Genomic Tools And Strategies For Studying Hereditary Conditions In Horses

S. Mikko[*], L.S. Andersson[*], S. Eriksson[*], J. Axelsson[*], and G. Lindgren[*]

Introduction

The era of genomic studies presents new opportunities to find the genetic background of heritable diseases and conditions in horses. Before the horse genome was sequenced, molecular genetic research mainly focused on monogenic traits, such as Severe Combined Immunodeficiency (SCID) (Bailey et al., 1997; Shin et al., 1997), Overo Lethal White foal Syndrome (OLWS) (Metallinos et al., 1998; Santschi et al., 1998), and Hyperkalaemic Periodic Paralysis (HYPP) (Bowling et al., 1996; Rudolph et al., 1992) Many of these genes were found using a candidate gene approach, together with knowledge from comparative studies. In many cases, causative mutations were first found in the human or mouse, giving valuable information to horse geneticists. However, the search for genes associated with diseases, and defects, was slow and tedious. In the field of quantitative horse genetics, the main focus has been to characterize heritability levels and to calculate breeding values of traits important for performance, such as longevity, jumping ability, gaits, and conformation. Now that the horse genome sequence is published, new opportunities to find genes responsible for such performance related traits, as well as genes causing multifactorial diseases are available.

Human genetic research has been supported with large resources, in terms of funding, but also via commercially available tools in the form of assays and kits. Although animal studies, and especially horse genetics, has long suffered from low funding, but one advantage compared to human studies is that often large half-sib families are available. This resource would never be possible to generate in human studies. Due to inbreeding and strong selection for certain desirable phenotypes, linkage disequilibrium (LD) is much stronger in horses compared to humans (Wade et al., 2009). Given the high frequency of interesting traits in the general population, it is possible to utilize differences in LD and the available family material to reduce the number of markers to find linkage or association between genes and traits.

The total length of the EquCab2 assembly is 2.68 Gb, and comprises more than 20,000 protein-coding genes found in the 6.8X coverage assembly (Wade et al. 2009). 54,602 single nucleotide polymorphisms (SNP:s) were selected for the Equine 50K SNP chip, produced by the Broad Institute together with Illumina (San Diego, CA). To find genes of interest, substantial statistical analyses together with biological knowledge is necessary when performing genome wide scans and whole genome sequencing for causative mutations.

[*]Dept. of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, S-75007 Uppsala, Sweden

The selection of data, individuals and samples, is crucial for the success of the projects. The choice of strategy and design for the project is highly dependent on trait characteristics and complexity, e.g. if it is a monogenic or multifactorial trait. The degree of linkage disequilibrium (LD) of selected breeds will also be of great importance. In the past, large half-sib families segregating for traits of interest were used in linkage studies. With the advent of the genome reference sequence and the Illumina Equine 50K SNP chip, there has been a move towards association studies. These later experiments can be performed with fewer individuals that do not need to be related. SNPs are now the preferred marker for current studies given the ease of typing, however they are less informative than microsatellites and more of them are required to generate the same level of information.
The collection of samples is often time-consuming and costly. Blood samples are preferred for archival material however hair roots from the tail are easier and cheaper to collect, although it gives lower quality of the prepared DNA. Despite this, a hair sample used in one of our studies was typed for 0.97 of all markers on the Equine 50K SNP chip. Sample collection requires both a well established relationship with breeding associations and horse owners, as well as knowledge about breeding structure. In our case we have collected new samples, but also had very good use of samples collected over the years at the Animal Genetics Laboratory at our department.
Linkage analysis is preferred if half-sib family material is available. This method can easily detect linkage even with very few markers per chromosome. Large scale genome wide association studies are valuable when the pattern of inheritance is not known. If matched e.g. half-sib groups are not available for the affected and non-affected groups, respectively, equally unrelated individuals in both groups can be used if there is a careful control for stratification. One of the frequently used software tools in use today is PLINK (Purcell et al., 2007). PLINK was developed for large scale whole genome association analysis, and not for more general population genetic analyses. In the association analysis performed in PLINK, each marker or haplotype is analyzed separately. To be able to handle very large amounts of SNP data and estimate all marker effects simultaneously, Bayesian methodology can be used.
Below I will describe the different approaches we have used for some of our ongoing projects; mapping genes for Silver Dapple coat color & MCOA, Skeletal Atavism, Summer eczema, and Performance traits.

Silver Dapple coat color and MCOA

The Silver Dapple coat color phenotype in horses is inherited as an autosomal dominant trait caused by a single nucleotide mutation in *PMEL17* (Brunberg et al. 2006), thus, it is a monogenic trait. The causative mutation was therefore relatively easy to find even with a small number of individuals, and few markers. In this study we had access to a very well characterized family material, with good phenotypic information, and a sufficient number of individuals (one heterozygous stallion, 34 of his offspring, and 29 non-silver dams). In this way it was possible to perform linkage analysis using CRI-MAP (Green et al., 1990). Similar phenotypes were present in other species so we could use a comparative candidate gene approach to choose the most probable markers to be able to find significant linkage. In total we did genotype 41 microsatellite markers. One of the selected markers showed significant linkage to the candidate gene *PMEL17* responsible for the merle coat color in dogs (Clark et

al. 2006), silver in mouse (Martínez Esparza et al. 1999), and the dun and smoky phenotypes in chicken (Kerje et al. 2004) (Figure 1). There had also been reports about visual defects associated to *PMEL17* mutations, such as in the zebra fish *fading vision* (Schonthaler et al. 2005).

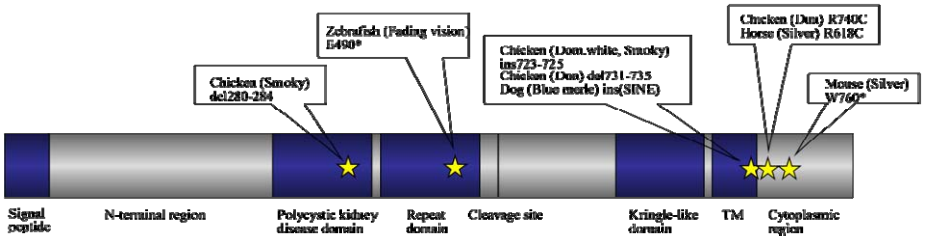

Figure 1. A schematic picture of the *PMEL17* protein with domains and known mutations. The transmembrane (TM) protein *PMEL17* has previously been shown to regulate hypopigmented phenotypes in mouse, chicken, dog, and zebra fish. The locations of known mutations associated with hypopigmentation in these species are indicated. R740C in chicken (Dun) is at the same location as the R618C in the horse (Silver).

In Rocky Mountain ponies the silver coat color was associated with an ocular defect, Equine Multiple Congenital Ocular Anomalies (MCOA), where the heterozygotes show an intermediate phenotype. MCOA has now shown to be associated with the same chromosomal region as *PMEL17*. This association was found by linkage analysis within 17 paternal half-sib families of the Rocky Mountain pony breed (Andersson et al. 2008). This study also showed that MCOA is inherited in a codominant fashion. So far, there are only two mutations found in this region, but further analysis by haplotype or "Identical By Decent" (IBD) mapping has now shortened the region down to hundreds of kilobases. Still, it is difficult to prove the association of these two mutations to MCOA. One way of shortening the interval would be to find a recombinant Silver colored individual, preferably from another breed. Since the Silver dapple coat color is not that common, we have so far not been able to identify such an individual. With the emergence of less expensive sequencing techniques, it will be possible to perform deep sequencing of this region to detect any so far undetected SNP:s. To ultimately prove that the mutation is causative to MCOA, we may design a knock-in mouse experiment.

Skeletal Atavism in Shetland ponies

Another monogenic trait that we recently started to study is Skeletal Atavism in Shetland ponies. This defect is recognized as fully elongated ulna and fibula that results in a shortening of radius, and tibia respectively, as well as a vertical outward rotation of the leg (Figure 2). The trait shows an autosomal recessive inheritance. Since the affected foals will suffer from this defective conformation, they are often utenized early in life, and not always reported as affected to the breed association. This is because in some breed associations, carriers are eliminated from breeding, and e.g. in the case of approved stallions, this could give significant economical loss. The rules for exclusion of breeding animals also differ

between different breed associations, causing trouble when exporting and importing breeding animals. Since the affected foals are often not reported, it is quite difficult to get hold of samples from them, and not all breeders want to cooperate. Also, if affected foals are not reported, possible carriers will remain undetected. This will be of great importance when selecting non-carrier individuals for the genetic studies.

Figure 2. A 19 year old Shetland pony with skeletal atavism. Notice the extremely short legs with the carpus and hock located almost directly under the elbow, and patella, respectively. Photo: Lisa Andersson.

For the monogenic trait Skeletal Atavism, we collected samples from affected individuals, carriers, and non carriers. It was very difficult to find unrelated individuals. Also, we did not want to choose matched half-sibs to affected individuals, as the risk of them being carriers, was significant. Instead we selected samples from six affected individuals and their respective parents, as well as another six known carriers. The 24 non-carriers were all stallions with more than 100 offspring, and selected not to be related to the known carriers, or affected. This was to eliminate the risk of them being undetected carriers. In an association study where the carriers are labeled as diseased, we will therefore get stratification. We were aware of this problem but chose this way to avoid undetected carriers. Furthermore, we cannot run a linkage analysis, since we do not have any non-affected half-sibs to control for the segregating disease allele. Results from the 50K SNP chip are currently under analysis. Average sample success rate is 0.974 which is quite good since most samples were old blood samples that had been stored in -20^{o}C for 10-20 years. In our further analysis we have to be aware of that not all non carriers are true non-carriers. One approach in the analysis of this disease is to label the carriers as affected, although they are in reality healthy, i.e. we designate the trait a dominant inheritance, and the minor allele is probably associated with the atavism. Another approach is to perform homozygosity mapping. There could be problems if the haplotype where the mutation originally arose is very common in the population. Regions where the affected individuals are homozygous are currently under analysis.

Summer eczema (insect bite hypersensitivity)

Summer eczema is a multifactorial disease that causes severe suffering for the horse and considerable economic loss for the horse owner. The outbreak of the disease is linked to a major environmental factor since the eczema is due to hypersensitivity to primarily biting midges of *Culicoides spp*., which are not known to be present in Iceland. The symptoms usually do not show until the horse is at least three years old, but the breakout can be at any age after that, thus it sometimes has a quite late onset, causing problems to select suitable non-affected control individuals.

In this project the strategy was to first evaluate if the condition was heritable, and after that, start the search for genes associated with the disease. It is likely to be inherited in a complex fashion with many associated genes, most of them with small effects. The highest number of affected individuals (30-50 %) is among the ones born in Iceland and exported to a country where the midges exist. Among Icelandic horses born in Sweden, the prevalence drops to about 8 %. To avoid the large environmental influence that horses imported from Iceland are subjects to, we chose to focus our genetic study on Icelandic horses born in Sweden. To do this we performed a survey study where the horse owners themselves graded the phenotype of the horse in four classes from healthy to severe eczema. The survey could either be answered by sending in a form or on the Internet. In total this study included 1250 horses sired by 33 stallions. The trait was shown to have a heritability of about 30 % on the underlying, continuous scale (Eriksson et al. 2008).

Although unrelated individuals are preferred in genome wide association studies, this was not possible in our case as the affected individuals were all related to some extent. To avoid stratification, a lot of effort was put into matching the controls so that they showed the same kind, as well as level, of relatedness. In total we collected 104 cases, and 105 matched half-sib controls. No completely unrelated individuals were included in the genotyping. Instead, due to this family material, we can see a deflation of marker effects. No stratification could be detected, and due to the relatedness we could not correct for Hardy-Weinberg Equilibrium (HWE). To avoid false negatives, horses were regarded as healthy only if they could be assumed to have been exposed to biting midges several seasons without showing symptoms, i.e. if they were older and were located in regions where *Culicoides spp* are common.

PLINK is used to analyze SNP data retrieved from the Equine 50K SNP chip. After frequency and genotyping pruning, there are 44,168 SNPs (MAF<0.01) left to analyze. Wade et al., (2009) have estimated that 100,000 SNP:s would be sufficient for association mapping within all breeds. Compared to this figure we may need more than these 44,000 markers to find an association to the complex trait we are studying. The analysis is still going on, both by using PLINK, and other modeling methods. The methodologies are constantly developing and new more efficient methods will soon be available, like the next generation sequencing.

Performance traits in Swedish trotters

The origin of the North Swedish Trotter (NST) provides a unique opportunity to identify genes influencing body constitution and racing performance. The NST has the same derivation as the North Swedish Horse (NSH), a draught horse used in farming and forestry. There has been some crossbreeding between Standardbreds (STB) and NST before paternity control was required in Sweden. The definition of "racing performance" in trotters may not

be the same as in thoroughbreds and may give different results when studying performance. Horse maintenance and training methods are examples of environmental factors of large importance. A remarkable improvement in racing performance of the NST has occurred during the last fifty years. This process should leave "genetic footprints" in the genome of NST in the form of chromosome segments originating from STB. Our hypothesis is to detect a previous gene flow between STB and NST, and a subsequent strong selection for racing performance in the NST. We are using a population genetics approach to identify chromosomal regions under positive selection. These regions are likely to harbor genes affecting both morphological and physiological traits important for physical performance.
Originally, 10 individuals from each of the three breeds were selected for microsatellite genotyping. We chose a rather small number of individuals in order to be able to genotype more markers. In total these 30 individuals were genotyped for 117 microsatellites. Higher resolution in this study has been achieved with the use of Illumina Equine SNP50 Bead Chip. The genotypes from the chip are currently being analyzed using PLINK, as well as population genetics software like GENEPOP on the Web 3.4 (Raymond et al. 1995), GENETIX 4.05 (Belkhir et al. 1996-2004), STRUCTURE 2.2 (Pritchard et al. 2000), and BOTTLENECK (Cornuet J.M. and Luikart G.,1997). PLINK does not have that many population genetics functions, but at the same time the more commonly used population genetics software have difficulties handling the massive amount of data generated from the 50K SNP chip. So far we have found some interesting chromosome regions to be further analyzed (Figure 3).

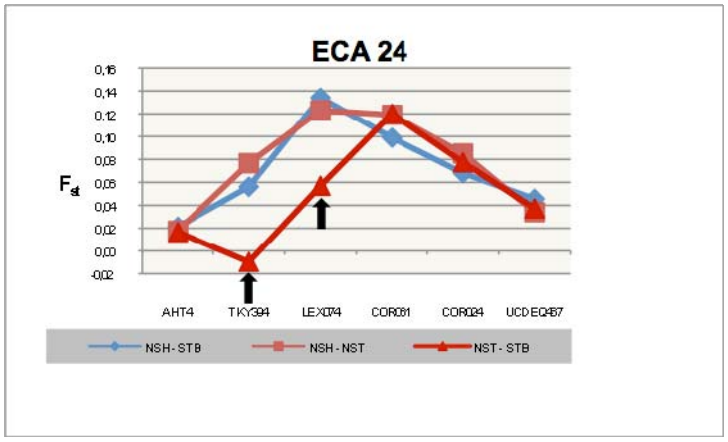

Figure 3. Pairwise F_{st} comparisons of loci at ECA 24, between the three breeds. The markers on the horizontal axis are ordered as they are placed on their respective chromosome. Black arrows indicate loci where the North Swedish Trotter is more similar to the Standardbred than to the North Swedish Horse.

References

Andersson, L.S., Juras, R., Ramsey, D.T. *et al.* (2008). *BMC Genet.*, 9:88-97.

Bailey, E., Reid, R.C., Skow, L.C., *et al.* (1997). *Anim. Genet.*, 28:268-273.

Belkhir, K., Borsa, P., Chikhi, L. *et al.* (1996-2004). GENETIX 4.05, Université de Montpellier II, Montpellier (France).

Bowling, A.T., Byrns, G., Spier, S. (1996). *Anim. Genet.,* 27:279-281.

Brunberg, E., Andersson L., Cothran, G. *et al.* (2006). *BMC Genet.,* 7:46-55.

Clark, L.A., Wahl, J.M., Rees, C.A., *et al.* (2006). *Proc. Natl. Acad. Sci. USA*, 103:1376-81.

Cornuet, J.M., and Luikart G. (1997). *Genetics,* 144:2001-2014.

Eriksson, S., Grandinsson, K., Fikse W.F. *et al.* (2008). *Animal* 2(3):360-365.

Green, P, Falls, K.A., Crooks, S. (1990). Documentation for CRI-MAP, version 2.4., Washington University School of Medicine, St. Louise, MO, USA.

Kerje, S., Sharma, P., Gunnarsson, U. *et al.* (2004). *Genetics*, 168:1507-1518.

Martínez Esparza, M., Jiménez Cervantes, C., Bennett, D.C. *et al.* (1999). *Mamm. Genome.*, 10:1168-1171.

Metallinos, D.L., Bowling, A.T., Rine, J. (1998) *Mamm. Genome,* 9:426-431.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). STRUCTURE 2.2, *Genetics*, 155:945-959.

Purcell, S., Neale, B., Todd-Brown, K. *et al.* (2007). *Am. J. Hum. Genet.* 81:559-575. URL: http://pngu.mgh.harvard.edu/purcell/plink/

Raymond, M., and Rousset, F. (1995). GENEPOP on the Web, *J. Heredity*, 86:248-249.

Rudolph, J.A., Spier, S.J., Byrns, G. *et al.* (1992). *Nat. Genet.* 2(2):144-147.

Santschi, EM., Purdy, AK., Valberg, SJ. et al. (1998). *Mamm. Genome* 9:306-309.

Schonthaler, H.B., Lampert, J.M., von Lintig, J. (2005). *Dev. Biol.*, 284:421-36.

Shin, E.K., Perryman, L.E., Meek, K. (1997). *J. of Imm.* 158:3565-3569.

Wade, C.M., Giulotto, E., Sigurdsson, S. (2009). *Science*, 326:865-867.