diurnal and nocturnal species. Our finding suggests that equine Th1 humoral responses may be favoured when antigen exposure occurs in the evening. This has clear implications regarding the potential optimal time of day for vaccination in the horse, emphasizing the importance of further research in this area. In summary, this work highlights the complexity of circadian–immune regulation in a large diurnal mammal and potentially explains phenomena such as diurnal variation in resolution of inflammatory insult (Marpegan et al. 2009).

Conflicts of interest

The authors have declared no potential conflicts.

References


Genetic diversity in the Belgian Draught Horse breed as revealed by pedigree analysis and molecular marker data


Department of Biosystems, KULeuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium

Accepted for publication 11 July 2010

Source/description: The native Belgian draught horse (BTP) was bred as a large and well-muscled horse to serve in agriculture and transportation. The reduction in population size after 1950 and the unequal usage of stallions may have reduced the breed’s diversity. Genetic diversity was assessed using both pedigree data and molecular marker genotypes, making a comparison between inbreeding and multilocus heterozygosity feasible.  

Animals: Pedigree analysis was performed on horses born from 2003 to 2009, which corresponds to an approximate generation interval of one (computed in previous analysis). PEDIG and own software routines were used.  

Blood or hair root samples were collected at national shows in 2008 and 2009 (n = 127), mainly from stallions, and these were genotyped for 14 microsatellites, namely AHT4, AHT5, ASB2, ASB17, ASB23, CA425, HMS3, HMS6, HMS7, HTG4, HTG6, HTG7, HTG10, VHIL20. These include the nine core microsatellites of the ISAG Horse Panel and five additional markers that might be added to this panel in the future.  

Pedigree diversity: Inbreeding levels in the breed are low (Table S1), but in view of the limited pedigree depth and completeness (Table 1), these levels should be interpreted as a lower limit. Frequent import of animals that are not well integrated with the Belgian data (because of their foreign id numbers) has reduced pedigree depth and may also explain the minor loss of founder diversity (ratio of effective founders to effective ancestors). Although more founders were counted in mares than in stallions, effective diversity is equal between the sexes. Effective size, computed from individual inbreeding, varies from 83 to 102, which is in the range recommended for livestock. The fluctuations in effective size and in inbreeding may be related to the introduction of recent founders.  

Table 1 Diversity measures of Belgian draught horses (born 2003–2009) computed from the registered pedigree.

<table>
<thead>
<tr>
<th></th>
<th>Mares</th>
<th>Stallions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference population</td>
<td>3453</td>
<td>2559</td>
</tr>
<tr>
<td>PEC5</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td>Founders</td>
<td>1247</td>
<td>1123</td>
</tr>
<tr>
<td>Effective founders</td>
<td>85</td>
<td>84</td>
</tr>
<tr>
<td>Effective ancestors</td>
<td>42.7</td>
<td>42.9</td>
</tr>
<tr>
<td>Contribution of the single most important ancestor</td>
<td>8.2%</td>
<td>8.1%</td>
</tr>
<tr>
<td>Variability explained by the 6 most important ancestors</td>
<td>34%</td>
<td>31%</td>
</tr>
<tr>
<td>Effective founder genomes</td>
<td>20.0</td>
<td>19.7</td>
</tr>
</tbody>
</table>

However, a genetic concentration is expected in the coming generations, because six ancestors explain 31% and 34% of the remaining variability for stallions and mares, respectively. These ancestors were born from 1961 to 1995, which is only 4–6 generations from the sample studied here. Interestingly, the largest contribution to genetic variability is attributed to one mare (8%, born in 1976).  

Molecular data: Allele frequencies, polymorphism information content, heterozygosity and Hardy–Weinberg equilibrium (HWE) were computed using Genepop⁴ (Table S2).  

On average, 6.07 alleles per locus were counted (standard deviation 1.73), which is on the upper side of the number of alleles reported in German heavy horse breeds (6). The least polymorphic locus in the BTP was HTG6 (three alleles) and the most variable was ASB17 (10 alleles), which is in line with overall allele counts. Heterozygosity values per locus ranged from 0.805 to 0.264. The overall expected and observed heterozygosities (0.657 ± 0.035 and 0.663 ± 0.012) are very similar to values reported previously in draught horses. HWE was not met at locus level for ASB17, HMS6 and HTG10, but a global test did not reject HWE.  

The average multilocus heterozygosity⁵ of 0.656 (±0.14) was significantly correlated with the level of inbreeding computed from the pedigree (Pearson correlation = 0.27, P-value 0.004). Our estimate is equal to the mean correlation computed from 12 animal populations. The low level of inbreeding in the Belgian draught horse and consequently the low variation in inbreeding levels might explain the poor correlation between (pedigree) inbreeding level and multilocus heterozygosity.  

Comments: Pedigree analysis and marker data indicate an ‘intermediate’ level of diversity in the Belgian draught, when compared to other horse breeds. Current effective size is in the recommended range and seems to be improving. However, marginal contributions from ancestors indicate a genetic concentration, which may compromise genetic diversity in the coming generations. The two approaches to assess diversity have led to the same general conclusions.  

Acknowledgements: We thank the breeding associations VFTB and KMBT for providing the pedigree file and for cooperation during sample collection. Funding was provided by FWO (Research Foundation – Flanders).

Conflicts of interest
The authors have declared no potential conflicts.

References

Correspondence: S. Janssens (Steven.Janssens@biw.kuleuven.be)
Conflicts of interest
The authors have declared no potential conflicts.

Supporting information
Additional supporting information may be found in the online version of this article.
Table S1 Diversity measures of Belgian draught horses based on pedigree information.

Table S2 Descriptive statistics for 14 microsatellite loci typed on 127 Belgian draught horse individuals.
As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.
Association analysis of candidate SNPs in TRPM1 with leopard complex spotting (LP) and congenital stationary night blindness (CSNB) in horses


*Department of Biology, University of Tampa, Tampa, FL 33606, USA. †Quill Lake, Saskatchewan S0A3E0, Canada. ‡Faculty of Veterinary Sciences, The University of Sydney, NSW Australia. §Institute of Genetics, University of Berne, Bremgartenstr. 109a, 3001 Bern, Switzerland. ‡Department of Biomedical Sciences and Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, SK S7N5B4, Canada. **Department of Small Animal Clinical Sciences Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, SK S7N5B4, Canada

Accepted for publication August 15, 2010

Description:
Leopard complex spotting occurs in several breeds of horses and is caused by a single incompletely dominant autosomal locus, LP.1 In the Appaloosa breed, homozygosity for LP has been associated with congenital stationary night blindness (CSNB).2 Decreased expression of TRPM1 has been implicated as the cause for both LP and CSNB.3 LP and CSNB were fine-mapped to a 173-kb haplotype on ECA1 and Illumina sequencing identified six SNPs for further investigation (Table 1).4 In this study, we investigate these six SNPs for association with CSNB in the Appaloosa and for association with LP in several breeds of horses.

Genotyping:
DNA was isolated from blood or hair of unrelated horses from breeds segregating for LP: Appaloosa (N = 205), Knabstrupper (N = 66), Noriker (N = 112), American Miniature (N = 63), Pony of America (N = 20), British Spotted Pony (N = 25), and Australian Spotted Pony (N = 10). DNA was isolated from the Thoroughbred (N = 37) and American Quarter Horse (N = 3), which are breeds not segregating for LP. LP genotype and CSNB status were determined as previously described.1 Initially, all six SNPs were investigated by PCR amplification and sequencing DNA from 10 individuals (Table 1). Three of these SNPs did not show complete association with LP genotype in this panel and were excluded as the causative mutation and thus not investigated further. The other three SNP showed complete association and were either genotyped by direct sequencing or by custom TaqMan genotyping assays in 531 additional individuals (Assay IDs AHS0Q19, AHRRSV1, and AHT9058, Applied Biosystems) (Table 1). All TaqMan assays were performed in 5-µl reactions on a Mastercyler® ep realplex thermocycler (Eppendorf).

SNP Association with LP and CSNB:
ECA1 g.108281765T>C, ECA1 g.108288853C>T, and ECA1 g.108337069T>G genotypes were analysed for association with LP genotype and CSNB status by chi-squared analysis. All three SNPs were completely associated with LP (N = 513, X² = 1026, P < 0.0005) and CSNB (N = 28, X² = 28, P < 0.0005). It is possible that one of these SNPs is the causative mutation for LP and CSNB. However, none of these are located in exonic regions that have previously been characterized, and thus further investigation is warranted.4 It is also likely that these SNPs may simply be associated and are not causative. Nevertheless, any of these SNPs could be used as a DNA test for LP and CSNB until the causative mutation has been identified or confirmed.

Acknowledgements:
The authors thank all of the horse owners who provided samples. We thank Karla Brown, Scott Lawson, Elizabeth Kowalski and Janelle Nelson for their assistance in DNA extraction and Dominic Trillizio for his technical assistance. We acknowledge Eppendorf for use of equipment. This study was supported by the Heather Ryan and L. David Dube Veterinary Health and Research Fund and a Dana Faculty Development Grant from the University of Tampa.

Conflicts of interest
The authors have declared no potential conflicts.

References
2 Sandmeyer L. et al. (2007) Vet Ophthalmol 10, 368–75
3 Bellone R.R. et al. (2008) Genetics 179, 1861–70

Correspondence: R R Bellone rbellone@41.edu

Table 1 Candidate SNPs investigated for LP and CSNB. PCR conditions and results of SNP association with LP genotype in 10 animals from five different breeds are presented.

<table>
<thead>
<tr>
<th>SNPs tested</th>
<th>Forward primer sequence</th>
<th>Reverse primer sequence</th>
<th>Annealing temp. (°C)</th>
<th>PCR product size (bp)</th>
<th>Association with LP genotype in small panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECA1 g.108281765T&gt;C</td>
<td>5’-AGTGGGAGTCCCTGCACTG-3’</td>
<td>5’-CTTAGCTCAAGCCCTTCTC-3’</td>
<td>65</td>
<td>326</td>
<td>10/10 associated</td>
</tr>
<tr>
<td>ECA1 g.108288853C&gt;T</td>
<td>5’-CATGTAACCGGCCGCTTAAA-3’</td>
<td>5’-GCAGCCTCCAAAATCAGCTC-3’</td>
<td>58</td>
<td>223</td>
<td>10/10 associated</td>
</tr>
<tr>
<td>ECA1 g.108337069T&gt;G</td>
<td>5’-CTTAGTGGAGGCCAGGTCCAG-3’</td>
<td>5’-ACCTGGCAAGCCCTTCTC-3’</td>
<td>65</td>
<td>485</td>
<td>10/10 associated</td>
</tr>
<tr>
<td>ECA1 g.108489901G&gt;A</td>
<td>5’-TGCTCTCGGCACGCCCC-3’</td>
<td>5’-CTAGCCCCCCCGACCCCAT-3’</td>
<td>65</td>
<td>153</td>
<td>8/10 associated</td>
</tr>
<tr>
<td>ECA1 g.108497669C&gt;A</td>
<td>5’-CCCTGACGCTCGGACTCTT-3’</td>
<td>5’-GCAAAGTGGGAAGAACACTC-3’</td>
<td>60</td>
<td>246</td>
<td>9/10 associated</td>
</tr>
<tr>
<td>ECA1 g.108497990C&gt;T</td>
<td>5’-TTCCTCTAGCAGGTGGCACTT-3’</td>
<td>5’-GATGCTGAGTTCAGCCACTAA-3’</td>
<td>62</td>
<td>306</td>
<td>8/10 associated</td>
</tr>
</tbody>
</table>

*SNPs with complete association were investigated further and were completely associated in a panel of 541 horses. CSNB, congenital stationary night blindness.
LP, leopard complex spotting.

© 2010 The Authors, Journal compilation © 2010 Stichting International Foundation for Animal Genetics, 41 (Suppl. 2), 207
ANIMAL GENETICS
Volume 41 Supplement 2 December 2010

CONTENTS

Foreword

1. Horse genomics and the Dorothy Russell Havemeyer Foundation
   E Bailey

Articles

2. Genome-wide SNP association–based localization of a dwarfism gene in Friesian dwarf horses

8. Linkage disequilibrium and historical effective population size in the Thoroughbred horse
   J I Corbin, S C Ibbott, J E Swinburne, M Vaudin, S C Bishop, J A Wooliams

16. RT-qPCR Comparison of mast cell populations in whole blood from healthy horses and those with laminitis
   S A Brooks, E Bailey

23. Transcription of L1seq-derived sequences in exercise-induced stress in horses
   S Capomaccio, A Verini-Supplizi, G Galla, N Vitulo, G Barsacchia, M Felicetti, M Silvestrelli, K Gappelli

28. A genome-wide association study for racing performances in Thoroughbreds clarifies a candidate region near the MSTN gene
   T Tsuzaki, T Miyake, H Nakano, N Ishida, K Hirota, Y Nakano

36. Refinement of quantitative trait loci on equine chromosome 10 for radiological signs of navicular disease in Hanoverian warmblood horses
   M S Lopes, U Braemer, A da Câmara Machado, O Distl

41. Molecular heterogeneity of XY sex reversal in horses
   T Raudsepp, K Dar transmitted to the

43. Genetic diversity in the Maremmano horse and its relationship with other European horse breeds
   M Felicetti, M S Lopes, A Verini-Supplizi, A da Câmara Machado, M Silvestrelli, D Mendonça, O Distl

56. Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance
   E W Hill, J Conway, B A McGivney, D E MacHugh

64. Genetics of swayback in American Saddlebred horses
   D Cook, P C Gallagher, E Bailey

72. Comparative human–horse sequence analysis of the CYP3A subfamily gene cluster
   A Schmitz, S Donnell, L M Peters, T Leib, M Mevissen, B Hauxer

80. A genome-wide scan for tying-up syndrome in Japanese Thoroughbreds
   T Tsuzaki, K Hirota, S Sužeta, N Ishida, T Miyake, H Oki, T Hayagana

Contents continued overleaf

This journal is available online at Wiley Online Library. Visit www.wileyonlinelibrary.com to search the articles and register for the table of contents e-mail alerts.