

# Proposal to sequence the genome of the domestic horse, *Equus caballus*

Ernest Bailey (University of Kentucky), Doug Antczak (Cornell University), David Adelson (Texas A&M University), Matthew Binns (Royal Veterinary College, London), Bhanu Chowdhary (Texas A&M University), Gérard Guérin (INRA, France), Telhisa Hasagawa (JRA, Japan), Teri Lear (University of Kentucky), Tosso Leeb (School of Veterinary Medicine Hanover, Germany), James N. MacLeod (University of Kentucky), James Mickelson (University of Minnesota), James Murray (University of California, Davis), Cecilia Penedo (University of California, Davis), Loren Skow (Texas A&M University), Stephanie Valberg (University of Minnesota), Mark Vaudin (Animal Health Trust, UK).

## I. Introduction and Rationale for the Project

We are applying to National Human Genome Research Institute (NHGRI) for funding to produce a draft-2X reference sequence for the genome of the horse that can be integrated with physical and genetic maps. A whole genome shotgun approach is ideal for this approximately 2.5-3.0 billion base pair genome since preliminary data suggests that the genome is not highly repetitive and that most repeats are sufficiently diverged to ensure a random representation and low level assembly of the whole genome with this approach. The fosmid library produced from the reference horse in connection with the shotgun sequencing procedure will also be a valuable tool for the applications described below. Another important part of the project should be discovery of single nucleotide polymorphisms (SNPs) generated by conducting further low coverage sequencing of horses from diverse breeds. These SNPs will be essential for investigations using haplotypes in horse populations with well-recorded, deep pedigrees and other quantitative trait loci (QTL) studies.

Main points supporting inclusion of the horse among NHGRI-supported mammals with low-coverage full genome sequencing include:

1. The horse is proposed as the representative species of the order Perissodactyla. No other species from this order have yet been sequenced. The order includes both domesticated and wild species, some of which are highly endangered. Investigations of the horse genera have been especially important in evolutionary biology, particularly for the well-preserved fossil record and the karyotypic evidence of rapid chromosomal rearrangements for the genus *Equus* (MacFadden, 2005).
2. A distinguishing characteristic of the horse compared to other reference animals is its extensive breeding records. Horse breeders have maintained meticulous pedigree records for many horse breeds, some extending back over 300 years. Deep pedigree records coupled with excellent health and performance records provides excellent material for (QTL) studies. The cooperation of horse breeders with participants in the horse genome workshop has been excellent.
3. An organized international community of basic research scientists and veterinary clinicians work in the area of horse genetics. Scientists from over 25 laboratories meet regularly and have been collaborating on equine genome definition and application for the past decade. The major forums for meetings have been in connection with the USDA-NRSP8 program, the International Society for Animal Genetics and the Dorothy Russell Havemeyer Foundation Workshops on horse genomics. More than 150 multi-institutional publications have resulted from these collaborations.
4. Comparison with other mammalian genomes will contribute significantly to the NHGRI ENCODE project to develop an encyclopedia of DNA elements in the human genome. This especially includes the non-coding functionally important regions.
5. Tools developed during the last 10 years include synteny, linkage, cytogenetic, RH, comparative and integrated maps, 3 BAC libraries, collections of EST data and several databases and websites for sharing information (<http://www.uky.edu/AG/horsemap/>). By the end of 2006, over 4000 genes and markers should be mapped using RH and linkage data. Funding has just been approved for BAC fingerprinting and BAC end sequencing for 150,000 clones of the CHORI-241 library (Leeb,

- personal communication). Applications to the USDA Genomic Tools Program for generation of an additional 360,000 ESTs and an oligoarray for functional genomics research are under review.
6. Several aspects of the physiology and pathology of the horse are highly relevant to human biology and medicine. These include examples from immunology (Severe Combined Immunodeficiency Disease, a fatal genetic disease of Arabian horses), virology (Equine Infectious Anemia retrovirus), zoonoses (viral encephalitides such as West Nile Virus and others), neurology (Cerebellar Abiotrophy and others), muscle disorders (glycogen storage diseases, periodic paralysis and others), musculoskeletal diseases, connective tissue/skin disorders (Junctional Epidermolysis Bullosa and others), reproduction, orthopedics, and respiratory biology. A full genome sequence would support these and other equine research studies with direct application to human health.
  7. The equine genome sequence would also be applied in studies of equine disease and physiology for the benefit of the horse. Gene discovery and quantitative trait loci analyses are paramount in this endeavor. This would include investigations of simple inherited autosomal and sex linked conditions and complex diseases that have both genetic and environmental components. In addition, the sequence data would be used in the development of expression microarrays for functional genomic studies.
  8. Finally, the applications of the horse genome sequence would benefit the equine industry, which has a yearly economic impact of \$102 billion Gross Domestic Product (GDP) and 1.4 million Full Time Employees (FTE) in the United States alone. The population size of domestic horses is estimated at 9.2 million in the United States and 115 million worldwide.

### **Domestication and Uses of Horses**

Horses were domesticated approximately 4000-6000 years ago in Eurasia and provided much of the power and transportation used in agriculture and industry until replaced by the internal combustion engine during the 20<sup>th</sup> Century. Since then, the horse continues to be used for sports and recreation and the horse industry remains important in the United States and around the world. A 2005 study by the American Horse Council (AHC, 2005) reported the annual direct, indirect and induced economic impact of the horse industry on the US economy at \$102 billion GDP and 1.4 million FTE. The horse remains economically important even though its role in human society has changed over time.

Today, horses are selected for racing performance, riding characteristic, pulling ability, size, behavior, intelligence, color and diverse, unique other traits, e.g., cow-cutting ability, sprinting ability and special gaits. Longevity and health are also important characteristics because horses are often trained for several years before achieving their potential use. Horses suffer from a number of Mendelian genetic diseases although the true extent of such defects is not yet known. Several known disorders in horses are homologous to human conditions while others, such as laminitis (inflammatory disease of the hoof) have no clear human homologue. Horses appear to have a lower incidence of cancer than dogs, cats and humans. Horses also suffer complex allergic and developmental diseases that become apparent with use and age. Advances in health monitoring and health care of horses would provide social and economic benefits and will likely shed new light on mechanisms of human disease.

Elite horses are also among the most widely traveled domestic animals. Horses used for dressage or event-riding competitions, as for the Olympics, are competitive in their teen years and travel between Europe, America and Asia. Likewise, Thoroughbred, Quarter Horses and other racehorses travel abroad for competition or for use in breeding programs. One consequence of this international traffic is the importance of understanding the possibility of epidemic infectious diseases among horses and being able to develop effective vaccines, therapies and preventative strategies. Most infectious diseases of horses do not affect people; however there are several viral encephalopathies (Western Equine Encephalitis, Eastern Equine Encephalitis, and Venezuelan Equine Encephalitis) and West Nile virus infections that can be fatal in people. Research on these diseases is important to prevent human disease.

## Systematics of the genus *Equus*

The horse (*Equus caballus*) is a member of the order Perissodactyla of odd toed mammals made up by the three families: Equidae (two horse species, three hemione species, two donkey species and at least three zebra species), Rhinocerotidae (five species) and Tapiridae (four species). Among the Perissodactyla, only three species have been domesticated: the horse (*Equus caballus*) and the two donkeys (*Equus asinus* and *Equus africanus somalicus*). Hybrids have been produced for most species pairs among the Equidae, however the hybrids are almost always infertile. The best known hybrid is the cross of the horse mare with a donkey jack to produce a mule. Mules were prized for their intelligence and vigor and played a major role in agriculture and industry. The reason for the infertility is probably differences in genome organization between the species of Equidae and failure of meiosis within hybrid animals. Chromosome numbers range from 33 pairs for *Equus przewalskii* (Mongolian Wild Horse) to 16 pairs for the *Equus zebra hartmannae* (Hartmann's zebra, a.k.a., Mountain zebra). These species diverged 3-5 million years ago, so chromosome evolution has been relatively rapid for this family. Comparative mapping studies indicate that chromosome evolution among the Equidae has probably been the result of chromosome fusion and centromere repositioning (Lear, 2005; Guilotto, unpublished)

## Applications

Gene maps and comparative approaches have been used to assign several coat color traits and inherited disorders to horse chromosomes. In some cases, this has led to the identification of the responsible gene and the causal mutation, while in others it has helped to identify markers tightly linked to the trait. We are aware of at least 6 other disorders currently being analyzed by researchers in the US (University of Kentucky, UC Davis, University of Minnesota), Japan, France and Switzerland.

<i>Trait/condition</i>	<i>Marker</i>	<i>Chr.</i>	<i>Reference</i>
Hyperkalemic periodic paralysis- <b>HYPP</b>	SCN4A	ECA11p	Rudolph et al., 1992.
Severe comb. Immunodef. <b>SCID</b>	– PRKDC	ECA9p	Bailey et al. 1997; Shin et al., 1997.
Overo lethal white foal syndrome <b>OLWS</b>	EDNRB	ECA17q	Metallinos et al., 1998; Santschi et al., 1998; Yang et al., 1998.
Herlitz Jnct. Epider. bullosa, <b>H-JEB</b>	LAMC2	ECA5p	Spirito et al., 2002; Milenkovic et al. 2003
Glycogen storage disease IV, <b>GBED</b>	GBE1	ECA26q	Ward et al., 2003; Ward et al., 2004.
Extension (red/black coat color)	MC1R	ECA3p	Marklund et al., 1996.
Cream dilution coat color	MATP	ECA21q	Locke et al., 2001; Mariat et al., 2003
Gray coat color	G	ECA25q	Henner et al., 2002; Locke et al., 2002; Swinburne et al., 2002
Tobiano (white spotting)	KIT?	ECA3q	Brooks et al., 2002
Agouti (black pigment distribution)	ASIP	ECA22q	Rieder et al., 2001
Appaloosa (white spotting)	TRPM1	ECA1q	Terry et al., 2004
Dominant white	WNT1	ECA6	Mau et al., 2004.

The horse is a monogastric animal that mimics human biology to a large extent and that has over 80 hereditary conditions in common with humans. A number of equine conditions hitherto studied share common genes/mutations with those studied in humans. This makes an important and compelling case to obtain detailed genome information for the horse, in order to develop an improved understanding of a range of conditions for which little is known in humans (allergies, inflammatory processes, infectious diseases, etc.)

## **II. Specific biological/biomedical rationales for utility of the sequence data**

### *1. Improving human health by increasing understanding of disease or relevance to development of methods for diagnosis, treatment or prevention of human disease.*

The genetic basis for diseases in horses are very similar to those found in people. To date, the mutations associated with five major breed-specific Mendelian disorders, OLWS (Santschi et al., 1998), SCID (Shin et al., 1997), JEB (Spirito et al., 2002), GBED (Ward et al., 2003), and HYPP (Rudolph et al., 1992), have been defined through candidate gene approaches and diagnostic DNA tests developed to control their propagation. The interest in conducting QTL studies on complex traits is very high among scientists and breeders. Although complex traits such as chronic obstructive pulmonary disease, osteochondrosis dissecans, and susceptibility and resistance to infectious disease are predicted to have a genetic component, very little is known of the heritability or nature of the contributing genes. A similar situation exists in the areas of equine athletic performance, conformation, and behavior.

### *2. Informing Human Biology: How this will lead to better understanding of biological function in humans.*

Examples: Muscle diseases in horses are particularly apparent because they are usually not afforded the opportunity to be sedentary. A number of exertional myopathies that appear to be inherited in a Mendelian fashion are now known in horses. These include Recurrent Exertional Rhabdomyolysis in Thoroughbred racehorses and that affects approximately 5 – 10 % of the population (MacLeay et al., 1999) and Polysaccharide Storage Myopathies in Quarter Horses and Draft Horse breeds that may affect more than 5 and 25% of the population, respectively (Valberg et al., 1996; Firshman, 2005). Although the molecular basis is still unknown, the RER defect lies in some aspect of intracellular calcium regulation and muscle contractility (Lentz et al. 1999) and the defect for PSSM lies in some aspect of the regulation of glucose homeostasis and insulin sensitivity (Annandale et al., 2004). Determining the genes responsible for these conditions is one way in which insights into equine pathobiology would shed new light on human myopathies and even diabetes.

### *3. Expanding our understanding of biological processes relevant to human health.*

In a manner similar to other domestic species, horses have been selectively bred for hundreds of years with careful maintenance of pedigree data. Unlike other domestic species, however, selective breeding practices in horses have been based primarily on athletic performance traits. As such, horses provide an opportunity to investigate genetic determinants related to exercise physiology and diseases that compromise athletic performance. Genome sequence data and informative polymorphisms will permit the identification and characterization of quantitative trait loci for disease traits of the musculoskeletal, neuromuscular, cardiovascular, and respiratory system that appear very similar in clinical presentation and pathogenesis to diseases in humans. Specific examples include osteochondrosis dissecans, angular limb malformation, trauma induced osteoarthritis, tendon and ligament injury, stenosis and other malformations of the vertebral canal, exercise associated cardiac arrhythmias and conduction abnormalities, and exercise induced pulmonary hemorrhage.

Horses represent a superior mammalian model to define genetic determinants related to the etiology and predisposition of diseases associated with exercise and athletic performance, along with the opportunity to study the molecular pathogenesis of the condition and molecular mechanism related to both therapy and healing.

### *4. Providing additional surrogate systems for human experimentation; new disease models, etc.*

As noted above, many of the hereditary diseases of horses have human homologues making it an appropriate model for studying diseases relevant to human health. Furthermore, horses are large mammals and in some cases may be most appropriate subjects for some studies. As a case in point, one member of the horse genome workshop is investigating chondrocyte biology in horses. When chondrocytes are cultured, they change their physiological characteristics. Therefore primary cultures are necessary to investigate some questions. It is not possible to obtain sufficient chondrocytes from

humans for doing primary studies. However, it is possible from horses. It is also possible to take relatively large and multiple muscle biopsies for analysis from horses due to their large muscle mass.

The NIH has supported a project on equine pregnancy immunology (Antczak, Cornell) since the early 1980s because of its relevance to this aspect of human reproduction.

#### 5. *Facilitating ability to do experiments by direct experiments in additional organisms.*

As evidenced by the recent sequencing of the canine genome, ready access to contiged and annotated equine genome sequence would enable a huge step forward in the ability to develop and study heritable conditions in horses and make them relevant spontaneous models of human disease. For example, polymorphic simple tandem repeat markers from any gene or location of interest would be immediately available to test in case-control genetic association studies in populations.

Experiments are underway on a wide range of health traits in horses that also have human homologues. However, these experiments are hampered by the lack of basic genomics tools. Many scientists and their students begin a project on a functional trait by cloning and sequencing relevant genes. Considerable time and funds are expended before the fundamental question is addressed. If whole genome sequence information were available scientists would be better able to do their research, use fewer horses and possibly do many of the experiments *in silico*.

### **III. Strategic Issues in acquiring new DNA sequence data**

#### 1. *Demand for data; size of research community, enthusiasm and will this stimulate the expansion of the community.*

The Dorothy Russell Havemeyer Foundation convened the first Equine Gene Mapping Workshop in October 1995. Over 70 scientists from 20 countries participated in that meeting. In 1996 the USDA-National Animal Genome Program, under the auspices of the NRSP8 program invited the new horse gene mapping community to join their program, underway since 1992. The horse technical committee has been a funded part of the USDA-NRSP8 since 1998 and meets every January at the Plant and Animal Genome Conference in San Diego. These scientists have also conducted and participated in workshops on horse genomics at every Conference of the International Society of Animal Genetics since 1996. The Havemeyer Foundation has continued to convene workshop meetings, now called The Equine Genome Workshop to reflect the broader interests of the participating scientists. Currently, more than 200 scientists from more than 25 countries participate in the workshop activities. The next Havemeyer Workshop will be held from July 12-14, 2005 in Dublin, Ireland. The agenda includes gene mapping, functional genomics and whole genome sequencing. This agenda was set during discussions at the NRSP8 meeting at PAG in January 2005. Leadership at the meetings has rotated among laboratories. Currently Matthew Binns (UK) is chair of the USDA-NRSP8 committee, Telhisa Hasegawa (Japan) is chair of the ISAG workshop and John Flynn (Ireland) is the chair of the Havemeyer Workshop meeting in 2005.

Veterinary scientists who have not participated directly in the workshop have been keenly interested in the development of horse genomic tools for their applications. We anticipate that improvement of the number and quality of tools will accelerate research and, concomitantly, improve health care for horses. Conversely, if we do not develop these tools, many bright scientists will not find research on horses attractive and will work on species in which more fundamental experiments can be conducted.

#### 2. *Suitability of the organism for experimentation:*

There are over 9 million horses in the United States, alone (AHC, 2005). Most of those horses are domestic animals in private ownership. Their health is usually under close scrutiny. Horse pedigrees are important records and often determine the value of a horse. Furthermore, horses are domestic animals, quite tame, readily managed and various types of samples can be readily collected to assay a wide range of biological parameters for experimentation or study.

Although full-sibling families of horses are rare, families with many hundreds to over a thousand half-siblings are observed, and result from the practice of repeated breeding popular stallions and the propagation of desirable sire lines. This often allows for the design of gene mapping

experiments with excellent statistical power. However, it is also the case that breeders may be reluctant to participate in genetic studies that may cast doubt on the health of valuable animals. Extensive pedigree records for most breeds will also facilitate whole genome genetic association approaches that do not rely on collection of large closely-related kindreds.

3. *Rationale for complete sequence of the organism. Why would this be better than other approaches... extensive ESTs, etc.*

Genomics tools/resources for horses are modest by comparison to human or mouse. As a result, it is tempting to suggest improving linkage maps or constructing a physical map of the horse genome using BAC libraries. These resources would be very useful to the equine genome research community, particularly with respect to mapping QTL and positionally cloning loci. However, despite considerable agreement in marker positioning and order between maps, significant differences do exist which are unlikely to be resolved solely by the addition of more markers.

Low level sequencing of the equine genome will probably deliver more information than increasing map density/resolution or building a better physical map. A 2X sequence of the equine genome could be aligned to the human and cattle genomes to generate very useful proxy genome assemblies for the equine community, thus facilitating positional cloning. In addition, biomedical researchers would be the real winners, by being able to glean relevant data regarding loci of importance to disease resistance and exercise physiology.

4. *Cost of sequencing and the state of readiness of the organism's DNA for sequencing.*

Considerable work has been done on the linear and physical map for the horse. Extensive comparative mapping information exists based on RH mapping and cytogenetic mapping. In the near future we expect to have BAC fingerprinting and BAC end sequencing completed for 150,000 clones from the CHORI-241 library. Relating a whole genome sequence to the human sequence would be of great interest to the horse genomics research community and the tools exist to do this work.

Whole genome sequencing using the shotgun genome approach would benefit from using females with limited genetic variation. As with cats and dogs, many horse breeds were established from small numbers of foundation animals. Based on studies of biochemical and molecular markers we know that one of the horse breeds with the least genetic variability is the Thoroughbred horse (Cothran, personal communication). That the Thoroughbred is among the most highly inbred, valuable, and recognizable horse breeds in the world, makes it an ideal choice for generating the whole genome sequence.

Based on this premise, several appropriate horses are available for use in whole genome shotgun studies. The CHORI-241 BAC library was made from a male thoroughbred horse that was the product of a full-sibling mating at Cornell University. There are additional female thoroughbred horses available at Cornell University that are also the product of sibling matings. Furthermore, at the University of Kentucky there is one female thoroughbred horse that is the product of a father daughter mating.

#### STATE OF THE MAP AND RELEVANT RESOURCES

**Chromosomes** 2N= 64, XY

**Genome Size:** ~3,000 Mb, by reference to other mammalian species; preliminary evidence from BAC end sequencing suggests the genome size may only be 2,500 Mb (Leeb, personal communication).

**BAC Libraries:** A total of 3 equine BAC libraries are presently available. These libraries have been extensively used by researchers worldwide.

**INRA:** The INRA-LGCB BAC library was obtained from a fibroblast cell culture and from leukocytes of one stallion. Partial HindIII digest, vector pBeloBac11. The coverage of this library is 3.4X with 110 000 BACs with a mean size of 100Kb (Godard et al., 1998; Milenkovic et al., 2002). The library is distributed in 96 and 389 well plates and pooled in a three dimensions protocol for PCR screening. It is fully available for screening. See: <http://www-crb.jouy.inra.fr/bacyac.htm>

**TAMU:** The TAMU BAC library was constructed by Dr. Loren Skow. The library consists of ~39,000 clones in pBELOBAC11 from gelded Quarter Horse. Average size is 110 kb for approximately 2X genome coverage. DNAs are arrayed in hierarchical pools for PCR screening.

**CHORI-241:** The CHORI-241 BAC library has been constructed by Dr. Baoli Zhu in BACPAC. The DNA was digested with EcoRI and cloned in pTARBAC2.1. The library consists of 190,652 clones with an average insert size of 171 Kb. See: <http://bacpac.chori.org/>

**Synteny map:** The UC Davis horse x hamster somatic cell hybrid panel (Shiue et al. 1999) was instrumental in mapping ~450 markers and assigning syntenic groups to all equine chromosomes (Caetano et al. 1999; Shiue et al. 1999; Lindgren et al., 2001). The resource is almost depleted, and replaced by new, rapid, and more efficient techniques.

**Linkage map:** Three reference family resources, viz., the Uppsala half-sib family (Lindgren et al. 1998), the International Horse Reference Family Panel (IHRFP; Guérin et al. 1999; Guérin et al. 2003) and the Animal Health Trust (AHT) 3-generation full-sib family (Swinburne et al. 2000), have played a key role in developing linkage maps. The most recent linkage maps (Penedo et al. 2005; Swinburne et al. 2005) show a total of ~800 markers assembled in single linkage groups mapped to individual autosomes and the X chromosome. This map is currently being expanded to strengthen the current genome scanning panel for analysis of diseases and other traits of significance.

**Cytogenetic map:** The equine gene mapping community has published extensively to rapidly expand the cytogenetic map in the horse (e.g., see Breen et al. 1997; Godard et al. 2000; Lear et al. 2001; Lindgren et al., 2001; Mariat et al. 2001; Raudsepp et al. 2001; Milenkovic et al. 2002; Chowdhary et al. 2002; Raudsepp et al. 2002; Chowdhary et al. 2003; Lee et al. 2004; Raudsepp et al. 2004; Raudsepp et al. 2004; Brinkmeyer-Langford et al. 2005; Gustafson-Seabury et al. 2005). By 2006, a cytogenetic map with ~1000 markers will be available.

**RH & comparative map:** A 5000 rad horse x hamster RH panel was generated at Texas A&M University. The first generation RH and comparative map of the horse genome comprised 730 markers (258 type I and 472 type II; Chowdhary et al. 2003). The map is the first comprehensive framework in the horse that i) incorporates type I as well as type II markers, ii) integrates syntenic, cytogenetic and meiotic maps into a consensus map and iii) provides the most detailed genome-wide information to date on the organization and comparative status of the equine genome. Since then high-resolution maps for some of the horse chromosomes have been generated. These include ECA17 (Lee et al. 2003), ECAX (Raudsepp et al. 2002; Raudsepp et al. 2004), ECAY (Raudsepp et al. 2004), ECA22 (Gustafson-Seabury et al. 2005), parts of ECA7, 10 & 21 (homologues of HSA19; Brinkmeyer-Langford et al. 2005), ECA15 & 18 (homologues of HSA2; Wagner et al., *submitted*) and ECA14 & 21 (homologues of HSA5; Goh et al. *in preparation*). Additionally, analysis is being finalized for a number of other equine chromosomes, viz., ECA4, ECA10q, ECA16, ECA19, ECA29 and ECA26. The equine genomics community anticipates to report a 4000 marker (<1Mb resolution) second generation physical map of the equine genome by early 2006.

**Integrated Map:** Bernoco et al. 2005 ( Havemeyer Workshop, Ireland) integrated 660 markers from the linkage and 685 markers from the RH autosomal map in the first run of this analysis using the “defalco” option of the CarthaGène program (<http://www.inra.fr/bia/T/CartaGene/>). Altogether 1022 unique markers were included in the analysis, of which 323 (31.60 %; range: 7.89 % for ECA7 to 50.00 % for ECA30) were shared between the two data sets, 337 were unique for the linkage data and 362 were unique for the radiation hybrid data. The total number of unique markers integrated for each chromosome varied from 12 for ECA30 to 94 for ECA01. The total integrated autosomal linkage map span was 3536 cM. The average spacing between markers was estimated at 5.6 cM. Overall, the agreement between the two sets of data in the orientation and linear order of the markers is excellent.

**EST:** More ESTs are certainly needed and would be essential for annotating any low level sequencing effort. Gene prediction methods will not work well with the large number of small contigs expected from 2x WGS coverage. Therefore, any useful transcript annotation would have to come from EST alignments to the WGS sequence. This, in combination with comparative alignment to human, cattle,

dog and pig sequences will provide ample resources for positional cloning, particularly of candidate genes.

Collection of equine ESTs: Presently 16 cDNA libraries exist for the horse (see below). In total, there are material for ~64,300 clones.

TISSUE	Tissue details	Seq. clones	Reference
Blood	blood cells from a septic female	114	Pascual et al. 2002
	unstimulated peripheral blood leukocytes N6	947	Vandenplas et al. unpublished.
	monocytes	2341	Vandenplas et al. unpublished
Brain	peripheral blood lymphocytes	unknown	Binns et al. 1994
	fetal brain	7	Godard et al. 2000
Cartilage	articular cartilage	13,963	MacLeod unpublished
	embryonic stem cells	15	Saito et al. 2002
Embryo	foetal cDNA library directionally cloned	57	Brandon et al. unpublished
Liver	liver	613	Vandenplas et al. unpublished
Lymph	mesenteric lymph node	11,453	Watson et al. unpublished
Muscle	skeletal muscle	17,580	Chowdhary et al. unpublished
Ovary	preovulatory follicle	HSA2 genes	Stock et al. 2002
Placenta	placenta	3,622	Antczak 2005 unpublished
	skin	313	Lieto & Cothran 2001
Skin	intact skin subtract from wound margin	231	Lefebvre-Lavoie et al. 2005,
Testes	testes	13,000	Skow et al. unpublished

Recently, a total of 42,450 equine ESTs were analyzed using Phred, Phrap and Consed (Ewing and Green, 1998 ; Ewing et al. 1998; Gordon et al., 1998) to call bases, mask vector and known repeats, and cluster reads into contigs based on overlaps. Optimized settings for EST data were used for this analysis (Adelson et al., 2004). After clustering, the input data yielded 12,732 singlets and 7648 contigs. The singlets and contigs were then blasted against the human refseq proteins (28,707 sequences on Feb. 14 2005) using BLASTX. This provided 5,980 non-redundant hits (nr hits; e-value < 0.01), of which ~5,500 had e-values < 1e-10. Further, of the 2276 contigs with no hit at all, 1899 had three or more traces. Thus, on the basis of clustering sequence overlaps and identical BLAST hits, 5,980 + 1899 = 7879 unique gene candidates were identified for microarray development.

Dissemination of Data: Results are distributed via GenBank, ARKDB Horse Genome Database (<http://www.thearkdb.org/>), the INRA Horsemap Database (<http://dga.jouy.inra.fr/cgi-bin/lgbc/main.pl?BASE=horse>) and the Horse Genome Workshop website (<http://www.uky.edu/AG/Horsemap/>). We anticipate the results from whole genome sequencing will also be available on the website at the UCSC Genome Browser.

##### 5. *Are there other sources of funding available or being sought.*

During the past 10 years horse genomics research has enjoyed support from the Morris Animal Foundation, Grayson-Jockey Club, AQHA Research Foundation, Dorothy Russell Havemeyer Foundation, the USDA-NRSP8, USDA-NRI and numerous additional but smaller foundations and organizations interested in different aspects of the horse genome work. The costs of conducting whole genome sequencing exceed the more modest funding abilities of these organizations so additional fund raising activities are necessary and new donors must be found. At the July 12-14, 2005 Havemeyer Workshop meeting, Kellye Eversole, leader of whole genome sequencing efforts for other species, will discuss a plan for seeking funding and development of an additional 6X genome coverage for the horse.

A grant of 1.54 million EUR (1.83 million US\$) for the development of the equine physical map (BAC fingerprinting, BAC end sequencing) was approved by the German Volkswagen Foundation. It is expected that this grant will be officially allocated within the next month (Leeb, personal communication).

We enjoy wide industry support for these goals. Representative letters have been obtained from horse industry leaders representing The Jockey Club, American Quarter Horse Association, Morris

Animal Foundation, Gluck Equine Research Foundation, Les Haras Nationaux, The Horse Racing Levy Board and the German Equine Federation. The letters are included in an appendix to this proposal.

### **Bibliography**

- AHC, (2005) American Horse Council Study: *Economic Impact of the Horse Industry in the United States*, Deloitte Consulting, LCC, <http://www.HorseCouncil.org>.
- Adelson DL, Cam GR, DeSilva U, Franklin IR (2004) Gene expression in sheep skin and wool (hair). *Genomics* 83, 95-105
- Annandale EJ, Valberg SJ, Mickelson JR, Seaquist ER. (2004). Insulin sensitivity and skeletal muscle glucose transport in equine polysaccharide storage myopathy. *Neuromusc Disorders* 14: 666-674.
- Bailey E, Reid RC, Skow LC, Mathiason K, Lear TL, McGuire TC (1997) Linkage of the gene for equine combined immunodeficiency disease to microsatellite markers HTG8 and HTG4; Synteny and FISH mapping to ECA9. *Animal Genetics* 28: 268-273.
- Bernoco D, Chowdhary BP, Millon L, Bailey E (2005) First integration of equine linkage and radiation hybrid autosomal mapping data. *Proceedings of 6<sup>th</sup> Havemeyer Horse Genome Workshop, Ireland*, July 13, 2005.
- Binns M.M., Smith K., McManus A., Tavernor A., Butcher G., Mumford J.A. (1994). Identification of equine genes by characterization of clones from an equine peripheral blood lymphocyte cDNA library. In: *Equine Infectious Diseases VII. Proceedings of the Seventh International Conference*, Tokyo, June 8-11, 1994. Eds. H. Nakajuma and W. Plowright. Newmarket, R & W Publications, 1994, pp. 153-156.
- Breen M, Lindgren G, Binns MM, Norman J, Irvin Z, et al. (1997) Genetical and physical assignments of equine microsatellites--first integration of anchored markers in horse genome mapping. *Mamm Genome* 8, 267-273
- Brinkmeyer-Langford C, Raudsepp T, Lee EJ, Goh G, Schäffer AA, et al. (2005) A high-resolution physical map of equine homologues of HSA19 shows divergent evolution compared to other mammals. *Mamm Genome: in press*
- Brooks SA, Terry RB, Bailey E (2002) A PCR-RFLP for KIT associated with tobiano spotting pattern in horses. *Anim Genet* 33, 301-303
- Caetano AR, Shiue YL, Lyons LA, O'Brien SJ, Laughlin TF, et al. (1999) A comparative gene map of the horse (*Equus caballus*). *Genome Res* 9, 1239-1249
- Chowdhary BP, Raudsepp T, Honeycutt D, Owens EK, Piumi F, et al. (2002). Construction of a 5000<sub>rad</sub> whole-genome radiation hybrid panel in the horse and generation of a comprehensive and comparative map for ECA11. *Mamm Genome* 13: 89-94.
- Chowdhary BP, Raudsepp T, Kata SR, Goh G, Millon LV, et al. (2003) The first-generation whole-genome radiation hybrid map in the horse identifies conserved segments in human and mouse genomes. *Genome Res* 13, 742-751
- Ewing B and Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8, 186-194
- Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8, 175-185
- Firshman AM, Baird JD, and Valberg SJ. (2005). PhD thesis, University of Minnesota.
- Godard S, Vaiman A, Schibler L, Mariat D, Vaiman D, et al. (2000) Cytogenetic localization of 44 new coding sequences in the horse. *Mamm Genome* 11, 1093-1097
- Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8, 195-202
- Green P (1997) Against a whole-genome shotgun. *Genome Res* 7, 410-417
- Guerin G, Bailey E, Bernoco D, Anderson I, Antczak DF, et al. (1999) Report of the International Equine Gene Mapping Workshop: male linkage map. *Anim Genet* 30, 341-354

- Guerin G, Bailey E, Bernoco D, Anderson I, Antczak DF, et al. (2003) The second generation of the International Equine Gene Mapping Workshop half-sibling linkage map. *Anim Genet* 34, 161-168
- Gustafson-Seabury A, Raudsepp T, Goh G, Kata SR, Wagner ML, et al. (2005) High-resolution RH map of horse chromosome 22 reveals a putative ancestral vertebrate chromosome. *Genomics* 85, 188-200
- Henner J, Poncet PA, Guerin G, Hagger C, Stranzinger G, et al. (2002) Genetic mapping of the (G)-locus, responsible for the coat color phenotype "progressive greying with age" in horses (*Equus caballus*). *Mamm Genome* 13, 535-537
- Lear TL, Brandon R, Piumi F, Terry RR, Guerin G, et al. (2001) Mapping of 31 horse genes in BACs by FISH. *Chromosome Res* 9, 261-262
- Lear TL (2005) Using horse gene mapping tools to study chromosome evolution in Perissodactyls. *Proceedings of Plant and Animal Genome XIII*, W086, p23.
- Lee EJ, Raudsepp T, Kata SR, Adelson D, Womack JE, et al. (2004) A 1.4-Mb interval RH map of horse chromosome 17 provides detailed comparison with human and mouse homologues. *Genomics* 83, 203-215
- Lefebvre-Lavoie J, Lussier JG, Theoret CL (2005) Profiling of differentially expressed genes in wound margin biopsies of horses, using suppression subtractive hybridization. *Physiol. Genomics: in press*
- Lentz LR, Valberg SJ, Balog EM, Mickelson JR, and Gallant EM. (1999). Abnormal regulation of muscle contraction in equine recurrent exertional rhabdomyolysis. *Am J Vet. Res* 60: 992 - 999.
- Lieto LD and Cothran EG (2001) Characterization of expressed sequence tags generated from skin cDNA clones of *Equus caballus* by single pass sequencing. *Anim Biotechnol* 12, 87-97
- Lindgren G, Sandberg K, Persson H, Marklund S, Breen M, et al. (1998) A primary male autosomal linkage map of the horse genome. *Genome Res* 8, 951-966
- Lindgren G, Breen M, Godard S, Bowling A, Murray J, et al. (2001) Mapping of 13 horse genes by fluorescence in-situ hybridization (FISH) and somatic cell hybrid analysis. *Chromosome Res* 9, 53-59
- Locke MM, Ruth LS, Millon LV, Penedo MC, Murray JD, et al. (2001) The cream dilution gene, responsible for the palomino and buckskin coat colours, maps to horse chromosome 21. *Anim Genet* 32, 340-343
- Locke MM, Penedo MC, Bricker SJ, Millon LV, Murray JD (2002) Linkage of the grey coat colour locus to microsatellites on horse chromosome 25. *Anim Genet* 33, 329-337
- MacFadden BJ (2005) Evolution. Fossil Horses-Evidence for Evolution. *Sci.* 307:1728-1730.
- Macleay JM, Valberg SJ, Sorum SA, Sorum MD, Kassube T, Santschi EM, Mickelson JR, and Geyer CJ (1999). Heritability of recurrent exertional rhabdomyolysis in Thoroughbred racehorses. *Am J Vet Res* 60: 250-256.
- Mariat D, Oustry-Vaiman A, Cribiu EP, Raudsepp T, Chowdhary BP, et al. (2001) Isolation, characterization and FISH assignments of horse BAC clones containing type I and II markers. *Cytogenet Cell Genet* 92, 144-148
- Mariat D, Taourit S, Guerin G (2003) A mutation in the MATP gene causes the cream coat colour in the horse. *Genet Sel Evol* 35, 119-133
- Marklund L, Moller MJ, Sandberg K, Andersson L (1996) A missense mutation in the gene for melanocyte-stimulating hormone receptor (MC1R) is associated with the chestnut coat color in horses. *Mamm Genome* 7, 895-899
- Mau C, Poncet PA, Bucher B, Stranzinger G, Rieder S (2004) Genetic mapping of dominant white (*W*), a homozygous lethal condition in the horse (*Equus caballus*) *J Anim Breed Genet* 6: 374-383.
- Metallinos DL, Bowling AT, Rine J (1998) A missense mutation in the endothelin-B receptor gene is associated with Lethal White Foal Syndrome: an equine version of Hirschsprung disease. *Mamm Genome* 9, 426-431

- Milenkovic D, Oustry-Vaiman A, Lear TL, Billault A, Mariat D, et al. (2002) Cytogenetic localization of 136 genes in the horse: comparative mapping with the human genome. *Mamm Genome* 13, 524-534
- Milenkovic D, Chaffaux S, Taourit S, Guerin G (2003) A mutation in the LAMC2 gene causes the Herlitz junctional epidermolysis bullosa (H-JEB) in two French draft horse breeds. *Genet Sel Evol* 35, 249-256
- Pascual I, Dhar AK, Fan Y, Paradis MR, Arruga MV, et al. (2002) Isolation of expressed sequence tags from a Thoroughbred horse (*Equus caballus*) 5'-RACE cDNA library. *Anim Genet* 33, 231-232
- Penedo, MCT, Millon LV, Bernoco D, Bailey E, Binns MM, et al. (2005) International Equine Gene Mapping Workshop Report: A comprehensive linkage map constructed with data from new markers and by merging four mapping resources. *Cytogenet Genome Res: in press*
- Raudsepp T, Mariat D, Guerin G, Chowdhary BP (2001) Comparative FISH mapping of 32 loci reveals new homologous regions between donkey and horse karyotypes. *Cytogenet Cell Genet* 94, 180-185
- Raudsepp T, Kata SR, Piumi F, Swinburne J, Womack JE, et al. (2002a) Conservation of gene order between horse and human X chromosomes as evidenced through radiation hybrid mapping. *Genomics* 79, 451-457
- Raudsepp T, Lear TL, Chowdhary BP (2002b) Comparative mapping in equids: the asine X chromosome is rearranged compared to horse and Hartmann's mountain zebra. *Cytogenet Genome Res* 96, 206-209
- Raudsepp T, Lee E-J, Kata SR, Brinkmeyer C, Mickelson JR, et al. (2004a) Exceptional conservation of horse-human gene order on X chromosome revealed by high-resolution radiation hybrid mapping. *Proc Natl Acad Sci U S A* 101: 2386-2391.
- Raudsepp T, Santani A, Wallner B, Kata SR, Ren C, et al. (2004b) A detailed physical map of the horse Y chromosome. *Proc Natl Acad Sci U S A* 101, 9321-9326
- Rieder S, Taourit S, Mariat D, Langlois B, Guerin 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*). *Mamm Genome* 12, 450-455
- Rudolph JA, Spier SJ, Byrns G, Rojas CV, Bernoco D, et al. (1992) Periodic paralysis in quarter horses: a sodium channel mutation disseminated by selective breeding. *Nat Genet* 2, 144-147
- Saito S, Ugai H, Sawai K, Yamamoto Y, Minamihashi A, et al. (2002) Isolation of embryonic stem-like cells from equine blastocysts and their differentiation in vitro. *FEBS Lett* 531, 389-396
- Santschi EM, Purdy AK, Valberg SJ, Vrotsos PD, Kaese H, et al. (1998) Endothelin receptor B polymorphism associated with lethal white foal syndrome in horses. *Mamm Genome* 9, 306-309
- Shin EK, Perryman LE, Meek K (1997) A kinase-negative mutation of DNA-PK(CS) in equine SCID results in defective coding and signal joint formation. *J Immunol* 158, 3565-3569
- Shiue YL, Bickel LA, Caetano AR, Millon LV, Clark RS, et al. (1999) A synteny map of the horse genome comprised of 240 microsatellite and RAPD markers. *Anim Genet* 30, 1-9
- Spirito F, Charlesworth A, Linder K, Ortonne JP, Baird J, et al. (2002) Animal models for skin blistering conditions: absence of laminin 5 causes hereditary junctional mechanobullous disease in the Belgian horse. *J Invest Dermatol* 119, 684-691
- Stock AE, Bouchard N, Brown K, Spicer AP, Underhill CB, et al. (2002) Induction of hyaluronan synthase 2 by human chorionic gonadotropin in mural granulosa cells of equine preovulatory follicles. *Endocrinology* 143, 4375-4384
- Swinburne J, Gerstenberg C, Breen M, Aldridge V, Lockhart L, et al. (2000) First comprehensive low-density horse linkage map based on two 3-generation, full-sibling, cross-bred horse reference families. *Genomics* 66, 123-134
- Swinburne JE, Hopkins A, Binns MM (2002) Assignment of the horse grey coat colour gene to ECA25 using whole genome scanning. *Anim Genet* 33, 338-342

- Swinburne JE, Bournnell M, Hill G, Pettitt L, Allen T, et al. (2005) Single linkage group per chromosome genetic linkage map for the horse, based on two, three-generation, full- sibling, cross-bred horse reference families. *Genomics: in press*
- Terry RB, Archer S, Brooks S, Bernoco D, Bailey E (2004) Assignment of the appaloosa coat colour gene (LP) to equine chromosome 1. *Anim Genet* 35, 134-137
- Valberg SJ, Geyer C, Sorum SA, Cardinet GH 3rd. (1996). Familial basis of exertional rhabdomyolysis in quarter horse-related breeds. *Am J Vet Res* 57: 286-90.
- Wagner B, Miller DC, Lear TL, Antczak DF (2004) The complete map of the Ig heavy chain constant gene region reveals evidence for seven IgG isotypes and for IgD in the horse. *J Immunol* 173, 3230-3242
- Wagner B, Robeson J, McCracken M, Wattrang E, Antczak DF (2005) Horse cytokine/IgG fusion proteins--mammalian expression of biologically active cytokines and a system to verify antibody specificity to equine cytokines. *Vet Immunol Immunopathol* 105, 1-14
- Wagner ML, Raudsepp T, Goh G, Agarwala R, Schäffer AA, et al. (2005) A 1.3-Mb interval map of equine homologs of HSA2. *Cytogenetic and Genome Research*, submitted.
- Ward TL, Valberg SJ, Lear T, Guerin G, Milenkovic D, Swinburne J, Binns MM, Chowdhary BP, Raudsepp R, Skow S and Mickelson JR. (2003). Genetic mapping of GBE1 and its association with glycogen storage disease IV in American Quarter Horses. *Cytogenetics and Genome Research* 102: 201-206.
- Ward TL, Valberg SJ, Lear TL, Guerin G, Milenkovic D, et al. (2003) Genetic mapping of GBE1 and its association with glycogen storage disease IV in American Quarter horses. *Cytogenet Genome Res* 102, 201-206
- Ward TL, Valberg SJ, Adelson DL, Abbey CA, Binns MM, Mickelson JR. (2004). Glycogen branching enzyme (GBE1) mutation causing equine glycogen storage disease IV. *Mamm Genome*. 15: 570-7.
- Yang GC, Croaker D, Zhang AL, Manglick P, Cartmill T, et al. (1998) A dinucleotide mutation in the endothelin-B receptor gene is associated with lethal white foal syndrome (LWFS); a horse variant of Hirschsprung disease. *Hum Mol Genet* 7, 1047-1052