EXTENDED PDF FORMAT
SPONSORED BY
Stem Cell Marker Antibodies
>Learn More
R&D SYSTEMS
www.rndsystems.com

Science
AAAS

Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse

C. M. Wade *et al.*

Science **326**, 865 (2009);
DOI: 10.1126/science.1178158

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this information is current as of August 28, 2013):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/326/5954/865.full.html

Supporting Online Material can be found at:
http://www.sciencemag.org/content/suppl/2009/11/05/326.5954.865.DC1.html

This article **cites 17 articles**, 6 of which can be accessed free:
http://www.sciencemag.org/content/326/5954/865.full.html#ref-list-1

This article has been **cited by** 7 article(s) on the ISI Web of Science

This article has been **cited by** 36 articles hosted by HighWire Press; see:
http://www.sciencemag.org/content/326/5954/865.full.html#related-urls

This article appears in the following **subject collections:**
Genetics
http://www.sciencemag.org/cgi/collection/genetics

Downloaded from www.sciencemag.org on August 28, 2013

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2009 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

processes in the oceans. We posit that future studies will elucidate the relative importance of biotic (viral infection as in this study) and abiotic [phosphorus limitation (*17*)] stress conditions, in shaping community structure of marine microbes, through the detection of different classes of stress-specific lipids.

Although the origin of PCD in unicellular organisms is still unclear, its functional conservation among phylogenetically diverse phytoplankton lineages suggests key evolutionary and ecological drivers in aquatic environments (*29*). The retention and expression of a nearly complete, virus-based sphingolipid biosynthetic pathway, along with its requirement for viral replication and regulation of the host cell fate, now underscores the pivotal role of a chemical-based, coevolutionary "arms race" in mediating host-virus interactions.

References and Notes

1. M. J. Behrenfeld *et al.*, *Science* **291**, 2594 (2001).
2. G. Bratbak, J. K. Egge, M. Heldal, *Mar. Ecol. Prog. Ser.* **93**, 39 (1993).
3. D. C. Schroeder, J. Oke, G. Malin, W. H. Wilson, *Arch. Virol.* **147**, 1685 (2002).
4. J. L. Van Etten, M. V. Graves, D. G. Müller, W. Boland, N. Delaroque, *Arch. Virol.* **147**, 1479 (2002).
5. J. A. Fuhrman, *Nature* **399**, 541 (1999).
6. C. A. Suttle, *Nat. Rev. Microbiol.* **5**, 801 (2007).
7. D. Lindell, J. D. Jaffe, Z. I. Johnson, G. M. Church, S. W. Chisholm, *Nature* **438**, 86 (2005).
8. K. D. Bidle, L. Haramaty, J. B. E. Ramos, P. Falkowski, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6049 (2007).
9. *E. huxleyi* CCMP1516 Main Genome Assembly, http://genome.jgi-psf.org/Emihu1/Emihu1.home.html.
10. W. H. Wilson *et al.*, *Science* **309**, 1090 (2005).
11. A. Monier *et al.*, *Genome Biol.* 10.1101/gr.091686.109 (2009).
12. K. Hanada, *Biochim. Biophys. Acta* **1632**, 16 (2003).
13. Y. A. Hannun, L. M. Obeid, *Trends Biochem. Sci.* **20**, 73 (1995).
14. H. Liang *et al.*, *Genes Dev.* **17**, 2636 (2003).
15. M. J. Allen *et al.*, *J. Virol.* **80**, 7699 (2006).
16. G. Han *et al.*, *J. Biol. Chem.* **281**, 39935 (2006).
17. B. A. S. Van Mooy *et al.*, *Nature* **458**, 69 (2009).
18. Materials and methods are available as supporting material on *Science* Online.
19. D. Lynch, T. Dunn, *New Phytol.* **161**, 677 (2004).
20. B. Brügger *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2641 (2006).
21. M. J. Allen, J. A. Howard, K. S. Lilley, W. H. Wilson, *Proteome Sci.* **6**, 11 (2008).
22. H. Sakamoto *et al.*, *Nat. Chem. Biol.* **1**, 333 (2005).
23. A. Vardi *et al.*, *Curr. Biol.* **9**, 1061 (1999).
24. M. Frada, I. Probert, M. J. Allen, W. H. Wilson, C. de Vargas, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 15944 (2008).
25. M. Turkoglu, *J. Mar. Biol. Assoc. U. K.* **88**, 433 (2008).
26. A. Vardi *et al.*, *Curr. Biol.* **18**, 895 (2008).
27. A. Vardi *et al.*, *PLoS Biol.* **4**, e60 (2006).
28. A. Pagarete, M. J. Allen, W. H. Wilson, S. A. Kimmance, C. de Vargas, *Environ. Microbiol.* 10.1111/j.1462 (2009).
29. K. D. Bidle, P. G. Falkowski, *Nat. Rev. Microbiol.* **2**, 643 (2004).

30. The authors acknowledge M. Soule and E. Kujawinski and the funding sources of the Woods Hole Oceanographic Institution Fourier-Transform Mass Spectrometry Facility (NSF's Major Research Instrumentation Program OCE-0619608 and the Gordon and Betty Moore Foundation) for assistance with MS data acquisition. We thank D. Repeta for assistance and advice on pigment analysis, V. Starovoytov for TEM analysis, and C. Brown and M. Frada for helpful discussions and valuable feedback. This work was supported by NSF grant IOS-0717494 to K.D.B. and A.V., as well as NSF grant OCE-0646944 to B.V.M.

Supporting Online Material
www.sciencemag.org/cgi/content/full/326/5954/861/DC1
Materials and Methods
SOM Text
Figs. S1 to S4
References

5 June 2009; accepted 2 September 2009
10.1126/science.1177322

Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse

C. M. Wade,[1,2,3]* E. Giulotto,[4] S. Sigurdsson,[1,5] M. Zoli,[6] S. Gnerre,[1] F. Imsland,[5] T. L. Lear,[7] D. L. Adelson,[8] E. Bailey,[7] R. R. Bellone,[9] H. Blöcker,[10] O. Distl,[11] R. C. Edgar,[12] M. Garber,[1] T. Leeb,[11,13] E. Mauceli,[1] J. N. MacLeod,[7] M. C. T. Penedo,[14] J. M. Raison,[8] T. Sharpe,[1] J. Vogel,[15] L. Andersson,[5] D. F. Antczak,[16] T. Biagi,[1] M. M. Binns,[17] B. P. Chowdhary,[8] S. J. Coleman,[7] G. Della Valle,[6] S. Fryc,[1] G. Guérin,[19] T. Hasegawa,[20] E. W. Hill,[21] J. Jurka,[22] A. Kiialainen,[23] G. Lindgren,[24] J. Liu,[25] E. Magnani,[4] J. R. Mickelson,[26] J. Murray,[27] S. G. Nergadze,[4] R. Onofrio,[1] S. Pedroni,[14] M. F. Piras,[4] T. Raudsepp,[8] M. Rocchi,[28] K. H. Røed,[9] O. A. Ryder,[30] S. Searle,[15] L. Skow,[18] J. E. Swinburne,[31] A. C. Syvänen,[23] T. Tozaki,[32] S. J. Valberg,[26] M. Vaudin,[31] J. R. White,[1] M. C. Zody,[1,5] Broad Institute Genome Sequencing Platform,[1] Broad Institute Whole Genome Assembly Team,[1] E. S. Lander,[1,33] K. Lindblad-Toh[1,5]*

We report a high-quality draft sequence of the genome of the horse (*Equus caballus*). The genome is relatively repetitive but has little segmental duplication. Chromosomes appear to have undergone few historical rearrangements: 53% of equine chromosomes show conserved synteny to a single human chromosome. Equine chromosome 11 is shown to have an evolutionary new centromere devoid of centromeric satellite DNA, suggesting that centromeric function may arise before satellite repeat accumulation. Linkage disequilibrium, showing the influences of early domestication of large herds of female horses, is intermediate in length between dog and human, and there is long-range haplotype sharing among breeds.

As one of the earliest domesticated species, the horse, *Equus caballus*, has played an important role in human exploration of novel territories. Belonging to the order perissodactyla (i.e., odd-toed animals with hooves), the genus *Equus* radiated into 8 or 9 species around three million years ago (*1*). Members of the family equidae exhibit diverged karyotypes (*2*) and variable centromeric positioning (*1*). With over 90 hereditary conditions, which may serve as models for human disorders (*3*, *4*) (such as infertility, inflammatory diseases, and muscle disorders), the horse has much to offer as a model species.

DNA from a single mare of the Thoroughbred breed was sequenced to 6.8× coverage [supporting online material (SOM) text], resulting in a high-quality draft assembly (designated EquCab2.0) with a 112-kb N50 contig size (SOM) and a 46-Mb N50 scaffold size (tables S1 and S2), and >95% of the sequence anchored to the 64 (2N) equine chromosomes. The 2.5- to 2.7-Gb genome size is somewhat larger than the dog genome (2.5 Gb) and smaller than the human and bovine genomes (2.9 Gb) (*5*–*7*). Segmental duplications (*8*) make up <1% of the equine genome, and most are intrachromosomal duplications such as are seen in many other mammalian genomes (SOM). Repetitive sequences, many equine-specific, make up 46% of the genome assembly (SOM). The predominant repeat classes include long interspersed nuclear elements, dominated by L1 and L2 types (tables S3 and S4) (19% of bases), and short interspersed nuclear elements, including the recent ERE1 and ERE2 and the ancestral main immunogenic regions (7% of bases). Comparison of horse and human chromosomes reveals strong conserved synteny between these species (fig. S1). Indeed, 17 horse chromosomes (53%) comprise material from a single human chromosome (in the dog, it is 29%).

One unexpected feature of the horse genome landscape was the identification of an evolutionary new centromere (ENC) on chromosome 11 (ECA11), captured in an immature state. Several ENCs have been generated in the genus *Equus* by centromere repositioning (a shift of centromeric position without chromosome rearrangement) (*1*). Mammalian centromeres are typically complex structures characterized by the presence of satellite tandem repeats. ENCs are believed to form initially by unknown mechanisms in repeat-free regions and then progressively acquire extended arrays of satellite tandem repeats that may contribute to functional stability (*9*). The centromere of ECA11 resides in a large region of conserved synteny in many mammals, where the horse is the only species with a centromere present, strongly suggesting that this centromere is evolutionarily new. The ECA11 centromere is the only horse centromere lacking any hybridization signal in fluorescence in situ hybridization

Downloaded from www.sciencemag.org on August 28, 2013

[1]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA. [2]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA. [3]Faculty of Veterinary Sciences, University of Sydney, Sydney, New South Wales 2006, Australia. [4]Dipartimento di Genetica e Microbiologia, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy. [5]Department of Medical Biochemistry and Microbiology, Uppsala University, Box 582, SE-751 24 Uppsala, Sweden. [6]Dipartimento di Biologia, Università di Bologna, Via Selmi 3, 40126 Bologna, Italy. [7]Maxwell H. Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, KY 40546, USA. [8]University of Adelaide, Adelaide, South Australia 5005, Australia. [9]University of Tampa, 401 West Kennedy Boulevard, Box 3F, Tampa, FL, USA. [10]Helmholtz Centre for Infection Research, Braunschweig, Germany. [11]Institute of Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Bünteweg 17p, 30559 Hannover, Germany. [12]45 Monterey Drive, Tiburon, CA 94920, USA. [13]Institute of Genetics, Vetsuisse Faculty, University of Berne, Bremgartenstrasse 109a, 3001 Berne, Switzerland. [14]Veterinary Genetics Laboratory, University of California, Davis, CA, USA. [15]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [16]Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA. [17]The Royal Veterinary College, Royal College Street, London NW1 0TU, UK. [18]College of Veterinary Medicine, Texas A&M University, College Station, TX 77843, USA. [19]INRA, UMR 1313, Génétique Animale et Biologie Intégrative, Biologie Intégrative et Génétique Equine, Bâtiment 440.78350, Jouy-en-Josas, France. [20]Equine Research Institute, Japan Racing Association, 321-4 Tokami-cho, Utsunomiya, Tochigi 320-0856, Japan. [21]Animal Genomics Laboratory, School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland. [22]Genetic Information Research Institute, 1925 Landings Drive, Mountain View, CA 94043, USA. [23]Department of Medical Sciences, Uppsala University, 75185 Uppsala, Sweden. [24]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 597, SE-751 24 Uppsala, Sweden. [25]Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA. [26]College of Veterinary Medicine, University of Minnesota, St. Paul, MN 55108, USA. [27]Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis CA, USA. [28]Department of Genetics and Microbiology, University of Bari, Via Amendola 165, 70126 Bari, Italy. [29]Department of Basic Sciences and Aquatic Medicine, Norwegian School of Veterinary Science, N-0033 Oslo, Norway. [30]San Diego Zoo's Institute for Conservation Research, Escondido, CA 92029, USA. [31]Animal Health Trust, Suffolk, CB8 7UU, UK. [32]Department of Molecular Genetics, Laboratory of Racing Chemistry, 1731-2 Tsurutamachi Utsunomiya, Tochigi 320-0851, Japan. [33]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: c.wade@usyd.edu.au (C.M.W.); kersli@broadinstitute.org (K.L.-T.)

experiments probing with the two major horse satellite sequences (fig. S2A, table S8, and SOM text), as if it had not had enough time to acquire satellite DNA. We cytogenetically localized the primary constriction (fig. S2B), then precisely mapped, at the sequence level, the centromeric function using chromatin immunoprecipitation (ChIP)–on–chip experiments (fig. S5). In this region, we found only five sequence gaps [none >200 base pairs (bp)], no protein coding sequences, normal levels of noncoding conserved elements, and typical levels of interspersed repetitive sequences, but no satellite tandem repeated sequences (Fig. 1A). We also found no evidence of accumulation of L1 transposons (*10*) or KERV-1 elements (*11*), which were previously hypothesized to influence ENC formation. We propose that the ECA11 centromere was formed very recently during the evolution of the horse lineage, and, in spite of being functional and stable in all horses, has not yet acquired the marks typical of mammalian centromeres.

The equine gene set is similar to those of other eutherian mammals and has a predicted 20,322

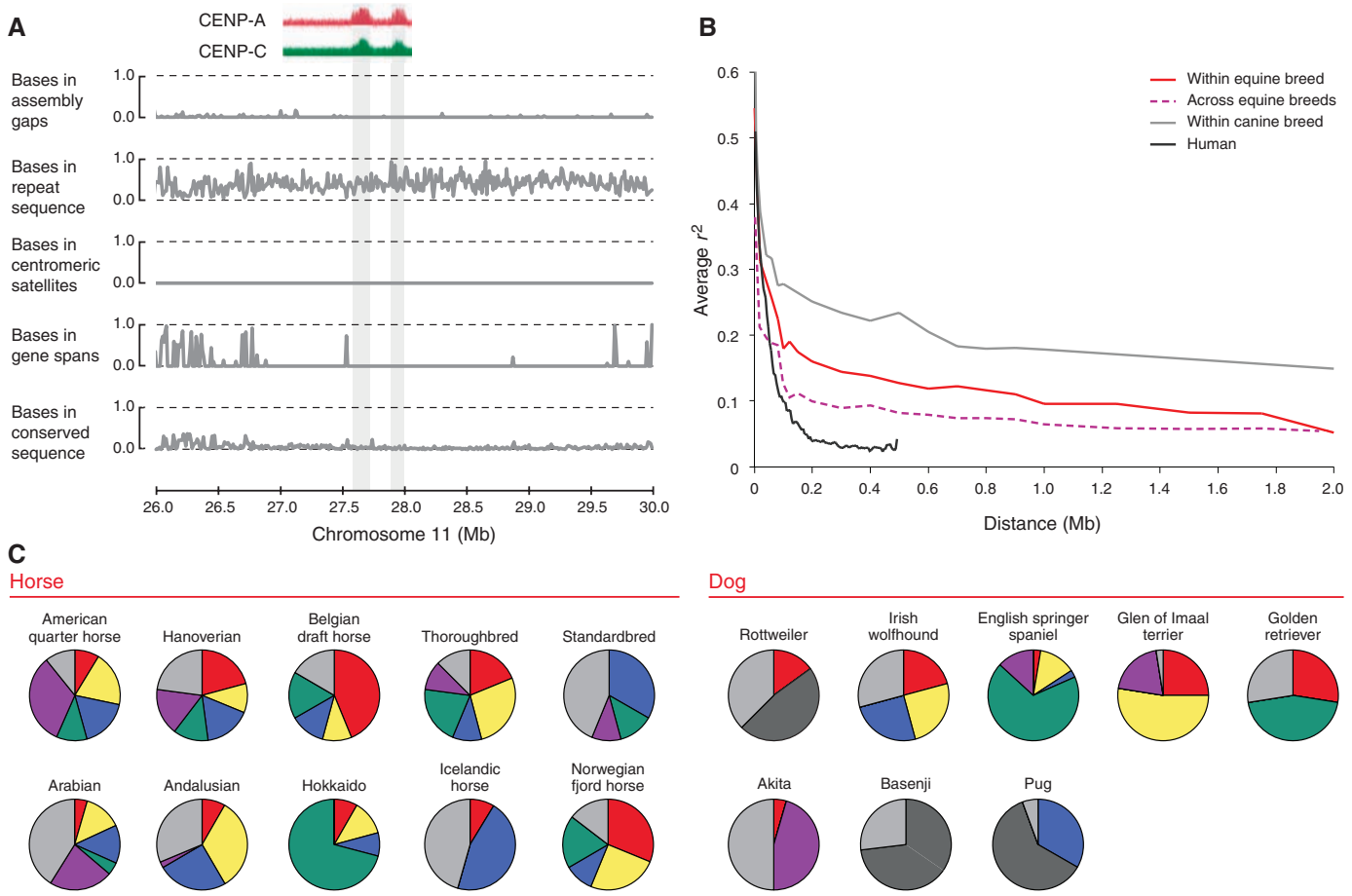

Fig. 1. Major findings of the genome analysis. (**A**) Analysis of the primary centromeric constriction of ECA11: 26,000,000 to 30,000,000 bases. ChIP-on-chip analysis with antibodies against centromeric proteins (CENP-A and CENP-C) shows two regions (136 and 99 kb) bound by kinetochore proteins. There are no uncaptured and few captured gaps, a normal fraction of bases in repeat sequences, no satellite tandem repeats, no protein-coding sequences present nearby, and normal levels of noncoding conserved elements (29 eutherians). (**B**) Horse LD is intermediate between human and dog. (**C**) Horses exhibit more long-range across-breed haplotype sharing than dogs. Haplotypes have the same color across breeds. Haplotypes in <5% of all individuals are light gray, and haplotypes in >5% of all individuals but a single breed are dark gray. Data show LD regions on ECA18 (first 100 kb) and dog chromosome 12 (first 100 kb), which are representative. Full data are in table S11.

Downloaded from www.sciencemag.org on August 28, 2013

protein-coding genes (ENSEMBL build 52.2b), of which 16,617, 17,106, and 17,106 have evidenced orthology to human, mouse, and dog, respectively. The remainder is composed of projected protein-coding genes, novel protein-coding genes, and pseudogenes. One-to-one orthologs with the human account for 15,027 horse gene predictions (SOM). Transcriptome analysis of eight equine samples confirmed the expression of 87% of the 18,039 nonoverlapping genes predicted by ENSEMBL and 88% of the 169,073 predicted exons. Gene family analysis shows paralogous expansion in horses as compared to both human and bovine (SOM) for several interesting families, such as keratin genes related to the condition of pachyonychia (nail bed thickening) in humans (*12*), perhaps affecting hoof formation; and opsin genes for photoreception, possibly advantageous for visual perception of predators (table S9).

The history of horse domestication, which has important implications for trait mapping strategies, differs in important ways from that of the domestic dog but is perhaps similar to that of the cow. Horses do not appear to have undergone a tight domestication bottleneck, and the presence of many matrilines in domestic horse history has been postulated (*13*). Screening the horse Y chromosome revealed a limited number of patrilines, consistent with a strong sex bias in the domestication process (*14*).

We first generated a single-nucleotide polymorphism (SNP) map of more than one million markers at an average density of one SNP per 2 kb by lightly sequencing seven horses from different breeds and by mining the assembly for SNPs (table S10).

We characterized the haplotype structure within and across breeds by genotyping 1,007 SNPs from 10 regions of the genome (SOM) in 12 populations, including 11 breed sets (each with 24 representatives), and 1 set of individual representatives from 24 other breeds and equids. 98% of SNPs were validated, with an average of 69% being polymorphic in alternate breeds (SOM). Like the bovine (*15*), within-breed linkage disequilibrium (LD) is moderate, dropping to twice the background levels (r^2) at 100 to 150 kb (Fig. 1B). The majority of breeds showed similar LD (SOM and fig. S7), and major haplotypes were frequently shared among diverse populations (Fig. 1C). Based on the length of LD in the horse, the number of haplotypes within haplotype blocks, and the polymorphism rate, power calculations suggest that ~100,000 SNPs are sufficient for association mapping within all breeds as well as across breeds (SOM and fig. S8).

Phylogenetic relationships among breeds were inconsistent across resequenced regions (fig. S9), which is most likely a consequence of the close relationships of horse breeds worldwide. We were unable to phylogenetically separate *E. przewalskii* from the domesticated horses, despite its different karyotype (2N = 66 versus 2N = 64 for the domesticated horse), which is in agreement with recent findings (*16*), whereas the donkey (*E. africanus*) is clearly a distinct taxon (fig. S9, table S14, and SOM text). This suggests that either intermixing of *E. przewalskii* and *E. caballus* occurred after subspecies separation or that *E. przewalskii* is recently derived from *E. caballus*.

We demonstrated the utility of the equine genome sequence and a SNP map by applying these resources to mutation detection for the Leopard Complex (LP) spotting locus (SOM). LP (Appaloosa spotting) is defined by patterns of white occurring with or without pigmented spots (fig. S10). Homozygosity confers a phenotype associated with congenital stationary night blindness in the Appaloosa breed (*17*). Fine mapping of a 2-Mb region followed by regional sequence capture and sequencing (300 kb) found no indications of associated copy number variants or insertions or deletions but found 42 associated SNPs. Of these, 21 reside within an associated haplotype near a candidate gene *melastatin 1* (*TRPM1*), which is expressed in the eye and melanocytes (*18*). Two conserved SNPs may be good candidates for the causal mutation.

Our analysis of the first high-quality draft sequence of a horse (*E. caballus*) distinguishes *E. caballus* from earlier eutherian genomes by its large synteny with humans and the identification of a centromere repositioning event that may provide an effective model to study epigenetic factors responsible for centromere function. Our results demonstrate that horse population history has led to across-breed haplotype sharing, increasing the feasibility of across-breed mapping. Mapping projects in the horse are likely to accelerate in the coming years and will identify mutations in genes related to morphology, immunology, and metabolism, which may benefit human health.

References and Notes

1. L. Carbone *et al.*, *Genomics* **87**, 777 (2006).
2. V. A. Trifonov *et al.*, *Chromosome Res.* **16**, 89 (2008).
3. Online Mendelian Inheritance in Animals (University of Sydney, 2009), http://omia.angis.org.au/.
4. B. P. Chowdhary, N. Paria, T. Raudsepp, *Anim. Reprod. Sci.* **107**, 208 (2008).
5. K. Lindblad-Toh *et al.*, *Nature* **438**, 803 (2005).
6. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
7. Bovine Genome Sequencing and Analysis Consortium, *Science* **324**, 522 (2009).
8. T. S. Mikkelsen *et al.*, *Nature* **447**, 167 (2007).
9. M. Ventura *et al.*, *Science* **316**, 243 (2007).
10. A. C. Chueh, E. L. Northrop, K. H. Brettingham-Moore, K. H. Choo, L. H. Wong, *PLoS Genet.* **5**, e1000354 (2009).
11. D. M. Carone *et al.*, *Chromosoma* **118**, 113 (2009).
12. J. W. Wu *et al.*, *J. Eur. Acad. Dermatol. Venereol.* **23**, 174 (2009).
13. C. Vila *et al.*, *Science* **291**, 474 (2001).
14. G. Lindgren *et al.*, *Nat. Genet.* **36**, 335 (2004).
15. Bovine HapMap Consortium, *Science* **324**, 528 (2009).
16. A. N. Lau *et al.*, *Mol. Biol. Evol.* **26**, 199 (2009).
17. L. S. Sandmeyer, C. B. Breaux, S. Archer, B. H. Grahn, *Vet. Ophthalmol.* **10**, 368 (2007).
18. R. R. Bellone *et al.*, *Genetics* **179**, 1861 (2008).

19. We thank the Kentucky Horse Park and L. Chemnick for samples, L. Gaffney for graphics, and M. Daly for useful discussions. Supported by the National Human Genome Research Institute, the Dorothy Russell Havemeyer Foundation, the Volkswagen Foundation, the Morris Animal Foundation, the Centro di Eccellenza di Genomica in Campo Biomedico e Agrario, and the Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale (PRIN-2006). K.L.T. is the recipient of a European Young Investigator award funded by the European Science Foundation. Sequences have GenBank accession numbers AAWR02000001 to AAWR02055316. 1,163,466 discovery SNPs have accession numbers rs6844103 to rs69617090 in dbSNP 130.

Supporting Online Material
www.sciencemag.org/cgi/content/full/326/5954/865/DC1
Materials and Methods
SOM Text
Figs. S1 to S11
Tables S1 to S14
References

23 June 2009; accepted 10 September 2009
10.1126/science.1178158

MafB/c-Maf Deficiency Enables Self-Renewal of Differentiated Functional Macrophages

Athar Aziz,[1,2,3]* Erinn Soucie,[1,2,3]* Sandrine Sarrazin,[1,2,3] Michael H. Sieweke[1,2,3]†

In metazoan organisms, terminal differentiation is generally tightly linked to cell cycle exit, whereas the undifferentiated state of pluripotent stem cells is associated with unlimited self-renewal. Here, we report that combined deficiency for the transcription factors MafB and c-Maf enables extended expansion of mature monocytes and macrophages in culture without loss of differentiated phenotype and function. Upon transplantation, the expanded cells are nontumorigenic and contribute to functional macrophage populations in vivo. Small hairpin RNA inactivation shows that continuous proliferation of MafB/c-Maf deficient macrophages requires concomitant up-regulation of two pluripotent stem cell–inducing factors, KLF4 and c-Myc. Our results indicate that MafB/c-MafB deficiency renders self-renewal compatible with terminal differentiation. It thus appears possible to amplify functional differentiated cells without malignant transformation or stem cell intermediates.

The nonproliferative state of terminally differentiated cells is assured by robust, often redundant mechanisms (*1*, *2*), and in rare exceptions where fully mature cells can re-enter the cycle, proliferation remains transient and/or involves de-differentiation (*3*). It remains unknown what renders differentiated cells refractory to the same mitogen signals that stimulate the proliferation of their direct precursors. For example, the proliferative response of myelomonocytic

Downloaded from www.sciencemag.org on August 28, 2013

www.sciencemag.org/cgi/content/full/326/5954/865/DC1

Supporting Online Material for

Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse

C. M. Wade, * E. Giulotto, S. Sigurdsson, M. Zoli, S. Gnerre, F. Imsland, T. L. Lear, D. L. Adelson, E. Bailey, R. R. Bellone, H. Blöcker, O. Distl, R. C. Edgar, M. Garber, T. Leeb, E. Mauceli, J. N. MacLeod, M. C. T. Penedo, J. M. Raison, T. Sharpe, J. Vogel, L. Andersson, D. F. Antczak, T. Biagi, M. M. Binns, B. P. Chowdhary, S. J. Coleman, G. Della Valle, S. Fryc, G. Guérin, T. Hasegawa, E. W. Hill, J. Jurka, A. Kiialainen, G. Lindgren, J. Liu, E. Magnani, J. R. Mickelson, J. Murray, S. G. Nergadze, R. Onofrio, S. Pedroni, M. F. Piras, T. Raudsepp, M. Rocchi, K. H. Røed, O. A. Ryder, S. Searle, L. Skow, J. E. Swinburne, A. C. Syvänen, T. Tozaki, S. J. Valberg, M. Vaudin, J. R. White, M. C. Zody Broad Institute Genome Sequencing Platform, Broad Institute Whole Genome Assembly Team, E. S. Lander, K. Lindblad-Toh*

*To whom correspodence should be addressed. E-mail: c.wade@usyd.edu.au (C.M.W.); kersli@broadinstitute.org (K.L.-T.)

Published 6 November 2009, *Science* **326**, 865 (2009)
DOI: 10.1126/science.1178158

This PDF file includes:

Materials and Methods
SOM Text
Figs. S1 to S11
Tables S1 to S14
References

Supporting On-line Materials

Materials and Methods 1
Data Generation 1
Assembly 2
Segmental Duplication Landscape 2
Repetitive Elements 3
Conserved Synteny 3
Centromeric Analysis 4
Localization of satellite DNA to all equine chromosomes 4
Cytogenetic localization of ECA11 primary constriction 4
Immunofluorescence on metaphase chromosomes 5
ChIP-on-chip analysis 5
ECA11 peri-centromeric region bioinformatic analysis 6
Gene Set 6
SNP Discovery 7
Population Genetics 7
Determination of Homozygous and Heterozygous blocks 7
Diversity Breeds 7
Linkage Disequilibrium and Haplotype Sharing 8
Power for gene mapping 9
Phylogeny 9
Variation within Przewalski's horses 9
Proof of Principle Mapping 10
Hybrid Capture for Massively Parallel Sequencing 10
Coverage 11
SNP and indel detection in Massively Parallel Sequencing 11
Functional ascertainment of sequenced intervals 11
Assessment of Copy Number Variation (CNV) in sequenced intervals 11
Supporting Text 12
Features of the equine genome 12
Genes 13
Equine population genetics and haplotype structure analysis 14
Proof of principle genetic mapping 16
Supporting Figures 17
Supporting Tables 29
Supplemental References 51

Materials and Methods

Data Generation

The Thoroughbred horse is a breed derived from few founders and has maintained breeding records since the 18th century. The breed is considered to be the epitome of equine athleticism and is a significant economic animal. The genome of a Thoroughbred

mare was sequenced using a whole genome shotgun approach using ABI 3730 sequencers (Applied Biosystems, Foster City, CA). Blood was collected and preserved in a BD Vacutainer tube with spray coated K_2EDTA to a ratio of 1.8mg EDTA per ml of blood, and the genomic DNA was extracted using Qiagen Blood and Cell Culture DNA kit with 100/G Qiagen Genomic Tips. During precipitation of the DNA with isopropanol the eluate was centrifuged at 12,000g for 20 minutes, and washed with 70% ethanol at 12,000g for 10 minutes. The resulting DNA was of high molecular weight with the great majority above 38kb. The anion-exchange resin, and gravity flow of the tips combined with the high centrifugation during precipitation ensured high purity, yield and molecular weight, with the majority of DNA above 38kb. The individual selected for sequencing was selected from an inbred herd of horses maintained by The Baker Research Institute at Cornell University, Ithaca, New York. Other equine samples were contributed by horse genetics researchers from around the world (Table S1).

A BAC library was made by Children's Hospital Oakland Research Institute (CHORI) from a half brother of the sequenced mare, and 314,972 BAC end sequences used in the assembly were generated by the University of Veterinary Medicine Hannover, Germany. Clone libraries (4kb, 10kb and Fosmid) were generated by the Broad Institute of Harvard and MIT. The proportions of sequence coverage contributed by the clone libraries were: 4Kb inserts 4.96× sequence coverage; 10Kb inserts 1.42×; 40Kb fosmid inserts 0.40×, resulting in a total coverage of 6.8×.

Assembly

The assembly of the horse genome used the package Arachne 2.0 (S*1*) and methods as previously described (S*2*). Ordering and orientation used combined probe sequences and positional information from the comprehensive radiation hybrid and FISH maps (S*3*). Remaining unplaced scaffolds greater than 1Mb in size were correctly positioned and oriented by Fluorescence In-Situ Hybridization by the University of Kentucky (S*4*). Summary statistics for the assembly are shown in Tables S2 and S3. The genome size estimate (2.689Gb) is based on the summed gapped scaffold sizes from all scaffolds (mapped and unassigned) in addition to the unplaced reads accounting for 260Mb (adjusted for expected coverage) that were not included in contigs.

Segmental Duplication Landscape

Segmental duplications were assessed using methods described previously (S*5*), Overall, 0.5% of the genome is involved in intra-chromosomal segmental duplications, and 0.1% is inter-chromosomally duplicated. An additional 0.4% of the genome shows duplications in unplaced contigs: most of these are likely to be artifacts of the assembly since such contigs typically comprise repetitive data that are difficult to resolve in the assembly process. The duplicated regions are small, with the largest being ~60Kb. Chromosome 25 was the most duplicated at 1.7%, and several chromosomes had only 0.2% duplication. Segmental duplications are defined as regions of >=90% similarity, extending for at least 1Kb.

Repetitive Elements

To detect novel equine repetitive sequences, we analyzed EquCab2.0 with PALS/PILER (S*6*) and RepeatScout (S*7*). We also compared the horse genome sequences to known mammalian repeats from Repbase using CENSOR (S*8*), and constructed horse-specific consensus sequences. Three runs with PALS/PILER were carried out in turn for chromosomes in linear order, permuted order for chromosomes and finally permuted order for the whole genome (including unassigned contigs). Each comparison required 90% identity for alignments and a minimum alignment length of 200 bp. In all 50,206 highly redundant repeat families were detected. A similar analysis with RepeatScout identified 24,813 repeat families genome-wide with the default running parameters.

All PALS/PILER families with greater than 95% identity and 90% length overlap from all three PALS runs were clustered to generate consensus sequences. We similarly clustered RepeatScout sequences genome-wide. Final clustering yielded 28,586 clusters, which were collapsed to consensus sequences. Consensus sequences were then compared to SwissProt sequences with WUBLAST (S9) to detect and remove protein coding, or non-repeat sequences. The final consensus set included 26,860 sequences. Many of these consensus sequences are less than 200bp in length. Intact known repeats accounted for 1,608 of the consensus sequences, while unknown repeats accounted for 6,531 of the consensus sequences each present at a minimum of 3 copies in the genome. The remainder of the consensus sequences consisted of chimeric or compound repeats (Table S4). These may have a biological origin either as novel, compound repeats or as a result of preferential, sequential tandem integration of particular repeats. Chimeric repeats often consist of embedded repeats of the same class. Most of the unknown repeats were present at very low copy number. When an arbitrary threshold of 100 copies was imposed, the number of unknown consensus sequences dropped to 48. An up to date repeat library was generated with the RepeatMasker library supplemented with our versions of the known repeats (probably improved consensus sequences) and our 48 unknown common repeats. Chimeric/compound repeats and repeats present in fewer than 100 copies were omitted from this library. All non-chimeric novel sequences were uploaded to Repbase (http://www.girinst.org/).

Counts of repeats clustered by type as determined in the chromosome scaffolds are shown in Table S5 and the distribution by chromosome is shown in Table S6. The repeat composition of the horse is quite different from that observed in other eutherian mammals. In dog tRNA derived SINEs are present at > 1,000,000. The horse is comparable to human in this regard, with a tiny fraction of SINEs derived from tRNAs (0.02% Table S5). The predominant SINEs are the recent ERE1/2 and the ancestral MIRs, while the predominant LINEs are L1 and L2. Note that the number of L1/L2 elements in the horse is about half that observed in human, while the percentage of the genome contributed by L1/L2 is comparable. This may reflect a larger number of intact LINE elements and perhaps more currently active LINE elements.

Conserved Synteny

Conserved synteny between mammalian species was assessed with the methods previously described (S5) at several resolutions and using blastz alignments as input. The genome builds HSA18, Btau4.0 , CanFam2.0 and mm8 were used for Human, Cow, Dog

and Mouse, respectively. The set of conserved blocks and segments are described for a resolution of 100 Kb (Table S7, Fig. S1). There are a total of 403 syntenic segments between human and horse as compared to 443 segments between human and dog, 1257 human to cow (the high number may be influenced by assembly quality) and 457 conserved segments between human to mouse. Not surprisingly, the horse has had the fewest rearrangements, which is consistent with a higher percent identity on the nucleotide level to human than even the dog (62 median percent identity versus 65 median percent identity). This is further reflected in anchors being longer (dog median: 297 while horse median at 374 bases) Anchors (at 100 Kb) also cover much more of territory: Horse anchors cover 902,880,738 (1/3 of the genome) of human bases, whereas dog anchors cover 793,321,127 bases. If one assumes that breaks fix at random then under any model (say Poisson) there is a higher probability of seeing more breaks between two species if the evolutionary distance separating the species is larger.

Centromeric Analysis

A novel evolutionary centromere was detected on ECA11. Here we present the details of the supporting analysis.

Localization of satellite DNA to all equine chromosomes

Two horse satellite sequences were isolated from a horse genomic library in lambda phage. One satellite (37 cen), consisting of 221 bp repeated units (Accession number: AY029358) is 93% identical to the horse major satellite family. The 2pI sequence, consisting of a 23 bp repeat, (Accession numbers: AY029359S1 and AY029359S2) belongs to the e4/1 family. One and two color FISH experiments were performed using the two satellites as probes. The localization of the two satellites on horse chromosomes is reported in Table S8. No satellite signal was seen on ECA11 (Fig. S2A).

Cytogenetic localization of ECA11 primary constriction

In order to localize the primary constriction of ECA11, we performed two or three color fluorescence in situ hybridization experiments on horse metaphase chromosome spreads prepared from primary fibroblasts derived from a healthy mare, with a panel of BAC clones as probe. The clones were identified with the BAC end sequences, ordered according to sequence homology with human chromosome 17 (S*10*), and the horse genome assembly. We first localized 6 clones, which mapped on the short arm of ECA11 (CHORI241-37D6 on p13; CHORI241-62D3, CHORI241-54B4, CHORI241-289O22, CHORI241-180K7 and CHORI241-10L5 on p12). We then identified 5 clones in the adjacent region, further away from the p arm telomere, separated from each other, on the horse genome assembly, by a distance of about 500 kb: CHORI241-262K23 mapped on p12; CHORI241-374A5, CHORI241-333G20, CHORI241-91H8 and CHORI241-21D14 were localized on the primary constriction region (Fig. S2B). These experiments were used to roughly identify the region responsible for the centromeric function, to be further investigated with ChIP-on-chip experiments.

Immunofluorescence on metaphase chromosomes

Horse metaphases were prepared according to standard methods (S*11*) spread on slides with a cytospin centrifuge, permeabilized at 37 °C for 15 minutes in KCM buffer (KCl 120 mM, NaCl 20 mM, Tris-HCl 10 mM, NaEDTA 0.5 mM and Triton X-100 0.1 % v/v). Incubation with the CENP-A or CENP-C primary antibodies (polyclonal antibodies against the human CENP-A and CENP-C proteins (S*12*), diluted 1:100 in KCM containing 1% BSA, was performed for 1 hour at RT. Slides were then washed three times for 5 minutes in KB- (Tris-HCl 10 mM, NaCl 150 mM and BSA 1%). The secondary antibody (TRITC conjugated anti-rabbit), diluted 1:100, was added and incubated for 1 hour at RT. After two washes in KB-, the slides were fixed in 4% formalin in KCM for 15 minutes at RT and then in cold methanol:acetic acid 3:1 for 15 minutes. Finally, the chromosomes were stained with Hoechst 33258 (Roche, Basel, Switzerland) and digital images were acquired with a Zeiss Axioplan fluorescence microscope (Zeiss,Jena, Germany) and a cooled CCD camera (Photometrics, Tucson, AZ, USA). The antibodies against human CENP-A (Fig. S3A) and CENP-C (Fig. S3B) specifically recognize all horse centromeres.

ChIP-on-chip analysis

A 2.7 Mb region between nucleotides ECA11: 25,566,599 and 28,305,611 was arrayed and analyzed by chip-on-chip.

The cross-linking agent formaldehyde was added at 1% final concentration to the culture medium of horse fibroblasts. Chromatin was extracted according to standard procedures (S*13*) and immuno-precipitated with rabbit polyclonal antibodies (S*12*) directed against human CENP-A or CENP-C centromeric proteins (S*12*) that have been demonstrated to recognize horse centromeres. Purified DNA fragments were amplified using the Whole Genome Amplification kit (Sigma-Aldrich, St.Louis, USA). DNA fragments obtained from the immuno-precipitated material and from total genomic DNA (input) were labeled and co-hybridized to a custom high-density tiling array (NimbleGen, Reykjavík, Iceland), with an average resolution of 100 bp, containing the 2.7 Mb of the ECA11 sequence (between nucleotides 25,566,599 and 28,305,611), where the centromere was cytogenetically localized. The immuno-precipitated DNA, before and after amplification, was proven to be enriched in centromeric sequences by Real-Time PCR performed with primers (described in figure legend Fig. S4.) were designed based on the known sequences of the two major horse satellite repeats (accession numbers AY029358, AY029359S1 and AY029359S2) (Fig. S4).

Hybridization with DNA array, signal scanning, data extraction and data normalization were carried out (Nimblegen, Reykjavík, Iceland). DNA binding peaks were identified with the statistical model and methodology described at http://chipanalysis.genomecenter.ucdavis.edu/cgi-bin/tamalpais.cgi (S*13*). The analysis was performed using very stringent conditions (98th percentile threshold and $P<0.0001$). With both antibodies, the analysis, showed two clear peaks of hybridization spanning about 136 kb (nt 27,643,400-27,779,400) and 99 kb (nt 27,950,800-28,050,000), respectively, separated by a region of about 165 kb (Fig. S5A). No hybridization was observed in the sequences flanking the two peaks, nor two unrelated regions from ECA20 and ECA26 that were used as negative controls (Fig. S5).

ECA11 peri-centromeric region bioinformatic analysis

The region of analysis (nt 26,000,000-30,000,000) was divided into 10kb windows. Bases occurring in assembly gaps (EquCab2.0), conserved elements over 29 mammals (S*14*), gene spans (University of Santa Cruz, genome table browser), repeat elements (equine specific library – see above), and centromeric satellite sequences as identified by BLAST (S*15*) of the repeats (accession no. AY029358, AY029359S1 and AY029359S2), were tallied and charted as a proportion of bases within the 10kb windows (Fig. S2C). Where genes had multiple transcripts, gene spans were derived for the largest possible isoform.

Based on previous reports of a correlation between neocentromeres and the presence or transcription of LI or KERV-1 elements, we examined these regions for the presence of these elements. Bases occurring in L1 transposons over ECA11 were taken from repeat summary tables. The region analyzed for the centromere was divided into 10Kb windows. The mean proportions of bases in L1 transposons per window were calculated for the same region that was interrogated for neocentromeric analysis. By examination of the RepeatMasker output, this region was also found to be free of KERV-1 (also known as MERK-1) elements that may also be associated with centromeres (S*16*).

Gene Set

Evidence based methods featured in the standard Ensembl mammalian pipeline were used to generate gene predictions for the *E. caballus* assembly. Priority was given to predictions from mammals over other vertebrates and non-vertebrates and horse proteins were given the greatest priority. We compared the set of predictions to 1:1 homologous genes in human and mouse, and missing homologs in the horse annotation were recovered with exonerate (S*17*). We used horse and human cDNAs to add untranslated regions to protein based predictions. Gene counts were derived from Ensembl biomart for database build 52.2b.

Gene family expansion analysis made use of the Ensembl biomart database (http://www.ensembl.org/biomart/) build 55. First, Equus caballus genes were filtered to include only genes with gene-types of "protein coding" and "paralog eqcab IDs only" in the multi-species comparisons filter. Human ensemble gene ids were selected in the attributes filter (N=15,167) for the equine genes and the output from this filter was saved as a text file. The equivalent process was followed for bovine (N=15,958) and human genes (N=15,420). Paralog counts were compared across species using the Ensemble human gene identifier. Gene identifiers demonstrating a two-fold increase in paralog count in horse over both human and cow (N=99) were output (Table S9). The Ensembl Biomart database was used to extract functional descriptors by gene ontology (Biological process, or molecular function) for the expanded horse genes.

Transcriptomic sequencing was carried out by Illumina Genome Analyzer RNA-seq protocols (Illumina, San Diego). Eight equine tissues (normal and LPS-stimulated articular cartilage, normal and LPS-stimulated synovial membrane, placental villous, testis, cerebellum, and a 34-day embryo) were studied. Sequence tags (35bp) were mapped to the equine genome sequence (EquCab2.0) with the ELAND module of Illumina's analysis pipeline. Of the total tags generated, 54% (158,787,455) had unique alignments. Nucleotide coordinates defining equine gene boundaries as indicated by

RNA-seq data were compared to Ensembl predictions and supported gene predictions were tallied.

SNP Discovery

SNP discovery breeds were selected to represent the global diversity of horse populations and included Andalusian, Arabian, Akhal Teke, Icelandic Horse, Standardbred, Thoroughbred and Quarter horse breeds. SNP were discovered using the SSAHA-SNP (S*18*) algorithm using the methods described previously (S*2*). Comparison between these sequences and the assembled genome as well as SNP discovery in the assembly process itself has resulted in a high quality SNP map of 1,162,753 distinct SNP (Table S10) SNPs discovered from the mulitple breeds totaled 416,680 SNP. SNP in unassigned portions of the genome are not counted towards these totals but represent a further 312,499 SNP.

The observed rate within the Thoroughbred assembly itself is 1 SNP per 3,000 bp (Table S10). This is partly a result of the closed breeding structure of the Thoroughbred horse breed and the fact that among Thoroughbreds, a relatively inbred horse was selected for the genome project, but is also overly conservative based on the stringent quality criteria for SNP calling during the assembly process (estimated to correspond to 1 per 2000bp when corrected for the ascertainment. SNP discovery within the assembly required that both alleles be observed twice. By Poisson, the chance of having cover of at least 4 reads when the mean cover is 6.8 × is ~90%. If we use binomial probabilities to account for the chance of seeing at least two alleles of each kind given that the read cover is sufficient, then the chance of this is 73%. In total, if the alleles are present we expect to observe them both twice about 66% of the time. This would give us an adjusted SNP rate of 1 per 2000bp which corresponds with a rate of 1 per 1050bp in the heterozygous portions of the genome.

Population Genetics

Determination of Homozygous and Heterozygous blocks

The homozygous or heterozygous state of each chromosome within the assembly was determined with the use of a Viterbi algorithm (S*19*) in a single Viterbi chain for each chromosome (Fig. S6). Observational probabilities were calculated separately for each 100kb window with a Poisson distribution with a mean dependent upon the state (Homozygous: expected bases/SNP =40,000; Heterozygous: expected bases/SNP=300), and the number of bases examined in the window (100,000). Transitional probabilities for a change from either homozygous to heterozygous, or vice versa, were set at 0.02. The chance of the state remaining the same was one minus this value. Homozygous and heterozygous block sizes were determined by merging adjacent windows with the same state.

Diversity Breeds

To examine the effects of population history on the genome structure, we characterized the haplotype structure within and across breeds. We genotyped 1,007 newly discovered SNPs from ten random regions of the genome in twelve populations,

including eleven breed sets (each with 24 representatives), and one across-breed set of individual representatives from 23 other breeds and equids. The breeds used in this analysis reflected global breed diversity (Table S1).

The breeds represented by 24 individuals in the study were the Arabian, Andalusian, Belgian Draft, French Trotter, Hokkaido, Hanoverian, Icelandic Horse, Norwegian Fjord, Quarter Horse, Standardbred, Thoroughbred. The mixed breed set included a donkey (*Equus africanus*) and Przewalskii horse (*Equus przewalski*) in addition to 21 individual horse-breed representatives (Belgian, Shire, Suffolk Punch, Quarter Horse, Standardbred, Thoroughbred, American Saddlebred, French Trotter, Selle Français, Trakehner, Paso Fino, Friesian, Akhal teke, Arabian, Andalusion, Lusitano, Icelandic, Norwegian Fjord, Exmoor Pony, Mongolian Horse, Misaki). SNPs for genotyping in these populations were ascertained from ten genomic regions each covering two megabases (Mb). The mixed breed set was sequenced in the first 5kb of each of the 10 regions and SNPs were ascertained. Two breeds (Misaki and Lusitano) failed sequencing resulting in 19 horse breed representatives that could be used in the 5kb haplotype sharing analysis.

Linkage Disequilibrium and Haplotype Sharing

Linkage disequilibrium was assessed by dispersing 100 SNP randomly throughout each of 10 regions. The regions chosen were (all in EqaCab2.0 coordinates): ECA 4: 44,820,715- 46,820,715, ECA 5: 52,839,996 - 54,839,996, ECA 6: 11,581,414 - 13,581,414, ECA 10: 17,762,861 - 19,762,861, ECA 11: 25,111,945 - 27,111,945, ECA 14: 86,952,151 - 88,952,151, ECA 15: 18,556,343 - 20,556,343, ECA 17: 62,038,338 - 64,038,338, ECA 18: 16,040,788 - 18,040,788, ECA 29: 2,427,654 - 4,427,654

We assessed the long-range correlation between markers by r^2, a traditional measure of LD, across the 2Mb regions (Fig. S7). An alternative method identified individuals homozygous within the first 5Kb of each region (n= 110 of 288 total on the basis of resequencing data), and examined the increasingly distant genotypes for each individual until fewer than 50% of the originally homozygous individuals maintained homozygosity. Similar methods were applied in the dog genome analysis (*2*).

Haplotype sharing was assessed for breeds represented by multiple individuals (24 per breed). For this analysis, genotyped SNPs with minor allele frequencies greater than 5% within the first 100Kb of each chromosomal region were phased with Fastphase (S*20*) into haplotypes and the observed haplotypes were tallied within and between breeds. In all 194 haplotypes were observed over the 10 loci examined for all breeds. Within breed, the average number of haplotypes observed was 5.3. The most frequent haplotype for all horses was observed in an average of 10.2 of 11 included breeds, while the second most frequent haplotype was observed in 9.1 breeds (Table S11).

We have also analyzed the degree of similarity between breeds within the 5kb resequenced regions using horses from the mixed breed set of the diversity panel (N=19). Within each segment, we tallied the number of times each haplotype was observed. More than 60% of haplotypes present at a locus are observed in more than one breed (Table S12).

Power for gene mapping

The number of SNPs required to capture horse genomic variation was tested by sampling the ten genotyped regions with 1,000 permuted maps for each of five different densities up to an equivalent of 100,000 SNPs. The maximum density of SNP present was equivalent to an array with 125,000 SNPs. SNPs not included in each map at densities lower than this were treated as hidden test SNPs. For each test SNP the maximum r^2 with any SNP included in the map was recorded for a range of breeds with different levels of LD and the average value was plotted against array density (Fig. S8).

The results of this analysis demonstrate well-supported trend lines that enable some prediction outside of the testing limits (Table S13). With a desired mean r^2 max of 0.8, the indications are that 75K SNP may be sufficient for breeds with long LD but are unlikely to fully represent variation in breeds with moderate or short LD at this stringency. At lower stringencies (e.g. r^2_{max}=0.5) >40,000 SNPs may be sufficient for all other than short LD breeds and across breed mapping.

Phylogeny

The phylogenetic distances among breeds were assessed by comparing the SNPs discovered in the 5Kb resequencing regions, where individual representatives from 22 breeds and 2 equids (*Equus africanus* and *Equus przewalskii*) were analyzed. The package PHYLIP (Phylogeny Inference Package) Version 3.5c (S*21*) including modules dnadist, kitsch and drawgram were used to assess the phylogenetic distances between individuals for SNPs with minor allele frequencies of greater than 5% over all horses. Heterozygous SNPs within individuals were coded with ambiguity codes. Bootstrapping (1000 permutations) was used to ascertain the stability of the phylogenies (Fig. S9).

Variation within Przewalski's horses

To test the hypothesis that the lack of divergence of the Przewalski's horse sample was due to sampling error or to use of an outlier individual we increased our sample size to include a wider sampling of the population. Multiple alignments of PCR traces from 8 Przewalski's horse representatives of several breeding lines was carried out with Consed (S*22*) for 7 regions (Table S14). Differences that were homozygous in the assayed individuals were noted as were the number of high quality bases examined. The observed divergence from the genome assembly for this group was one SNP per 550bp. This is similar to the divergence of the Przewalski representative from the re-sequencing assays (40 differences in 21,695 high quality bases or 1 SNP per 542bp), thus suggesting that the original Przewalski individual used in the breed phylogeny analysis was representative. Thus, we conclude that Przewalski's horses have similar divergence rates from domestic horses as wolves to dogs. However, unlike wolves the Przewalski's horses have few interspecific mutations that are not represented in the domestic horse population and this leads to the undifferentiated phylogenies observed at many loci.

Proof of Principle Mapping

The Leopard Complex (often referred to as Appaloosa spotting) is defined by striking patterns of white that can occur with or without oval pigmented spots (Fig. S10A). These patterns are observed in several breeds. Horses with pigmented spots in the white patterned areas are heterozygous for a mutation, which in the homozygous state confers a phenotype with few or no pigmented spots that is associated with Congenital Stationary Night Blindness (CSNB) (S*23*). A ~10 Mb region on ECA1 had previously been identified using linkage analysis (S*24*).

We performed fine–mapping of this interesting coat-color locus (Leopard-complex) to show the utility of the SNP resource (Fig. S10B). We then performed hybrid selection followed by Illumina sequencing of associated regions. Fine mapping of the Leopard-complex entailed genotyping of 70 SNPs across the region ECA1:107,194,138-109,299,508 in 192 horses [Appaloosa: (125) including 13 with confirmed CSNB; Quarter Horse: (2); Knabstrupper (29); Noriker (36)]. Of these, only horses phenotyped as solid (control) or else in the phenotypic classes of "few-spot", "snow-cap", or night blind (case) were evaluated in the final analysis. This set comprised: Appaloosa (40/18 case/control, respectively]; Appaloosa phenotyped for Congenital stationar night blindness (CSNB) (13/14); Noriker (2/18); Knabstrupper (13/0). This strategy used the maximally differentiated phenotypes for association. Mapping that included presumed heterozygous phentoypes also resulted in the association of the same haplotype although with lower probability. Association analysis of SNPs and haplotypes was carried out with Haploview (S*25*) resulting in the identification of a 173Kb haplotype of association (ECA1: 108,197,355- 108,370,150) containing the two most associated SNP: BIEC2-48043 ECA1: 108,248,113 Pr $6.15E^{-23}$; and BIEC2 48086_ECA1:108,370,091). Two samples (Knabstrupper "few-spot" and Appaloosa "snow-cap") presumed to be homozygous for the associated allele based on SNP genotypes were chosen for hybrid capture (NimbleGen, Reykjavík, Iceland) and mutation detection with Illumina Genome Analyzer (Illumina, San Diego, CA) over the non-repetitive sequence within the interval ECA1:108,200,000- 108,500,000.

The Grey coat color mutation resulting in premature hair greying was used as a positive control for copy number detection. Horses with this phenotype are prone to the development of melanoma. While, the mutation has been already described as a 4.6Kb duplication in the 6th intron of STX17 on ECA25 (S*26*), we were interested to see if the new sequencing technologies could readily detect the duplication by comparing a homozygote for the duplication with a homozygote for the ancestral haplotype upon which the duplication arose. An analysis of SNP and sequence coverage in the 350Kb region clearly identified the known duplication and three SNPs in disequilibrium with the duplication but none of these in conserved or transcribed regions. Comparison of this analysis with those of other regions demonstrates an absence of observable copy number variation in the Leopard complex samples.

Hybrid Capture for Massively Parallel Sequencing

Genomic DNA was fragmented and purified on QIAquick columns (Qiagen, Valencia, CA). Fragment ends were blunted and 5'-phosphorylated, and a 3' overhang of a single adenosine was added. Illumina adapters (Illumina, San Diego, CA) were ligated to the DNA fragments, which were subsequently hybridized to NimbleGen Sequence

Capture Arrays (Nimblegen, Reykjavík, Iceland), following manufacturer's recommendations. Following enrichment for selected sequences on the arrays, Illumina adapter ligated fragments were amplified with a high-fidelity polymerase and purified (QIAquick). Enrichment of samples was assessed by quantitative PCR comparison to the same samples prior to hybridization.

Coverage

The coverage varied somewhat between samples, likely as a function of the efficiency of the hybridization. The six horses included in the analysis had a mean coverage of 53.5 ± 35.4 reads per base with most of the observed variation arising from differences between horses rather than between loci. A minimum of 5-fold coverage was required to call a polymorphic base where both alleles were observed twice. The Icelandic horse had the greatest coverage with 99% of tiled bases achieving greater than five fold cover. The average horse with successful hybridization had 94% of repeat masked and tiled bases with >5× cover.

SNP and indel detection in Massively Parallel Sequencing

Genome Analyzer I single reads (each 35bp) were aligned to the EquCab2.0 using 12-mer seeding followed by a local Smith-Waterman alignment, allowing for insertions or deletions in the alignment. Output, reported all discrepancies between the reads and the reference. Focusing only on the mismatch calls in this dataset resulted in the set of potential SNPs: the minimum coverage to call the mismatch was set to 5; the minimum quality to call the mismatch was set to be the highest quality observed in the reads.

Discrepant lengths of alignment of k-mers (k=24) between reads and the reference were analyzed for the presence of insertion/deletion events. Windows of alignment moved in single base increments. The EquCab2.0 assembly base adjacent to insertion start point was recorded by breed if the alignment passed the quality criteria above.

Functional ascertainment of sequenced intervals

Intervals identified as conserved elements using SiPhy (S*14*) in 29 eutherian mammals were marked within the sequenced regions. Associated SNPs or indels falling within 5 bases of the marked intervals were considered to be “conserved” and therefore candidate mutations. Potential transcripts identified by Illumina Genome Analyzer RNA-SEQ of 8 equine tissues (see section on Genes) were compared to the associated polymorphic sites. Bases with overlap with RNA-SEQ transcripts were treated as if they are transcribed. The general level of transcription within the Grey target region was somewhat higher than in the Leopard region in the tissues studied thus far.

Assessment of Copy Number Variation (CNV) in sequenced intervals

For all pairs of samples the mean-normalized ratio of cumulative coverage in 1kb windows across the target region for the trait was calculated. CNVs appear in this metric as a sharp deviation from the background signal. The ratio of signals is differentiated to enhance the signal and to handle the fluctuations in coverage across different genomic regions. The standard deviation from the mean of the distribution is then calculated from the differentiated signal, and histograms were used to assess results (Fig. S11). The

cumulative coverage both forward and backward across the target region is used to alleviate instability during the initial passes of the algorithm over the data.

This analysis demonstrates that no significant copy number variants were present in the sequenced regions for horses other than the grey Lippizaner, for which the mutation is known to be a duplication event.

Supporting Text

Features of the equine genome

DNA from a single Thoroughbred mare was used in the construction of a 6.8× coverage high-quality draft assembly, with added contiguity generated by the inclusion of BAC end sequences from a related male Thoroughbred horse, from which a BAC map has been produced (S*10*). The assembly (designated EquCab2.0) is of high quality and contiguity with a 112 kb N50 contig size and a 46 Mb N50 scaffold size (Tables S2, S3), and with >95% of the sequence anchored to the 64 (2N) equine chromosomes.

The estimated euchromatic genome size of *Equus caballus* based on the total lengths of scaffolds is 2.47 Gb (Table S3). This is somewhat larger than the dog and smaller than the human genome (2.9 Gb) and bovine (2.9 Gb) genome (S*2, S27*) (S*28*). Segmental duplications in the assembly, determined using standard methods (*5*) comprise less than 1% of the equine genome, and the majority of these (with ~80% mapped to chromosomes) are intra-chromosomal duplications such as those seen in the dog and mouse genomes. However, the assembly has many unplaced sequence reads suggesting that the true genome size may be up to 2.7 Gb (Table S3) similar to the mouse (S*29*). The unassembled sequence is highly repetitive in nature. Using standard mammalian repeat libraries, 39% of the genome assembly is annotated as repetitive, however, using custom repeat libraries that include horse-specific repeats, 46% of the assembled sequence is identified as repetitive, a quantity comparable with that of the human genome. The predominant repeat classes include LINEs dominated by L1 and L2 types (Tables S5, S6) (19% of bases) and SINEs including the recent ERE1/2 and the ancestral MIRs (7% of bases). Novel equine repeats account for a large fraction of the observed consensus repeat elements, but only 48 elements are present in greater numbers (>100 copies). Chimeric repeats are an important newly identified source of repetitive elements in the horse and are difficult to classify unambiguously. Frequently, these repeats appear to stem from the random placement of new repeats within existing repeats, and the different repeat classes occur in proportion to their overall frequency in the genome.

In the horse genome an evolutionary new centromere (ENC), has been captured in an immature state. A previous study (S*11*) showed that, during the last 3 million years, several evolutionary new centromeres (ENCs) were generated in the genus *Equus*, by centromere repositioning (shift of centromeric position without chromosome rearrangement). In particular, it was shown that the centromere of ECA11 is an ENC. Mammalian centromeres are typically complex structures characterized by the presence of satellite tandem repeats. It is commonly thought that ENCs, following seeding of

satellite tandem repeat DNA into a single-copy DNA region, progressively acquire the characteristic extended arrays of satellite tandem repeats and that presumably these stabilize the centromeric function (S*30*). However, centromeres lacking satellite sequence have been described only in occasional human clinical cases, where they stabilize an acentric fragment (S*31, S32*) or, occasionally in normal chromosomes of a normal individual, in which the centromere was repositioned (S*33-S35*) as is the case in evolution. The centromere of ECA11 was the only horse centromere lacking any hybridization signal in FISH experiments performed using the two major horse satellite sequences as probes. (Fig. S2A and Table S8). The absence of satellite signals in the ECA11 centromere suggests that this ENC may not have yet "matured" to the point of being endowed with satellite DNA. The availability of the horse genome sequence enabled us to test this hypothesis. Using a panel of BAC clones as probes in FISH experiments, we cytogenetically localized the primary constriction (Fig. S2B). We then fine-mapped the centromeric function by ChIP-on-chip analysis, using antibodies against the kinetokore proteins CENP-A and CENP-C, and conclusively demonstrated that centromeric function is localized within this region (Fig. S5). The ~400kb region to which the centromere is localized (Chr11: 27,643,400-28,049,600) contains only five sequence gaps in the EquCab2.0 assembly and none is larger than 200 bp (all are spanned by paired-end reads from multiple clones). It has no protein coding sequences, normal levels of non-coding conserved elements and typical levels of interspersed repetitive sequences, but no satellite tandem repeated sequences (Fig. S2C). A recent study of a human neocentromere (S*36*) identified a transcriptionally active LINE_L1 (L1) retrotranposon. We have not ascertained whether the L1 elements within the ECA11 ENC are transcribed, but we note that the ECA11 ENC has fewer L1s than the rest of the genome. The entire genome has ~16.3% of bases in L1 repeats; all of ECA11 has ~9.3% bases in L1, while the neocentromeric region from 26.7 to 28.1Mb has only 5% bases in L1 elements. While this genomic region resides in a large region of conserved synteny with many mammals, the only species in which a centromere is present is the horse. In conclusion, we propose that the ECA11 centromere was formed very recently during the evolution of the horse lineage and, in spite of being functional and stable in all horses, has not yet acquired all the marks typical of mammalian centromeres.

Genes

The equine gene set is, not surprisingly, similar to that of other eutherian mammals. Gene structure annotation by the ENSEMBL pipeline predicts 20,322 protein-coding genes (Ensembl build 52.2b). The number of genes orthologous to these predictions is comparable in human (16,617), mouse (17,106) and dog (16,159). The remainder comprises projected protein coding genes, novel protein-coding genes, and pseudogenes. Gene family expansion relative to other species seems to be active in olfactory and immune genes genes (Table S9) as has been seen in other non-primate mammals.

Interesting expansions are present in gene families relating to vision (particularly opsins) and keratins, which form a significant component of hair and nails (hooves). The expansion of the sulfotransferase gene family may generate disulfide bonds for the keratin cysteine amino acids, which likely lend rigidity to hooves (S*37, S38*). Among the expanded keratin genes are those which when mutated in humans give rise to the condition pachyonychia congenita (S*39*). The symptoms of this condition include

thickening of nail beds and skin on the hands and feet and it seems likely that changes in these genes may have led to the development of the hoof.

Equine population genetics and haplotype structure analysis

To facilitate trait mapping in the horse, we set out to generate the necessary resources and to characterize the haplotype structure among horse populations. The history of horse domestication differs in important ways from that of the domestic dog. First, whereas dogs went through a tight bottleneck at the time of domestication from the wolf, horses do not appear to have been subjected to a tight domestication bottleneck. It is likely that horses were first domesticated between 4,000 and 6,000 years ago (S*40*) either as a source of food (S*41*) or for transport. Mitochondrial evidence (S*42*) suggests the presence of many matrilines in domestic horse history; perhaps due to herds of mares being captured for food production, or else as a result of the geographical distribution of horse genetic material during periods of military conquest. Screening for polymorphisms on the horse Y chromosome has revealed a limited number of patrilines, consistent with a strong sex-bias in the domestication process (S*43*). Second, some horse breed registries (with exceptions such as the Thoroughbred) have fewer restrictions than dog kennel clubs regarding introgression of limited genetic stock from other breeds and many breeds have used out-crossing in their initial establishment.

We first generated a single nucleotide polymorphism (SNP) map of more than one million markers, comprised of >700,000 SNPs discovered by comparing the two chromosomes within the genome assembly and ~400,000 SNPs were discovered from ~100,000 whole genome shotgun reads from each of seven horse breeds. The resulting SNP map has an average density of one SNP per 2kb (Table S10). The mean SNP rate between horses of different breeds is similar to humans and dogs at 1 SNP per 1,200 bp. Within the sequenced horse 46% of the genome was determined to be homozygous using a 100kb window size (Fig. S6). This amount of homozygosity is likely somewhat higher than in the average horse, since the chosen individual came from a herd that had been inbred.

98% of SNPs were polymorphic (true SNP) within the genotyped panels. Among all horses the mean minor allele frequency for polymorphic SNPs was 0.22 (+/-0.15) in a flat distribution from zero to 0.5, suggesting that horse populations have been subject to neither massive expansion, contraction nor bottlenecks throughout their collective histories (S*44*).

When we assessed the likelihood that SNPs were polymorphic in a third breed based on the discovery breed used the mean polymorphism rate for SNPs from any discovery breed comparison was 69%. The breed with the highest proportion of polymorphic SNPs in any third breed (besides the discovery breed and the thoroughbred) was the Standardbred (74.3% polymorphic), while the lowest was the Akhal teke (61% polymorphic in a third breed). The low rate at which Akhal teke discovery SNPs were observed to be polymorphic in other breeds suggests that the Akhal teke breed is genetically diverged from the other eleven breeds, since these SNP seem to be Akhal teke specific.

The within-breed linkage disequilibrium (LD) in the horse is moderate. It is approximately five times shorter than the within-breed LD for dog and five times longer than in a human population of European Ancestry (Centre d'Etude du Polymorphisme Humain (CEPH)) population (Fig. 1B). The LD is of similar extent to that observed in the bovine (S*45*). While the majority of horse breeds show similar patterns of LD, the Thoroughbred appears to have LD akin to that of dog breeds (Fig. S7). This was expected because the Thoroughbred breeding population has maintained a closed breeding structure for well over a century and was derived from relatively few founders (S*41*). Some ancient breeds, such as the Norwegian Fjord Horse and Hokkaido, show somewhat shorter LD. Interestingly, the across breed LD in the horse is only slightly shorter (50-70 kb) than the within breed values, likely reflecting the absence of strong bottlenecks during breed formation and the requirement of many mares to maintain population size, due to the limited number of offspring per mare.

The horse's particular history is further illustrated by the frequent sharing of major haplotypes among diverse horse populations, emphasizing their common history. On average we observe five haplotypes for each breed within the first 100 kb segment of the LD region, and nineteen haplotypes across breeds. The two most common haplotypes across breeds, for a given segment, are typically present in almost all of the eleven breeds and 98% of 100 kb haplotypes with a minor allele frequency (MAF) >0.05 (of all haplotypes) are shared between at least two of the breeds studied (Fig. 1C, Tables S11, S12). By contrast, human haplotypes are short and are highly shared, while dog haplotypes are long within breeds and these breed-specific haplotypes show relatively little sharing across breeds (65% of 100kb haplotypes are shared at MAF >0.05). Ancestral canine haplotypes are short and are highly shared and these 10 kb haplotypes are always present in multiple dog breeds. Thus, equine genetic history is reflected both in across breed LD and haplotype sharing properties.

The marker density required for effective genome-wide association mapping can be estimated from the length of LD in the horse, the number of haplotypes within haplotype blocks, and the polymorphism rate. Effective tagging of haplotypes for genetic mapping purposes in the average breed will require an estimated map density of greater than 100,000 SNP (for 5 haplotypes in 18,000 blocks of 150kb when 70% of SNP are polymorphic within a given breed), and more markers will be required for effective mapping in ancient breeds as well as in those with a large effective population size. The estimate is supported by stochastic sampling over the 2 Mb regions in the breeds examined for LD (Table. S13), where with a sampling density equivalent to 100,000 polymorphic SNPs we were able to obtain mean maximum r^2 values of >0.5 for tested SNPs in all breeds as well as the across breed groups (Fig. S8).

Phylogenetic analysis of the SNPs in the re-sequenced regions shows varying relationships between breeds among the different loci tested (Fig. S9). The heterogeneity is most likely a consequence of the close ancestral relationship of the horse breeds studied even though they were sourced from geographically dispersed populations. Because the breeds are from a common global population; small differences in genotype

or local gene history caused by drift or selection, create large effects on the branching structure of the phylogeny. This creates a varying phylogeny across loci where few nodes are significantly supported at bootstrap values of > 0.90.

The Przewalski's horse, *E. przewalskii,* is considered a more ancestral equid than the domestic horse and has a different karyotype (2N=66). In agreement with recent findings (S*46*), we are unable to phylogenetically separate *E. przewalskii* from the domesticated horse breeds, whereas donkey (*E. africanus*) clearly has its own branch in phylogenic comparison (Fig. S9). We were concerned that the sample we had chosen first may have been an outlier to the wider *E. przewalskii* population and so evaluated the similarity of this individual horse to other Przewalski horses using 8 widely ascertained samples of *E. przewalskii* selected to represent all extant lineages. We speculate that there was either inter-mixing of *E. przewalskii* and *E. caballus* after sub-species separation or that *E. przewalskii* is recently derived from *E. caballus*.

Proof of principle genetic mapping

The Leopard Complex (often referred to as Appaloosa spotting) is defined by striking patterns of white that can occur with or without oval pigmented spots (Fig. 10A). These patterns are observed in several breeds. Horses with pigmented spots in the white patterned areas are heterozygous for a mutation which in the homozygous state that in the Appaloosa horse breed, confers a phenotype with few or no pigmented spots that is associated with Congenital Stationary Night Blindness (CSNB) (S*23*). A ~10 Mb region on ECA1 had previously been identified using linkage analysis (S*24*). We performed fine mapping on the 2 Mb region with the strongest LOD scores in the linkage analysis; it contains five known genes including the melastatin 1 (*TRPM1*) gene, which is a good candidate based on its expression in the eye and in melanocytes (S*47*). We genotyped 70 SNPs in 92 individuals with extreme phenotype (few spot or solid) from three breeds, Appaloosa (85 individuals), Noriker (20), and Knabstrupper (13), and identified a 173 kb haplotype that was strongly associated with disease across breeds (Fig. S10B). After microarray capture and targeted sequencing we found no indications of copy number variants or insertion-deletions associated with the phenotype, but found 42 associated SNP variants, of which 21 reside within the associated haplotype. Two SNPs are located within conserved elements and may be good candidates for the causal mutation. The candidate mutations for the Leopard complex will now be examined for association with phenotype in additional horses as well as for functional consequences.

Supporting Figures

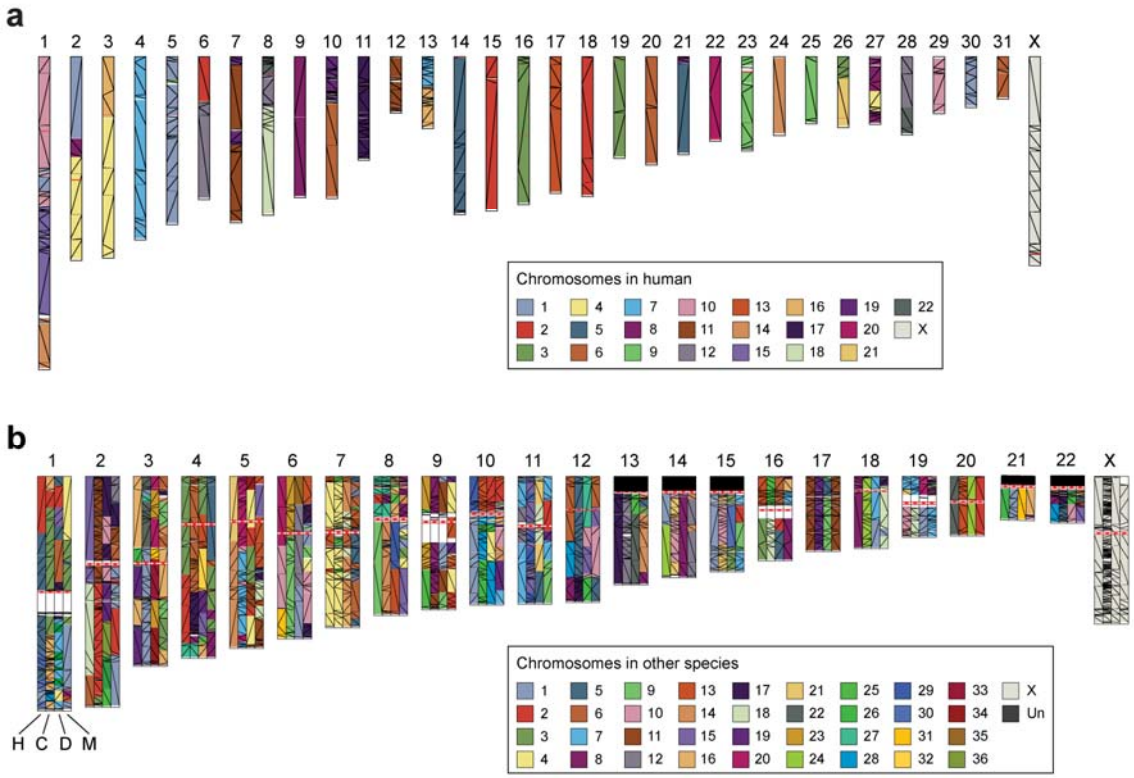

Fig. S1. Conserved synteny a. Horse chromosomes are colored by corresponding Human chromsomes. **b.** Human chromosomes are colored by corresponding chromosomes from Horse, Cow, Dog and Mouse (H, C, D and M, respectively). Corresponding chromosomes for other species are colored according to the key shown. Diagonal lines indicate sequence orientation.

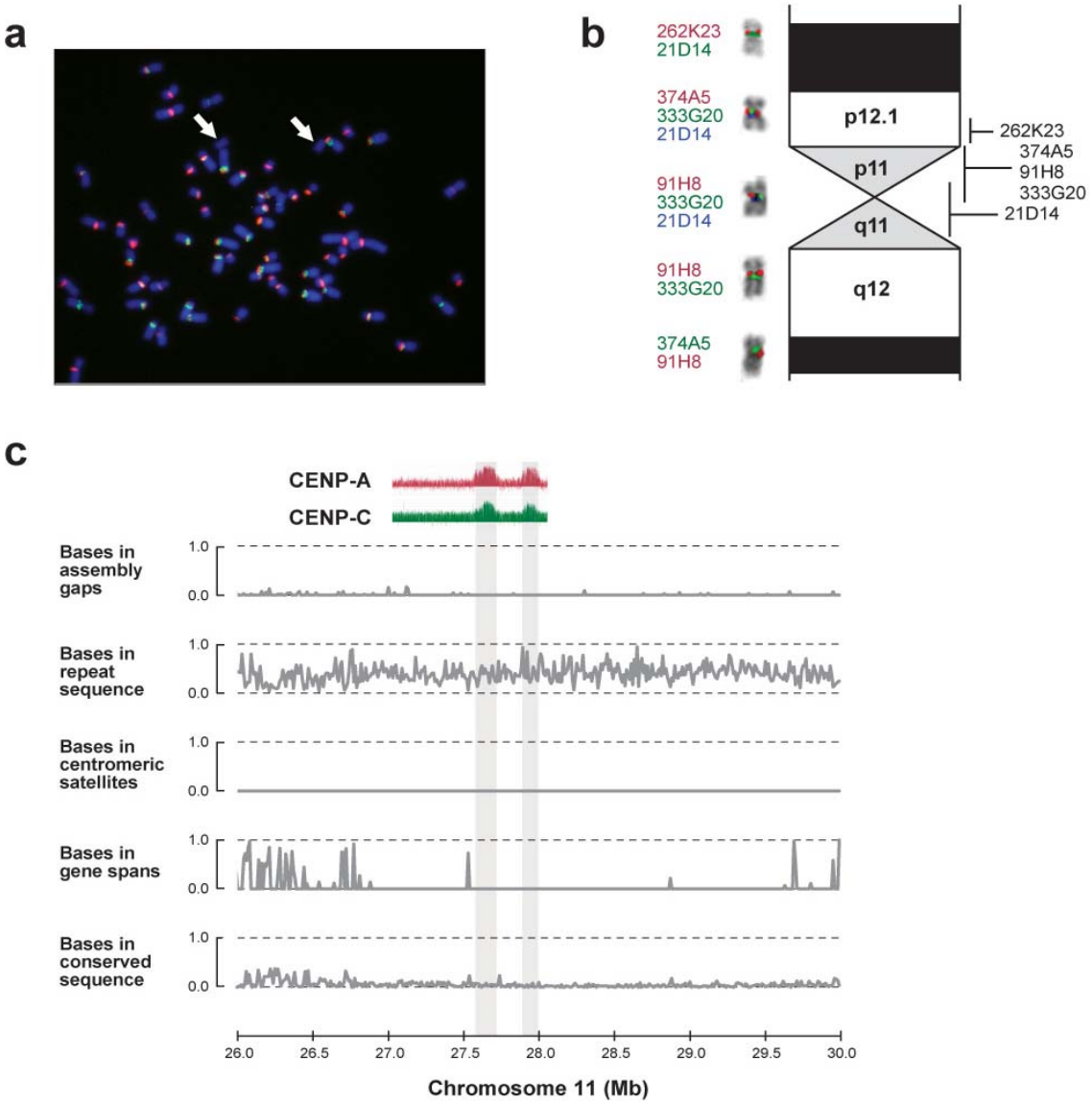

Fig. S2. Localization and sequence analysis of the ECA11 centromere. (**a)** Hybridization of the two major horse satellite sequences on horse chromosomes; the 23 bp repeat (2p1) is labeled in red and the 221 bp repeat (37cen) is labeled in green. All centromeres are labeled with one or both satellite probes except chromosome 11 (arrows). (**b)** Schematic representation of the cytogenetic localization of the primary constriction within a 2.7 Mb region occurring between the BAC clones CHORI241-262K23 and CHORI241-21D14; only the pericentromeric region of chromosome 11 is shown; numbers correspond to names of informative BAC clones. (**c)** Bioinformatic analysis of the sequence comprising the primary constriction of ECA11, between nucleotides 26,000,000 and 30,000,000; in panel one the results of the ChIP-on-chip analysis performed with antibodies against centromeric proteins (CENP-A and CENP-C) are shown, indicating that two regions of 136 and 99kb, respectively, are bound by kinetochore proteins; there are essentially no uncaptured and few captured gaps in EcaCab2.0 of this sequence (panel two); a normal fraction of bases occur in repeat

sequences (panel three) but the region is completely devoid of satellite tandem repeats (panel four). No protein coding sequences are present in the region binding centromeric proteins as well as in the nearby regions (panel five), but there are normal levels of non-coding conserved elements based on conservation among 29 eutherian mammals (panel six).

a b

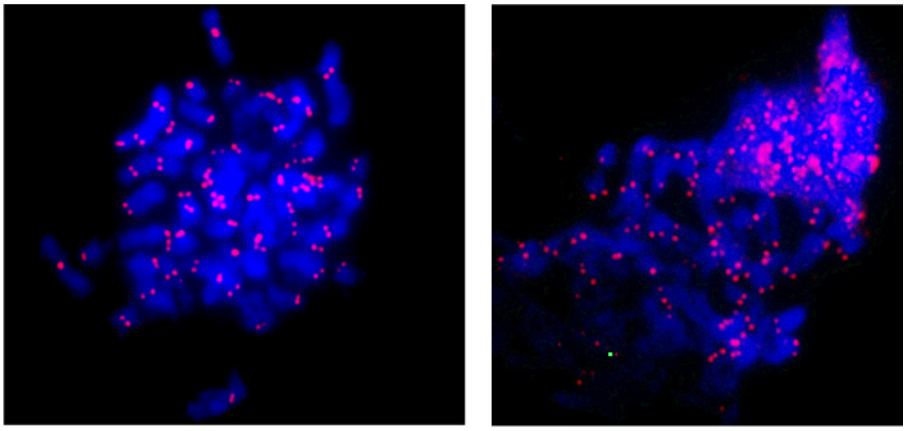

Fig. S3. Immuno-localization of CENP-A and CENP-C on equine chromosomes
Antibodies against human centromeric proteins CENP-A (left) and CENP-C (right) bind horse centromeres (red fluorescence signals).

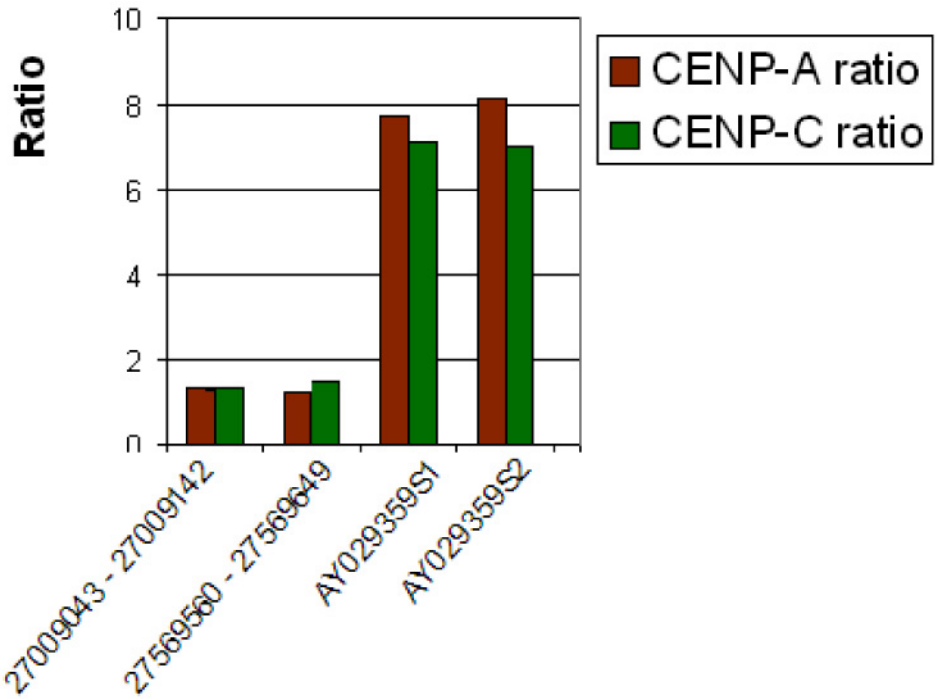

Fig. S4. RT-PCR analysis of the ChIP-chip samples. Results are presented as the log2 ratio between the hybridization signal, obtained with the immunoprecipitated DNA using anti-CENP-A (red) or anti-CENP-C (green) antibodies and that from the input DNA sample. The X-axis shows the genomic position of each primer used or the accession number of the sequence used for the design. The first two pairs of columns represent negative controls with primers that map into non centromeric regions of chr. 11, 27009043-27009142 (forward – TTGCCCTGATAGCGAGAAAT, reverse – CTATTCCTTGGGGCTTCTCC) and 27569560-27569649 (forward - ATGCCCTGGACTGTAAAACG, reverse - ATCCTCAAAGCTGAGCCAAA). They show no significant enrichment in comparison to the last two pairs of columns which correspond to primer pairs designed on the sequence of the major horse centromeric satellite DNA: AY029359S1 (forward – TGTGAAACCACATGCAGGA, reverse – CTGCCTGTTGTCTGGTGT) and AY029359S2 (forward – TGTGAAACCACATGCAGGAT, reverse CTGCCTGTTGTCTGGTGTGT). The lengths of all the PCR products are comparable (~100 bp).

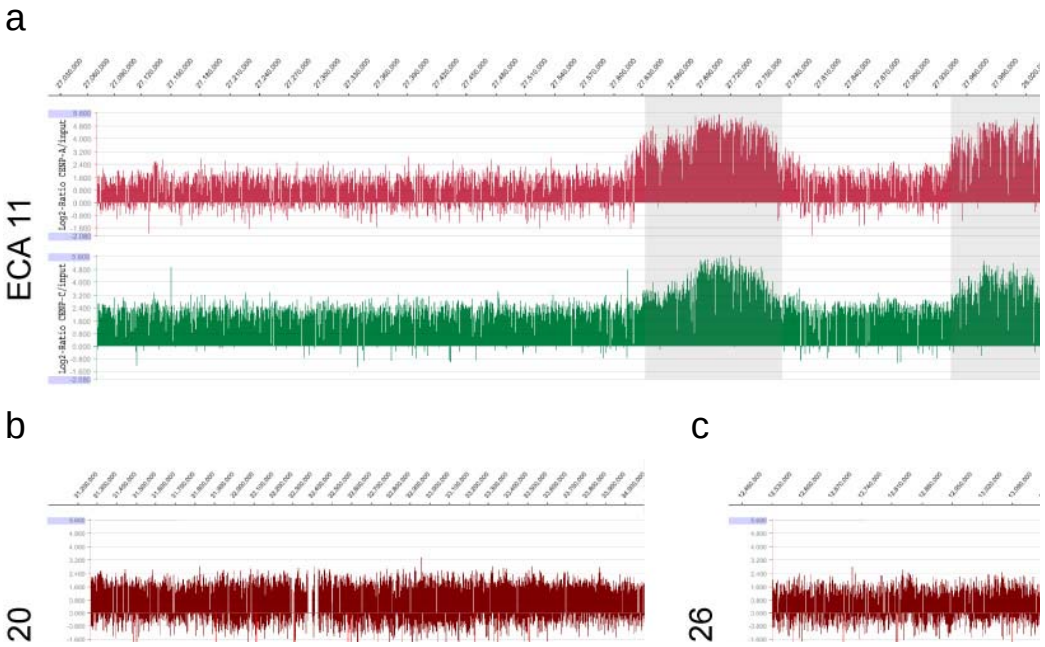

Fig. S5. Partial view of the ChIP-on-chip analysis data on ECA11, ECA20 and ECA26 with anti-CENP-A and -CENP-C antibodies. Results are presented as the log2 ratio between the hybridization signals obtained with DNA immunoprecipitated using anti-CENP-A or CENP-C antibodies and that from the input DNA sample. The X axis shows the genomic position of each oligo on the respective regions. The data are visualized by the SignalMap software (NimbleGen Systems, Reykjavík, Iceland). Details of the microarray structure are reported at the NimbleGen site (http://www.nimblegen.com). In ECA11 the CENP-A and CENP-C domains are indicated by shaded areas. In the regions flanking the two domains, in the segment between the two domains, as well as in the non-centromeric regions of ECA20 and ECA26 (used as negative controls) no significant hybridization is detected, using very stringent conditions (98th percentile threshold, P <0.0001).

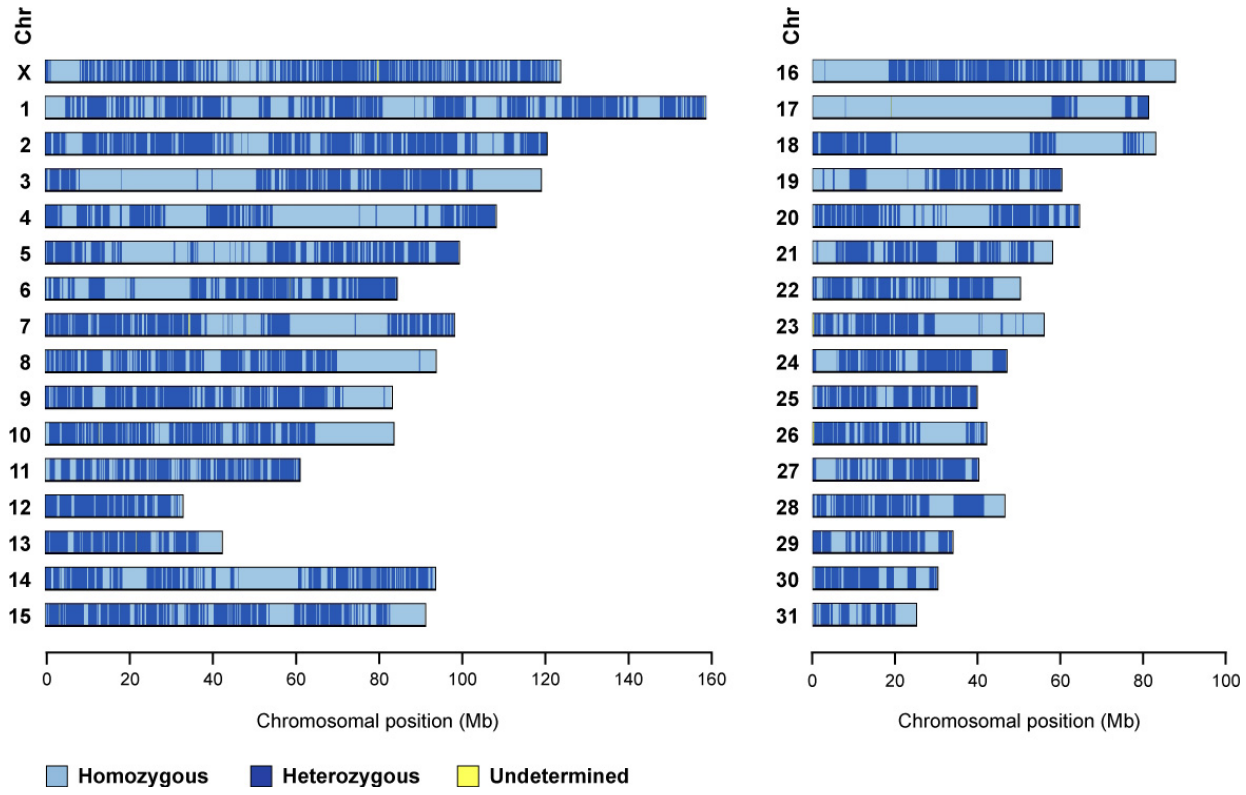

Fig. S6. Homozygosity in the sequenced horse. Regions of heterozygosity and homozygosity in the genome assembly were determined by the application of a Viterbi algorithm at 100kb resolution using regional distribution of SNP as the method of assessment. By this method, 46% of the genome is determined to be homozygous, which is higher than has typically been observed in other mammalian genomes sequenced to date.

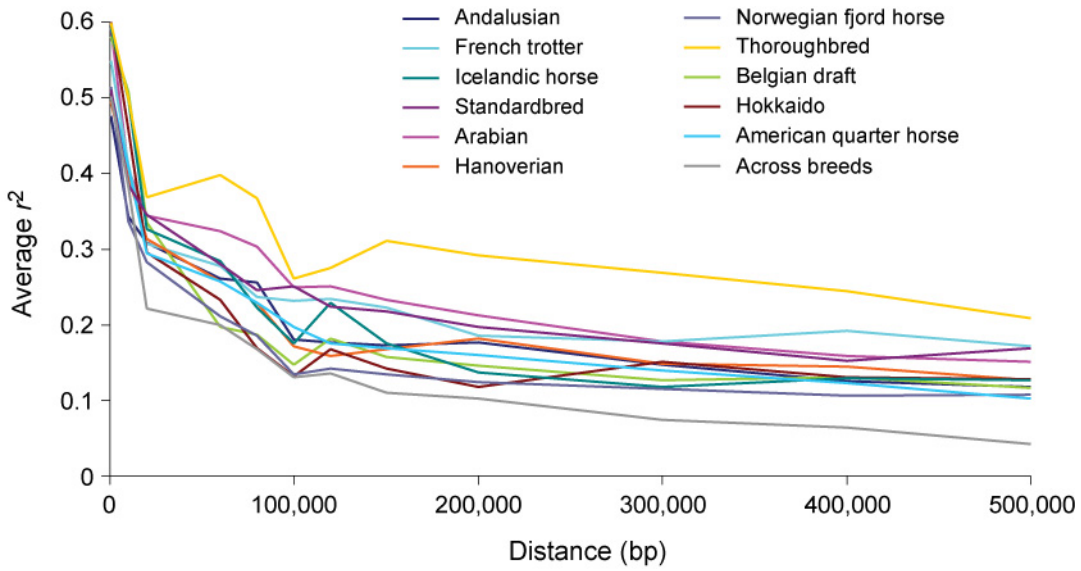

Fig. S7. Linkage disequilibrium (LD) variation by breed. Individual horse breeds examined have similar LD except for the Thoroughbred horse, which has unusually high LD akin to that of the average dog breed. In both species, data are based on binned means of all possible pair-wise r^2 among SNPs over all distances within assayed regions.

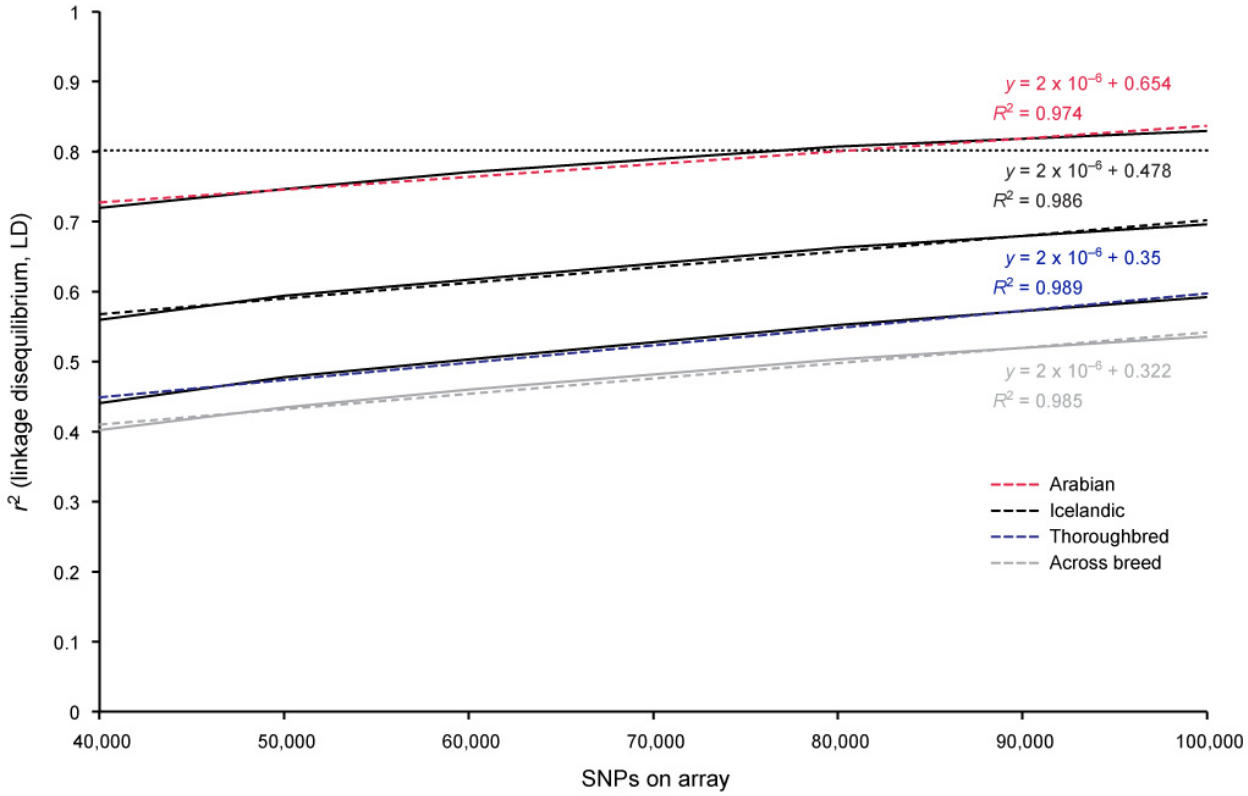

Fig. S8. SNP density required for genome-wide mapping Modeling suggests that a SNP genotyping density of 100,000 SNPs should enable association at mean maximum r^2>0.5 for most horse breeds as well as across breeds, but is insufficient to map with stringency of mean maximum r^2>0.8 for other than low effective population size breeds. The data was stochastically generated with hidden mutations within replicate (n=1000) maps generated with observed genotyping data at different densities.

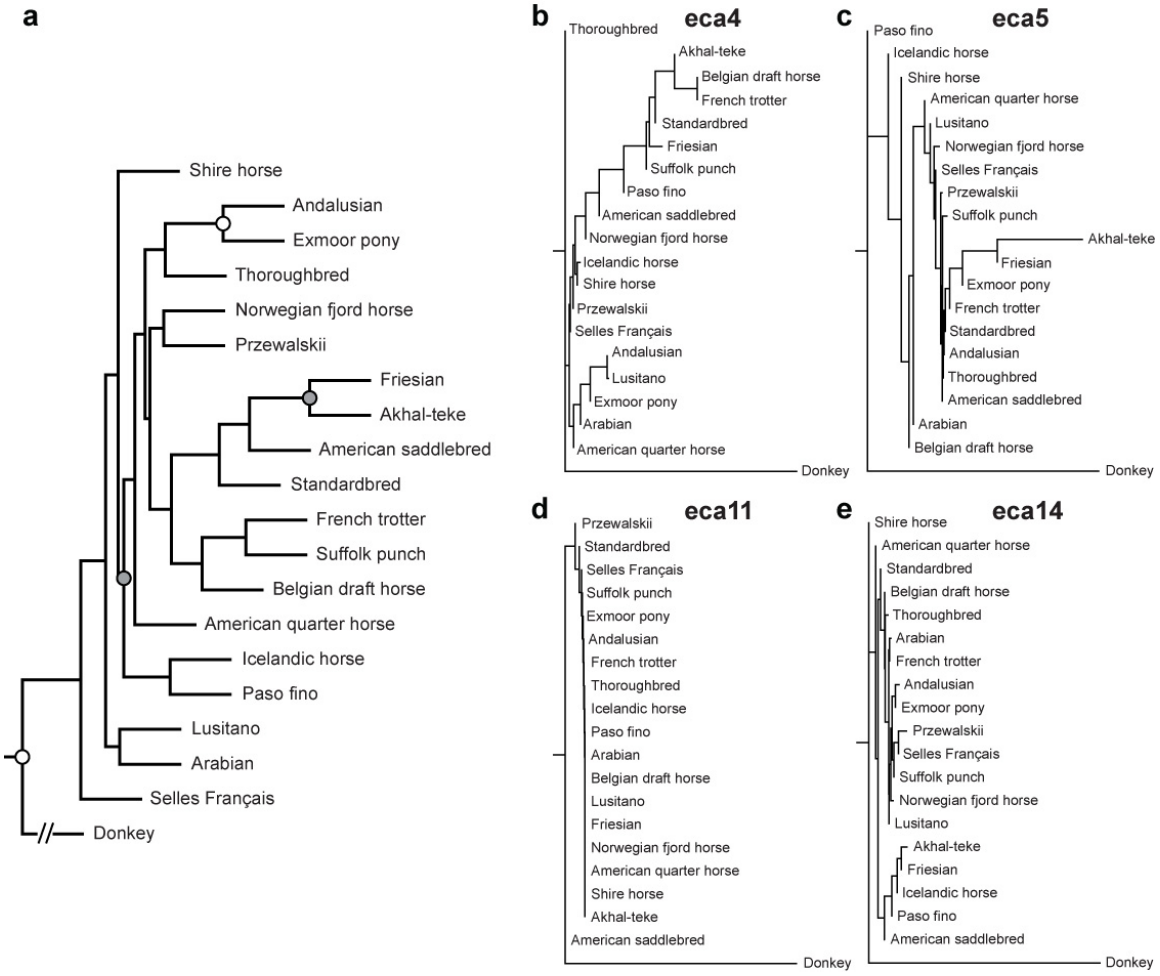

Fig. S9. Phylogeny of horse breeds at five random loci. (a) A phylogram among 19 horse breeds using all loci has few nodes with statistical support. The first supported node denotes separation from the donkey lineage. *E. przewalskii* does not separate from domestic horse breeds. White circles mark bootstrap values of greater than 95% from 1000 permutations, while gray circles denote those with values > 90%. (**b-e**) Phylograms from individual loci show that topologies among horse breeds are predominantly flat but are diverse across loci.

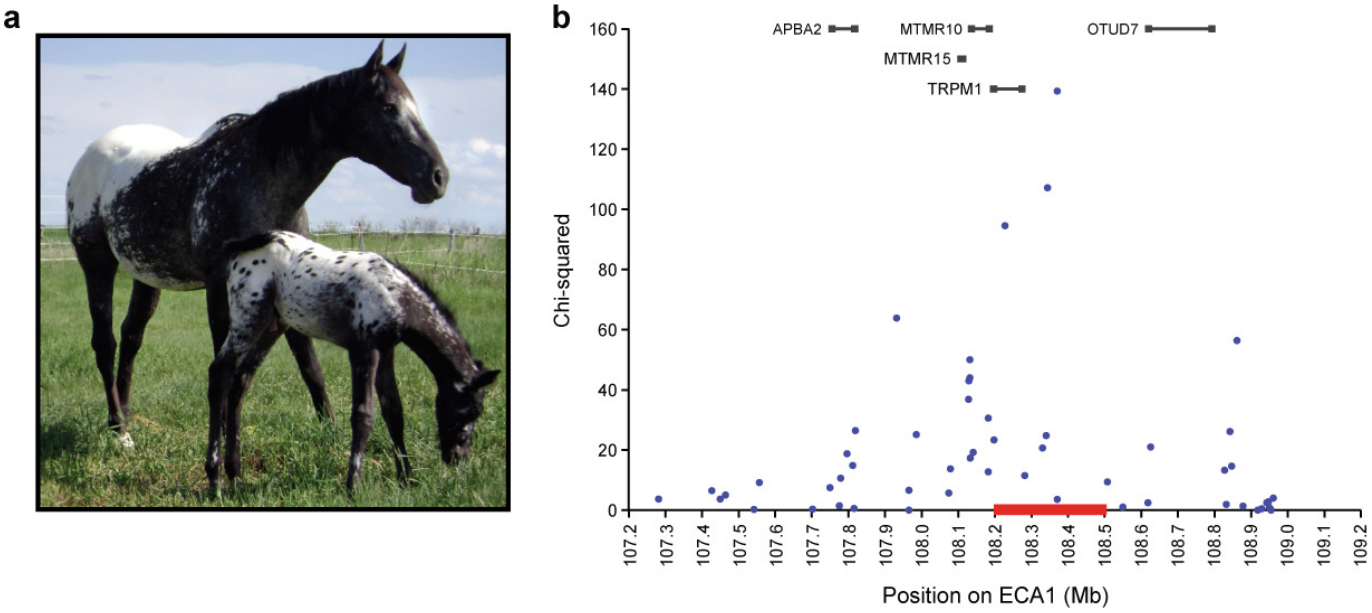

Fig. S10. Fine mapping of the Leopard Complex locus. (a) The mare represents the homozygous (few-spot and Congenital Stationary Night Blindness) Leopard Complex phenotype, while the foal represents the heterozygous phenotype. **(b)** Leopard Complex was fine mapped using 70 SNP over a 2Mb interval identifying a 173kb associated haplotype. A 300kb region was re-sequenced (red bar).

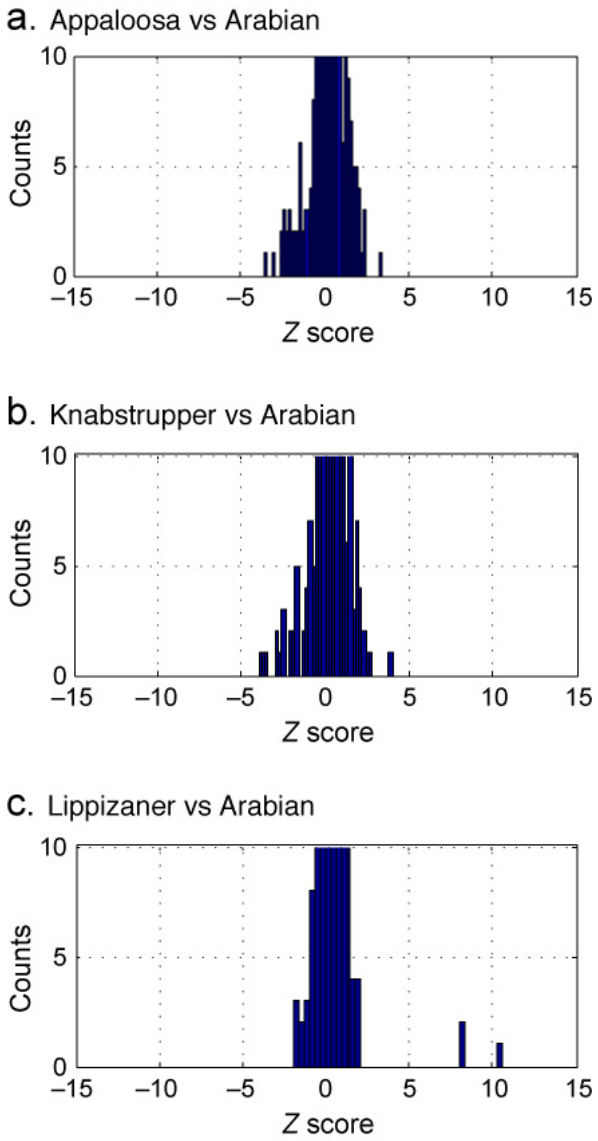

Fig. S11. CNV detection from resequencing data Z-scores for normalized coverage within 1kb windows are displayed in pair-wise comparisons of an Appaloosa (LP) versus Arabian (non-grey, non-LP, chestnut), Knabstrupper (LP) versus Arabian (non-grey, non-LP, chestnut) and Lippizaner (Grey) versus Arabian (non-grey, non-LP, chestnut). Only the Grey horse comparison reveals a deviation from the mean consistent with the presence of a copy number polymorphism.

Supporting Tables

Table S1. Sources of Samples

Strain	Number	Use †	Contributor
Twilight - Thoroughbred	1	GSDRL	Cornell University (USA)
Andalusion Horse	1	SDRL	University of Kentucky (USA)/Kentucky Horse Park
Icelandic Horse	1	SDRL	University of Kentucky (USA)/Kentucky Horse Park
Akhal teke	1	SDRL	University of Kentucky (USA)/Kentucky Horse Park
Thoroughbred Horse	1	SDRL	University of Kentucky (USA)
Standardbred-Pacer	1	SDRL	University of Kentucky (USA)
Quarter Horse	1	SDRL	University of California -Davis (USA)
Arabian Horse	1	SDRL	University of Kentucky (USA)/Kentucky Horse Park
Przewalskii Horse (Equid)	1	RL	San Diego Zoo's Institute for Conservation Research (USA)
Donkey (Equid)	1	RL	San Diego Zoo's Institute for Conservation Research (USA)
Paso Fino	1	RL	University of Kentucky (USA)/Kentucky Horse Park
Exmoor Pony	1	RL	University College of Dublin (Ireland)
Trakehner	1	RL	University of Kentucky (USA)/Kentucky Horse Park
Friesian	1	RL	University of Kentucky (USA)/Kentucky Horse Park
Belgian draft	1	RL	University of Minnesota (USA)
Norwegian Fjord	1	RL	University of Kentucky (USA)/Kentucky Horse Park
Shire	1	RL	University of Kentucky (USA)/Kentucky Horse Park
Suffolk Punch	1	RL	University of Kentucky (USA)/Kentucky Horse Park
Lusitano	1	RL	University of Kentucky (USA)/Kentucky Horse Park
American Saddlebred	1	RL	University of Kentucky (USA)
French Trotter	1	RL	INRA (France)
Selle Francais	1	RL	INRA (France)
Thoroughbred	24	L	University of Kentucky (USA)
Arabian	24	L	University of Kentucky (USA)
Icelandic Horse	24	L	Uppsala University (Sweden)
Hokkaido	24	L	Laboratory of Racing Chemistry (Japan)
Belgian draft	24	L	University of Minnesota (USA)
Quarter horse	24	L	University of California -Davis (USA)
French Trotter	24	L	INRA (France)
Standardbred-Pacer	24	L	University of Kentucky (USA)
Hanoverian	24	L	Institute of Animal Breeding and Genetics (Germany)
Norwegian Fjord	24	L	Norwegian School of Veterinary Science (Norway)
Andalusian	24	L	University of Kentucky (USA)
Przewalskii Horse	8	R	San Diego Zoo's Institute for Conservation Research (USA)

† Genome sequencing (G), SNP discovery (S), Re-sequencing (R), L=Linkage Disequilibrium and Haplotype analysis (L)

Table S2. Genome assembly characteristics

WGS Assembly (EquCab2.0)	
Number of sequence reads	28.8×10^6
Sequence redundancy (Q20 bases)	6.8×
Contig Length (kb; N50*)	112.4
Scaffold Length (Mb; N50)	46
Anchored bases in assembly (Mb)	2,336
Integration of physical mapping data	
Scaffolds anchored on chromosomes	83
Fraction of genome in anchored and oriented scaffolds (%)	95.6
Quality control	
Bases with quality score ≥ 40 (%)	98.6

*N50 is the size such that 50% of the assembly reside in contigs/scaffolds of a length at least X.

Table S3. Comparative assembly statistics

	Horse (EquCab2.0)	**Dog (CanFam2.0)**	**Cow (BTau4.0) †**
Coverage	6.8×	7.3×	7.1×
Total contig length (ungapped)	2.43Gb	2.31Gb	2.73Gb
Total scaffold length (gapped)	2.47Gb	2.33Gb	2.87Gb
N50 contig	112Kb	180Kb	49Kb
N50 scaffold	46Mb	45Mb	2Mb
Heterozygosity Rate within assembly*	~1/2000*	~1/1700*	1/1700
Anchored	96%	98%	90%
Unplaced (Gb)	0.26	0.075	0.28

* Adjusted for the SNP discovery protocol.
† ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20070913-freeze/README.Btau20070913.txt

Table S4. Occurrence of common equine repeat classes within chimeric repeat sequences by frequency

Repeat Class	Occurrence (N=2,229 chimeras)
L1	1,054
ERV	464
LTR	439
ERE	419
MER	406
MLT	247
MIR	242
L2	211
Tigger	70
SAT_EC	44
ES22	27
Total common	3,623

Table S5 Equine repeat counts by repeat class

Repeat_group	Count	Bases	% Genome
LINE_L1	464,621	402,200,888	16.25
DNA_All	330,947	77,908,269	3.15
LINE_L2	234,304	71,040,977	2.87
SINE_ERE1_2	296,962	70,493,153	2.85
LTR_MaLR	189,978	66,335,259	2.68
SINE_MIR	433,220	65,909,940	2.66
LTR_ERVL	97,995	45,135,591	1.82
LTR_ERV1	95,173	37,924,134	1.53
SINE_ERE3	159,847	35,928,664	1.45
tetra.penta_All	2,560,707	30,290,224	1.22
LINE_CR1	24,886	6,321,427	0.26
di_AC	455,500	5,154,925	0.21
di_AG	491,252	4,615,663	0.19
LINE_RTE	19,102	4,519,823	0.18
LTR_Other	16,778	4,205,795	0.17
di_AT	317,721	3,249,245	0.13
tri_AAT	239,773	2,478,174	0.10
tri_AAG	196,665	2,012,054	0.08
tri_AGG	182,780	1,830,576	0.07
LTR_ERVK	6,250	1,698,832	0.07
tri_AGC	118,372	1,179,278	0.05
tri_AAC	113,989	1,174,230	0.05
tri_ACC	97,191	972,959	0.04
tri_ATC	85,399	890,424	0.04
SINE_Other	6,507	647,743	0.03
SINE_tRNA	3,356	535,733	0.02
tri_ACT	29,303	284,872	0.01
tri_CCG	12,825	135,075	0.01
di_CG	6,671	60,736	0.00
tri_ACG	2,560	24,792	0.00
unclassified/chimeric		278,805,770	11.27
TOTAL		1,223,965,225	49.45

Table S6 Equine repeat counts by chromosome

Chromosome	Bases in Chromosome (Ungapped)	Bases in Custom Horse Repeat Library	%Bases	Bases in Repeat Standard mammalian repeat library	%Bases
ECA01	185,838,109	85,443,209	45.98	71,898,615	38.69
ECA02	120,857,687	54,761,556	45.31	47,449,429	39.26
ECA03	119,479,920	53,081,675	44.43	45,888,260	38.41
ECA04	108,569,075	48,749,160	44.90	40,502,306	37.31
ECA05	99,680,356	45,536,391	45.68	38,379,559	38.50
ECA06	84,719,076	37,006,521	43.68	31,281,116	36.92
ECA07	98,542,428	48,034,718	48.75	39,259,912	39.84
ECA08	94,057,673	42,512,680	45.20	35,105,738	37.32
ECA09	83,561,422	38,437,092	46.00	33,651,899	40.27
ECA10	83,980,604	41,053,406	48.88	32,849,434	39.12
ECA11	61,308,211	25,904,943	42.25	21,972,481	35.84
ECA12	33,091,231	15,989,901	48.32	12,126,721	36.65
ECA13	42,578,167	20,437,762	48.00	16,648,391	39.10
ECA14	93,904,894	42,251,798	44.99	36,215,084	38.57
ECA15	91,571,448	41,521,469	45.34	35,379,096	38.64
ECA16	87,365,405	39,087,829	44.74	34,037,732	38.96
ECA17	80,757,907	35,069,315	43.43	30,304,085	37.52
ECA18	82,527,541	36,606,822	44.36	31,370,240	38.01
ECA19	59,975,221	26,235,552	43.74	22,805,764	38.03
ECA20	64,166,202	30,104,769	46.92	24,406,416	38.04
ECA21	57,723,302	24,820,341	43.00	20,791,924	36.02
ECA22	49,946,797	22,919,109	45.89	19,839,539	39.72
ECA23	55,726,280	26,693,617	47.90	21,250,236	38.13
ECA24	46,749,900	20,575,606	44.01	17,427,933	37.28
ECA25	39,536,964	17,644,928	44.63	15,011,336	37.97
ECA26	41,866,177	17,923,392	42.81	14,932,430	35.67
ECA27	39,960,074	17,947,777	44.91	24,294,988	60.80
ECA28	46,177,339	20,647,313	44.71	17,968,238	38.91
ECA29	33,672,925	15,797,024	46.91	12,659,408	37.60
ECA30	30,062,385	13,484,851	44.86	11,530,197	38.35
ECA31	24,984,650	10,103,970	40.44	8,551,188	34.23
ECA X	124,114,077	73,541,234	59.25	59,996,902	48.34

Table S7. Comparative mammalian synteny by human

	Horse (EquCab2.0)	*Dog (Canfam2.0)*	*Cow (BTau4.0)†*	*Mouse (mm8)*
Total human bases within syntenic blocks	2,806,996,477	2,815,185,268	2,740,814,367	2,782,678,976
Number of syntenic blocks to human build 36	143	204	666	267
Mean size of syntenic block (human bp)	19,629,346	13,799,928	4,115,337	10,422,019
N50 size of syntenic block (bp)	49,960,324	29,533,072	13,931,946	24,094,919
Min size of syntenic block (human bp)	76,610	42,194	21,727	40,899
Max size of syntenic block (human bp)	111,937,458	123,805,846	88,057,757	145,402,504
Number of syntenic segments to human build 36	425	443	1,257	528
Mean size of syntenic segment (human bp)	6,494,112	6,245,596	2,130,424	5,176,048
N50 size of syntenic segment (human bp)	28,586,459	20,303,665	6,659,535	14,180,003
Min size of syntenic segment (human bp)	20,829	33,931	15,852	23,841
Max size of syntenic segment (human bp)	89,351,187	49,342,126	33,132,642	59,405,417

† Differences are likely technical rather than biological

Table S8. Results of hybridization with 37cen and 2P1 (the two major horse satellites).

Chr	37cen	2pl	Chr	37cen	2pl
ECA1	C†	/‡	ECA18	C	C
ECA2	/	C	ECA19	C	C
ECA3	C	C	ECA20	C	C
ECA4	C	/	ECA21	C	C
ECA5	C	/	ECA22	C	C
ECA6	C	C	ECA23	C	C
ECA7	C	C	ECA24	C	C
ECA8	C	C	ECA25	C	C
ECA9	C	/	ECA26	C	C
ECA10	C	C	ECA27	C	C
ECA11	/	/	ECA28	C	C
ECA12	C	/	ECA29	C	C
ECA13	C	C	ECA30	C	C
ECA14	C	C	ECA31	C	C
ECA15	C	C	ECAX	C	/
ECA16	C	C	ECAY	C	C
ECA17	C	C			

† Hybridized (C) ‡ No hybridization (/)

Table S9. Gene family expansion analysis for horse. Ensemble gene codes for genes with more than two-fold as many paralogs in horse as human and cow.

Horse ID	Human ID	Paralogs Horse	Paralogs Human	Paralogs Cow	GO Biological Process	GO Molecular Function	Names List	Ensembl annotation (human ID)
ENSECAG00000006092	ENSG00000226430	240	29	96		ligand-dependent nuclear receptor transcription coactivator activity,hydrolase activity,	XP_001494472.1USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33|USP42|USP36|USP22|USP49|A5YBM4_HORSE|XP_001494487.2|USP20|USP16|XP_001487991.1|USP44|USP45|USP33	Inactive ubiquitin carboxyl-terminal hydrolase 17-like protein 7
ENSECAG00000000647	ENSG00000223667	234	29	110			Names not provided	T-cell receptor beta V gene segment

ENSECAG00000007679	ENSG00000177693	213	29	56	XP_001502249.2XP_001918312.1|XP_001501803.2|XP_001502099.2||XP_001502096.2|XP_001501866.2|XP_001501873.2|XP_001501848.2|XP_001501813.2||XP_001502112.2|XP_001501822.2||XP_001502058.2|XP_001501914.2|XP_001502086.2||XP_001501894.2|XP_001502105.2|XP_001501936.2|XP_001502037.2|XP_001502336.2|OR4F15|XP_001501838.2|XP_001502322.2|XP_001502119.2|XP_001502447.2|XP_001502158.2|OR4K14|XP_001502033.2|XP_001501993.2|OR4K13|XP_001502032.2|OR4K5|XP_001502137.2|XP_001918302.1|XP_001502027.2|XP_001502288.2|XP_001501905.2|XP_001502299.2|XP_001502052.2|XP_001501918.2|XP_001502123.2|XP_001501902.2|OR4K15||XP_001502061.2|OR4K1|XP_001502065.2|OR4L1|XP_001502452.2|OR4K2|OR4Q3|XP_001500617.2|OR4D2|XP_001501811.2|OR4D1||XP_001492808.2|OR4D11|OR4D6|XP_001502465.2|OR4D9|XP_001502457.2|OR4N5|XP_001495003.1|XP_001501966.2|OR4N4|OR4Q2|XP_001504950.2|XP_001916079.1|XP_001918312.1|XP_001501803.2|XP_001502099.2||XP_001502096.2|XP_001501866.2|XP_001501873.2|XP_001501848.2|XP_001501813.2||XP_001502112.2|XP_001501822.2||XP_001502058.2|XP_001501914.2|XP_001502086.2||XP_001501894.2|XP_001502105.2|XP_001501936.2|XP_001502037.2|XP_001502336.2|OR4F15|XP_001501838.2|XP_001502322.2|XP_001502119.2|XP_001502447.2|XP_001502158.2|OR4K14|XP_001502033.2|XP_001501993.2|OR4K13|XP_001502032.2|OR4K5|XP_001502137.2|XP_001918302.1|XP_001502027.2|XP_001502288.2|XP_001501905.2|XP_001502299.2|XP_001502052.2|XP_001501918.2|XP_001502123.2|XP_001501902.2|OR4K15||XP_001502061.2|OR4K1|XP_001502065.2|OR4L1|XP_001502452.2|OR4K2|OR4Q3|XP_001500617.2|OR4D2|XP_001501811.2|OR4D1||XP_001492808.2|OR4D11|OR4D6|XP_001502465.2|OR4D9|XP_001502457.2|OR4N5|XP_001495003.1|XP_001501966.2|OR4N4|OR4Q2|XP_001504950.2|XP_001916079.1|	Olfactory receptor type 4

ENSECAG00000005288	ENSG00000196143	156	21	76	XP_001504682.2XP_001505176.2|XP_001502381.2|XP_001502514.2|XP_001502347.2|XP_001502579.2|XP_001502559.2||XP_001502230.2||XP_001502365.2|XP_001502217.2|XP_001502352.2|XP_001502530.2|OR11H6|XP_001502521.2|OR11H4|XP_001501945.2|XP_001505178.2|XP_001502338.2|XP_001502499.2|XP_001502541.2|XP_001502549.2|XP_001497033.2|XP_001501802.2||XP_001491504.2|XP_001490710.2|OR6Q1|OR6N1|XP_001497273.2|XP_001500666.2|OR6B1|OR6A2|XP_001499903.2|XP_001500074.2|OR6N2|XP_001500062.2||XP_001491710.2|XP_001491689.2|OR6P1|OR6Y1|OR6K6||OR6K2|XP_001915391.1|XP_001915385.1|OR6K3||XP_001495672.2|OR2AT4||XP_001505176.2|XP_001502381.2|XP_001502514.2|XP_001502347.2|XP_001502579.2|XP_001502559.2||XP_001502230.2||XP_001502365.2|XP_001502217.2|XP_001502352.2|XP_001502530.2|OR11H6|XP_001502521.2|OR11H4|XP_001501945.2|XP_001505178.2|XP_001502338.2|XP_001502499.2|XP_001502541.2|XP_001502549.2|XP_001497033.2|XP_001501802.2||XP_001491504.2|XP_001490710.2|OR6Q1|OR6N1|XP_001497273.2|XP_001500666.2|OR6B1|OR6A2|XP_001499903.2|XP_001500074.2|OR6N2|XP_001500062.2||XP_001491710.2|XP_001491689.2|OR6P1|OR6Y1|OR6K6||OR6K2|XP_001915391.1|XP_001915385.1|OR6K3||XP_001495672.2|OR2AT4||XP_001505176.2|XP_001502381.2|XP_001502514.2|XP_001502347.2|XP_001502579.2|XP_001502559.2||XP_001502230.2||XP_001502365.2|XP_001502217.2|XP_001502352.2|XP_001502530.2|OR11H6|XP_001502521.2|OR11H4|XP_001501945.2|XP_001505178.2|XP_001502338.2|XP_001502499.2|XP_001502541.2|XP_001502549.2|XP_001497033.2|XP_001501802.2||XP_001491504.2|XP_001490710.2|OR6Q1|OR6N1|XP_001497273.2|XP_001500666.2|OR6B1|OR6A2|XP_001499903.2|XP_001500074.2|OR6N2|XP_001500062.2||XP_001491710.2|XP_001491689.2|OR6P1|OR6Y1|OR6K6||OR6K2|XP_001915391.1|XP_001915385.1|OR6K3||XP_001495672.2|OR2AT4||XP_001491303.2|OR2A5|XP_001490956.2|XP_001504269.2|XP_001491009.2|XP_001491250.2||||OR2A25|OR2A4|XP_001497094.2|XP_001915816.1|XP_001503071.2||XP_001490735.2||XP_001490644.2|XP_001499876.2|XP_001497007.2|XP_001490902.2|XP_001489054.2|XP_001501467.2|OR2D3|XP_001490706.2|XP_001500032.2|XP_001917888.1|XP_001500126.2|XP_001497063.2|XP_001504662.2||XP_001490479.2|XP_001490434.2|XP_001490460.2|XP_001490333.2|XP_001499885.2||XP_001491303.2|OR2A5|XP_001490956.2|XP_001504269.2|XP_001491009.2|XP_001491250.2||||OR2A25|OR2A4|XP_001497094.2|XP_001915816.1|XP_001503071.2||XP_001490735.2||XP_001490644.2|XP_001499876.2|XP_001497007.2|XP_001490902.2|XP_001489054.2|XP_001501467.2|OR2D3|XP_001490706.2|XP_001500032.2|XP_001917888.1|XP_001500126.2|XP_001497063.2|XP_001504662.2||XP_001490479.2|XP_001490434.2|XP_001490460.2|XP_001490333.2|XP_001499885.2||	Olfactory receptor type 6
ENSECAG00000007285	ENSG00000228960	152	13	60	XP_001504682.2XP_001505176.2|XP_001502381.2|XP_001502514.2|XP_001502347.2|XP_001502579.2|XP_001502559.2||XP_001502230.2||XP_001502365.2|XP_001502217.2|XP_001502352.2|XP_001502530.2|OR11H6|XP_001502521.2|OR11H4|XP_001501945.2|XP_001505178.2|XP_001502338.2|XP_001502499.2|XP_001502541.2|XP_001502549.2|XP_001497033.2|XP_001501802.2||XP_001491504.2|XP_001490710.2|OR6Q1|OR6N1|XP_001497273.2|XP_001500666.2|OR6B1|OR6A2|XP_001499903.2|XP_001500074.2|OR6N2|XP_001500062.2||XP_001491710.2|XP_001491689.2|OR6P1|OR6Y1|OR6K6||OR6K2|XP_001915391.1|XP_001915385.1|OR6K3||XP_001495672.2|OR2AT4||XP_001491303.2|OR2A5|XP_001490956.2|XP_001504269.2|XP_001491009.2|XP_001491250.2||||OR2A25|OR2A4|XP_001497094.2|XP_001915816.1|XP_001503071.2||XP_001490735.2||XP_001490644.2|XP_001499876.2|XP_001497007.2|XP_001490902.2|XP_001489054.2|XP_001501467.2|OR2D3|XP_001490706.2|XP_001500032.2|XP_001917888.1|XP_001500126.2|XP_001497063.2|XP_001504662.2||XP_001490479.2|XP_001490434.2|XP_001490460.2|XP_001490333.2|XP_001499885.2||XP_001491303.2|OR2A5|XP_001490956.2|XP_001504269.2|XP_001491009.2|XP_001491250.2||||OR2A25|OR2A4|XP_001497094.2|XP_001915816.1|XP_001503071.2||XP_001490735.2||XP_001490644.2|XP_001499876.2|XP_001497007.2|XP_001490902.2|XP_001489054.2|XP_001501467.2|OR2D3|XP_001490706.2|XP_001500032.2|XP_001917888.1|XP_001500126.2|XP_001497063.2|XP_001504662.2||XP_001490479.2|XP_001490434.2|XP_001490460.2|XP_001490333.2|XP_001499885.2||	Olfactory receptor type 2

ENSECAG00000000825	ENSG00000215550	145	33	28		protein binding,structural constituent of cytoskeleton,structural molecule activity	XP_001504594.1|XP_001917443.1|XP_001496432.2|XP_001496861.1|KRT14|XP_001496491.1|KRT13|KRT16|KRT35|KRT36|KRT32|XP_001497057.1|O62661_HORSE|XP_001497111.2|XP_001917457.1|KRT23|KRT24|KRT20|KRT12|KRT40|KRT39|XP_001497020.1||XP_001918028.1|KRT28|KRT25|XP_001497038.1|KRT27|KRT26||XP_001917443.1|XP_001496432.2|XP_001496861.1|KRT14|XP_001496491.1|KRT13|KRT16|KRT35|KRT36|KRT32|XP_001497057.1|O62661_HORSE|XP_001497111.2|XP_001917457.1|KRT23|KRT24|KRT20|KRT12|KRT40|KRT39|XP_001497020.1||XP_001918028.1|KRT28|KRT25|XP_001497038.1|KRT27|KRT26||XP_001917443.1|XP_001496432.2|XP_001496861.1|KRT14|XP_001496491.1|KRT13|KRT16|KRT35|KRT36|KRT32|XP_001497057.1|O62661_HORSE|XP_001497111.2|XP_001917457.1|KRT23|KRT24|KRT20|KRT12|KRT40|KRT39|XP_001497020.1||XP_001918028.1|KRT28|KRT25|XP_001497038.1|KRT27|KRT26||XP_001917443.1|XP_001496432.2|XP_001496861.1|KRT14|XP_001496491.1|KRT13|KRT16|KRT35|KRT36|KRT32|XP_001497057.1|O62661_HORSE|XP_001497111.2|XP_001917457.1|KRT23|KRT24|KRT20|KRT12|KRT40|KRT39|XP_001497020.1||XP_001918028.1|KRT28|KRT25|XP_001497038.1|KRT27|KRT26|	Keratin (via paralog ID)
ENSECAG00000000871	ENSG00000138115	120	16	54	oxidation reduction	metal ion binding,Vitamin D3 25-hydroxylase activity	A7LGW8_HORSE|XP_001502306.2|XP_001500795.1|XP_001502229.2|XP_001502157.2|XP_001502043.1|XP_001502080.1|XP_001500623.2|A8WDL7_HORSE|CYP2G1P|XP_001916351.1|A9UHN2_HORSE|CYP2F1|XP_001498473.2|XP_001498439.1|||CYP2S1|CYP2J2|A8W4X2_HORSE|XP_001502906.1||XP_001502857.2|||CYP2U1|CYP2R1||CYP2W1||XP_001502306.2|XP_001500795.1|XP_001502229.2|XP_001502157.2|XP_001502043.1|XP_001502080.1|XP_001500623.2|A8WDL7_HORSE|CYP2G1P|XP_001916351.1|A9UHN2_HORSE|CYP2F1|XP_001498473.2|XP_001498439.1|||CYP2S1|CYP2J2|A8W4X2_HORSE|XP_001502906.1||XP_001502857.2|||CYP2U1|CYP2R1||CYP2W1||XP_001502306.2|XP_001500795.1|XP_001502229.2|XP_001502157.2|XP_001502043.1|XP_001502080.1|XP_001500623.2|A8WDL7_HORSE|CYP2G1P|XP_001916351.1|A9UHN2_HORSE|CYP2F1|XP_001498473.2|XP_001498439.1|||CYP2S1|CYP2J2|A8W4X2_HORSE|XP_001502906.1||XP_001502857.2|||CYP2U1|CYP2R1||CYP2W1||XP_001502306.2|XP_001500795.1|XP_001502229.2|XP_001502157.2|XP_001502043.1|XP_001502080.1|XP_001500623.2|A8WDL7_HORSE|CYP2G1P|XP_001916351.1|A9UHN2_HORSE|CYP2F1|XP_001498473.2|XP_001498439.1|||CYP2S1|CYP2J2|A8W4X2_HORSE|XP_001502906.1||XP_001502857.2|||CYP2U1|CYP2R1||CYP2W1|	Cytochrome P450 2C8 (EC 1.14.14.1)(CYPIIC8)(P450 form 1)(P450 MP-12/MP-20)(P450 IIC2)(S-mephenytoin 4-hydroxylase)

ENSECAG00000021518	ENSG00000213648	64	12	14		sulfotransferase activity	XP_001502088.1SULT1C3|Q0NXS2_HORSE|SULT1C4|SULT1C2|SULT4A1|SULT2B1|XP_001504181.2|XP_001504183.1|XP_001917522.1|XP_001487921.1|XP_001488018.2|XP_001487960.2|XP_001498245.1|SULT6B1|SULT1C3|Q0NXS2_HORSE|SULT1C4|SULT1C2|SULT4A1|SULT2B1|XP_001504181.2|XP_001504183.1|XP_001917522.1|XP_001487921.1|XP_001488018.2|XP_001487960.2|XP_001498245.1|SULT6B1|SULT1C3|Q0NXS2_HORSE|SULT1C4|SULT1C2|SULT4A1|SULT2B1|XP_001504181.2|XP_001504183.1|XP_001917522.1|XP_001487921.1|XP_001488018.2|XP_001487960.2|XP_001498245.1|SULT6B1|	Sulfotransferase 1A3/1A4 (EC 2.8.2.1)(Monoamine-sulfating phenol sulfotransferase)(Aryl sulfotransferase 1A3/1A4)(Sulfotransferase, monoamine-preferring)(M-PST)(Thermolabile phenol sulfotransferase)(TL-PST)(Placental estrogen sulfotransferase)(Catecholamine-sulfating phenol sulfotransferase)(HAST3)
ENSECAG00000018967	ENSG00000113262	63	8	16	learning	group II metabotropic glutamate receptor activity,calcium channel regulator activity,protein binding	XP_001492696.2XP_001502361.1|XP_001488463.1|XP_001495009.1|XP_001498083.2|XP_001916969.1|XP_001496052.2|XP_001502552.2|XP_001502361.1|XP_001488463.1|XP_001495009.1|XP_001498083.2|XP_001916969.1|XP_001496052.2|XP_001502552.2|XP_001502361.1|XP_001488463.1|XP_001495009.1|XP_001498083.2|XP_001916969.1|XP_001496052.2|XP_001502552.2|XP_001502361.1|XP_001488463.1|XP_001495009.1|XP_001498083.2|XP_001916969.1|XP_001496052.2|XP_001502552.2|XP_001502361.1|XP_001488463.1|XP_001495009.1|XP_001498083.2|XP_001916969.1|XP_001496052.2|XP_001502552.2|XP_001502361.1|XP_001488463.1|XP_001495009.1|XP_001498083.2|XP_001916969.1|XP_001496052.2|XP_001502552.2|XP_001502361.1|XP_001488463.1|XP_001495009.1|XP_001498083.2|XP_001916969.1|XP_001496052.2|XP_001502552.2|XP_001502361.1|XP_001488463.1|XP_001495009.1|XP_001498083.2|XP_001916969.1|XP_001496052.2|XP_001502552.2|	Metabotropic glutamate receptor 6 Precursor (mGluR6)
ENSECAG00000005188	ENSG00000215177	62	12	24			XP_001502963.2|XP_001494079.2|XP_001492611.2|XP_001493264.2|XP_001492990.1|XP_001494033.2||||XP_001915433.1|||XP_001488407.1|||XP_001493907.1|VPREB3||||XP_001502963.2|XP_001494079.2|XP_001492611.2|XP_001493264.2||XP_001492990.1|XP_001494033.2|XP_001915433.1|||XP_001488407.1|||XP_001493907.1|VPREB3|	Putative uncharacterized protein
ENSECAG00000021419	ENSG00000161643	54	13	10		protein binding	XP_001917550.1XP_001499308.1|XP_001496653.2|XP_001497053.2|XP_001917576.1|XP_001497072.1|MAG|SIGLEC1|CD22|XP_001499308.1|XP_001496653.2|XP_001497053.2|XP_001917576.1|XP_001497072.1|MAG|SIGLEC1|CD22|XP_001499308.1|XP_001496653.2|XP_001497053.2|XP_001917576.1|XP_001497072.1|MAG|SIGLEC1|CD22|	Sialic acid-binding Ig-like lectin 16 Precursor (Siglec-16)(Siglec-P16)

ENSECAG00000014464	ENSG00000203820	39	15	12	antigen processing and presentation of exogenous antigen	IgG receptor activity,IgE binding	XP_001916290.1|Q8MI30_HORSE|XP_001916314.1|XP_001916335.1|XP_001500974.2|FCRL6|FCRL5|FCRL4|FCRL3|FCRLA|FCRLB|Q8MI30_HORSE|XP_001916314.1|XP_001916335.1|XP_001500974.2|FCRL6|FCRL5|FCRL4|FCRL3|FCRLA|FCRLB|Q8MI30_HORSE|XP_001916314.1|XP_001916335.1|XP_001500974.2|FCRL6|FCRL5|FCRL4|FCRL3|FCRLA|FCRLB|	Putative high affinity immunoglobulin gamma Fc receptor IC Precursor (IgG Fc receptor IC)(Fc-gamma RIC)(hFcgammaR1C)(FcRIC)
ENSECAG00000000954	ENSG00000231819	39	15	13		structural molecule activity,protein binding,chaperone binding,	Q65YZ3_HORSE|MUC5B|VWF|MUC19|MUC2|MUC6|FCGBP|TECTA|BMPER|C12orf64|KCP|Q65YZ3_HORSE|MUC5B|VWF|MUC19|MUC2|MUC6|FCGBP|TECTA|BMPER|C12orf64|KCP|Q65YZ3_HORSE|MUC5B|VWF|MUC19|MUC2|MUC6|FCGBP|TECTA|BMPER|C12orf64|KCP|	Unknown
ENSECAG00000021048	ENSG00000165092	42	6	16		,oxidoreductase activity,3-chloroallyl aldehyde dehydrogenase activity	XP_001914883.1XP_001498666.1|ALDH2_HORSE|XP_001500420.1|XP_001915212.1|AL1A1_HORSE|XP_001490605.2|XP_001498666.1|ALDH2_HORSE|XP_001500420.1|XP_001915212.1|AL1A1_HORSE|XP_001490605.2|XP_001498666.1|ALDH2_HORSE|XP_001500420.1|XP_001915212.1|AL1A1_HORSE|XP_001490605.2|XP_001498666.1|ALDH2_HORSE|XP_001500420.1|XP_001915212.1|AL1A1_HORSE|XP_001490605.2|XP_001498666.1|ALDH2_HORSE|XP_001500420.1|XP_001915212.1|AL1A1_HORSE|XP_001490605.2|XP_001498666.1|ALDH2_HORSE|XP_001500420.1|XP_001915212.1|AL1A1_HORSE|XP_001490605.2|XP_001498666.1|ALDH2_HORSE|XP_001500420.1|XP_001915212.1|AL1A1_HORSE|XP_001490605.2|	Retinal dehydrogenase 1 (RALDH 1)(RalDH1)(EC 1.2.1.36)(Aldehyde dehydrogenase family 1 member A1)(Aldehyde dehydrogenase, cytosolic)(ALHDII)(ALDH-E1
ENSECAG00000007557	ENSG00000188755	45	13	5			XP_001489584.2USP6NL|TBC1D10A|TBC1D10B|TBC1D10C|GRTP1|USP6NL|TBC1D10A|TBC1D10B|TBC1D10C|GRTP1|USP6NL|TBC1D10A|TBC1D10B|TBC1D10C|GRTP1|USP6NL|TBC1D10A|TBC1D10B|TBC1D10C|GRTP1|USP6NL|TBC1D10A|TBC1D10B|TBC1D10C|GRTP1|USP6NL|TBC1D10A|TBC1D10B|TBC1D10C|GRTP1|USP6NL|TBC1D10A|TBC1D10B|TBC1D10C|GRTP1|USP6NL|TBC1D10A|TBC1D10B|TBC1D10C|GRTP1|	Putative uncharacterized protein TBC1D3P2
ENSECAG00000011139	ENSG00000089335	48	23	22			XP_001491130.2ZNF570|ZNF470|ZNF471|ZNF583|XP_001495178.1|ZNF383|ZNF566|ZNF345|ZNF461|ZNF565|ZFP14|ZNF527|ZFP82|ZFP30|ZNF790|ZNF331|ZFP28|ZNF582|ZNF404|XP_001500066.2|ZNF570|ZNF470|ZNF471|ZNF583|XP_001495178.1|ZNF383|ZNF566|ZNF345|ZNF461|ZNF565|ZFP14|ZNF527|ZFP82|ZFP30|ZNF790|ZNF331|ZFP28|ZNF582|ZNF404|XP_001500066.2|	Zinc finger protein 302 (ZNF135-like)(ZNF140-like)
ENSECAG00000005670	ENSG00000102128	51	19	17		protein binding,GTPase activity,myosin V binding,	XP_001490311.2RAB40C|RAB8A|RAB8B|RAB13|RAB15|RAB10|RAB3A|RAB3B|RAB3D|RAB26|RAB3C|RAB37|RAB27A|RAB27B|RASEF||RAB44|RAB40C|RAB8A|RAB8B|RAB13|RAB15|RAB10|RAB3A|RAB3B|RAB3D|RAB26|RAB3C|RAB37|RAB27A|RAB27B|RASEF||RAB44|RAB40C|RAB8A|RAB8B|RAB13|RAB15|RAB10|RAB3A|RAB3B|RAB3D|RAB26|RAB3C|RAB37|RAB27A|RAB27B|RASEF||RAB44|	Ras-related protein Rab-40A-like (Ras-like GTPase)

ENSECAG00000020129	ENSG00000150165	36	13	11		protein binding,calcium-dependent protein binding,eukaryotic cell surface binding,cytoskeletal protein binding,phospholipase A2 inhibitor activity,acetylcholine receptor activity,	XP_001500744.1ANXA6|ANXA4|ANXA11|ANXA5|B2B9I8_HORSE|ANXA3|ANXA1_HORSE|XP_001917046.1|ANXA13|ANXA10|ANXA7|ANXA9|ANXA6|ANXA4|ANXA11|ANXA5|B2B9I8_HORSE|ANXA3|ANXA1_HORSE|XP_001917046.1|ANXA13|ANXA10|ANXA7|ANXA9|ANXA6|ANXA4|ANXA11|ANXA5|B2B9I8_HORSE|ANXA3|ANXA1_HORSE|XP_001917046.1|ANXA13|ANXA10|ANXA7|ANXA9|	Annexin A8-like protein 1
ENSECAG00000018815	ENSG00000225766	36	13	10	retinal metabolic process	retinal dehydrogenase activity,testosterone 17-beta-dehydrogenase (NADP+) activity,NAD or NADH binding,NADPH binding,protein binding	XP_001489425.2XP_001489532.2|DHRS2|HSD17B14|HSD17B8|BDH2|CBR4|DECR1|HSD17B10|A2T0Y1_HORSE|DCXR|PECR|NME4|XP_001489532.2|DHRS2|HSD17B14|HSD17B8|BDH2|CBR4|DECR1|HSD17B10|A2T0Y1_HORSE|DCXR|PECR|NME4|XP_001489532.2|DHRS2|HSD17B14|HSD17B8|BDH2|CBR4|DECR1|HSD17B10|A2T0Y1_HORSE|DCXR|PECR|NME4|	dehydrogenase/reductase (SDR family) member 4 like 1
ENSECAG00000012835	ENSG00000140839	36	13	11		heparin binding,molecular_function	XP_001498514.2CRISPLD2|PI15|CRISPLD1|R3HDML|Q8HX97_HORSE|CRIS3_HORSE|PI16|GLIPR1|GLIPR1L2|GLIPR1L1|Q8MHX2_HORSE|CRISPLD2|PI15|CRISPLD1|R3HDML|Q8HX97_HORSE|CRIS3_HORSE|PI16|GLIPR1|GLIPR1L2|GLIPR1L1|Q8MHX2_HORSE|CRISPLD2|PI15|CRISPLD1|R3HDML|Q8HX97_HORSE|CRIS3_HORSE|PI16|GLIPR1|GLIPR1L2|GLIPR1L1|Q8MHX2_HORSE|	C-type lectin domain family 18 member B Precursor (Mannose receptor-like protein 1)
ENSECAG00000010859	ENSG00000226650	36	17	16		protein binding,protein kinase binding,microtubule motor activity	KIF21B|KIF21A|KIF27|KIF3A|KIF3C|KIF3B|CENPE|KIF17|KIF5C|KIF5A|KIF5B|KIF15|KIF9|KIF6|KIF11|KIF7|KIF21B|KIF21A|KIF27|KIF3A|KIF3C|KIF3B|CENPE|KIF17|KIF5C|KIF5A|KIF5B|KIF15|KIF9|KIF6|KIF11|KIF7|	Chromosome-associated kinesin KIF4B (Chromokinesin-B)

ENSECAG00000024824	ENSG00000131355	36	8	11		XP_001497658.2|XP_001491463.2|EMR1|XP_001495406.2|XP_001495362.2|XP_001495655.2|XP_001495255.2|XP_001491463.2|EMR1|XP_001495406.2|XP_001495362.2|XP_001495655.2|XP_001495255.2|XP_001491463.2|EMR1|XP_001495406.2|XP_001495362.2|XP_001495655.2|XP_001495255.2|	EGF-like module-containing mucin-like hormone receptor-like 3 Precursor (EGF-like module-containing mucin-like receptor EMR3)
ENSECAG00000008594	ENSG00000186470	33	13	14		|BTN1A1|XP_001492580.2|XP_001492552.2|XP_001917154.1|XP_001492171.2|BTNL2||BTNL9|MOG|ERMAP||BTN1A1|XP_001492580.2|XP_001492552.2|XP_001917154.1|XP_001492171.2|BTNL2||BTNL9|MOG|ERMAP||BTN1A1|XP_001492580.2|XP_001492552.2|XP_001917154.1|XP_001492171.2|BTNL2||BTNL9|MOG|ERMAP|	Butyrophilin subfamily 3 member A2 Precursor
ENSECAG00000023574	ENSG00000149742	33	11	13	organic anion transmembrane transporter activity	XP_001495240.2|XP_001489890.2|XP_001495264.1|XP_001495463.1|SLC22A8|XP_001502995.1|XP_001918258.1|XP_001488889.1|XP_001495240.2|XP_001489890.2|XP_001495264.1|XP_001495463.1|SLC22A8|XP_001502995.1|XP_001918258.1|XP_001488889.1|XP_001495240.2|XP_001489890.2|XP_001495264.1|XP_001495463.1|SLC22A8|XP_001502995.1|XP_001918258.1|XP_001488889.1|	Solute carrier family 22 member 9 (Organic anion transporter 4)
ENSECAG00000000709	ENSG00000165246	30	14	14	protein binding,neurexin binding,hydrolase activity,acetylcholinesterase activity	XP_001488067.1NLGN3|NLGN2|NLGN1|CHLE_HORSE|CEL|ACHE|CES2|XP_001491752.1|XP_001491576.2|XP_001491978.1|XP_001491160.1|CES3|CES7|Q95M80_HORSE|NLGN3|NLGN2|NLGN1|CHLE_HORSE|CEL|ACHE|CES2|XP_001491752.1|XP_001491576.2|XP_001491978.1|XP_001491160.1|CES3|CES7|Q95M80_HORSE|	Neuroligin-4, Y-linked Precursor (Neuroligin Y)
ENSECAG00000019932	ENSG00000171916	30	9	7	protein binding	XP_001915607.1|LGALS12|LGALS4|LGALS8|XP_001496714.1|LGALS3|XP_001496731.1|XP_001494811.2|XP_001915089.1|XP_001915607.1|LGALS12|LGALS4|LGALS8|XP_001496714.1|LGALS3|XP_001496731.1|XP_001494811.2|XP_001915089.1|XP_001915607.1|LGALS12|LGALS4|LGALS8|XP_001496714.1|LGALS3|XP_001496731.1|XP_001494811.2|XP_001915089.1|	Galectin-9C (Galectin-9-like protein B)
ENSECAG00000007548	ENSG00000189052	28	9	8	hormone activity	LSHB_HORSEFSHB_HORSE|TSHB_HORSE|GPHB5|FSHB_HORSE|TSHB_HORSE|GPHB5|FSHB_HORSE|TSHB_HORSE|GPHB5|FSHB_HORSE|TSHB_HORSE|GPHB5|FSHB_HORSE|TSHB_HORSE|GPHB5|FSHB_HORSE|TSHB_HORSE|GPHB5|	Choriogonadotropin subunit beta Precursor (CG-beta)(Chorionic gonadotrophin chain beta)
ENSECAG00000013008	ENSG00000205754	27	11	9	sodium-independent organic anion transmembrane transporter activity	XP_001498086.1Q206K0_HORSE|SLCO1C1|SLCO3A1|SLCO2A1|Q30BK7_HORSE|SLCO5A1|SLCO4A1|SLCO4C1|SLCO6A1|Q206K0_HORSE|SLCO1C1|SLCO3A1|SLCO2A1|Q30BK7_HORSE|SLCO5A1|SLCO4A1|SLCO4C1|SLCO6A1|Q206K0_HORSE|SLCO1C1|SLCO3A1|SLCO2A1|Q30BK7_HORSE|SLCO5A1|SLCO4A1|SLCO4C1|SLCO6A1|	liver-specific organic anion transporter 3TM12

ENSECAG00000022371	ENSG00000214163	24	7	4		cytidine deaminase activity,AU-rich element binding	XP_001499921.1AICDA|XP_001501883.1|XP_001916568.1|XP_001916562.1|APOBEC2|APOBEC1|AICDA|XP_001501883.1|XP_001916568.1|XP_001916562.1|APOBEC2|APOBEC1|AICDA|XP_001501883.1|XP_001916568.1|XP_001916562.1|APOBEC2|APOBEC1|AICDA|XP_001501883.1|XP_001916568.1|XP_001916562.1|APOBEC2|APOBEC1|	Probable C->U-editing enzyme APOBEC-2 (EC 3.5.4.-)
ENSECAG00000001856	ENSG00000183340	21	9	3		mRNA binding,DNA binding	JRKLJRK|TIGD7|Q4W445_HORSE|TIGD4|TIGD3|CENPB|	Jerky protein homolog-like (HHMJG)
ENSECAG00000007449	ENSG00000183675	21	9	7		non-membrane spanning protein tyrosine phosphatase activity,zinc ion binding,protein binding,	XP_001494699.2PTPN4|PTPN3|PTPN1|PTPN21|PTPN2|PTPN14|PTPN9|PTPN4|PTPN3|PTPN1|PTPN21|PTPN2|PTPN14|PTPN9|PTPN4|PTPN3|PTPN1|PTPN21|PTPN2|PTPN14|PTPN9|	Tyrosine-protein phosphatase non-receptor type 20 (hPTPN20a)(EC 3.1.3.48)
ENSECAG00000018684	ENSG00000102076	21	9	7	response to stimulus	photoreceptor activity,protein binding,G-protein coupled photoreceptor activity	OPSR_HORSEOPN1SW|Q95M85_HORSE|OPN3|OPN4|OPN5|RRH|A6P3D5_HORSE|OPN1SW|Q95M85_HORSE|OPN3|OPN4|OPN5|RRH|A6P3D5_HORSE|OPN1SW|Q95M85_HORSE|OPN3|OPN4|OPN5|RRH|A6P3D5_HORSE|	Red-sensitive opsin (Red cone photoreceptor pigment)
ENSECAG00000014654	ENSG00000215750	20	8	4			XP_001496767.1TEKT3|TEKT1|TEKT5|TEKT2|TEKT3|TEKT1|TEKT5|TEKT2|TEKT3|TEKT1|TEKT5|TEKT2|TEKT3|TEKT1|TEKT5|TEKT2|TEKT3|TEKT1|TEKT5|TEKT2|	hypothetical protein
ENSECAG00000000761	ENSG00000126952	18	4	6		protein binding	XP_001914856.1|XP_001488763.2|XP_001492619.2|A4UZ09_HORSE|XP_001488763.2|XP_001492619.2|A4UZ09_HORSE|XP_001488763.2|XP_001492619.2|A4UZ09_HORSE|	Nuclear RNA export factor 5 (TAP-like protein 1)(TAPL-1)
ENSECAG00000009473	ENSG00000082175	16	3	3	glucocorticoid metabolic process	transcription regulator activity,androgen receptor activity,sequence-specific DNA binding	Q5J7Z4_HORSEXP_001501712.2|Q8MJ07_HORSE|Q2MDJ1_HORSE|Q2YHQ1_HORSE|XP_001501712.2|Q8MJ07_HORSE|Q2MDJ1_HORSE|Q2YHQ1_HORSE|XP_001501712.2|Q8MJ07_HORSE|Q2MDJ1_HORSE|Q2YHQ1_HORSE|XP_001501712.2|Q8MJ07_HORSE|Q2MDJ1_HORSE|Q2YHQ1_HORSE|	Progesterone receptor (PR)(Nuclear receptor subfamily 3 group C member 3)
ENSECAG00000000970	ENSG00000182334	15	1	6			XP_001500502.1|XP_001504813.2|XP_001500412.2|XP_001500377.2|XP_001500330.2|XP_001500371.2|XP_001500443.2|XP_001500405.2|XP_001500357.2|XP_001500431.2|XP_001500451.2|XP_001500523.2|XP_001500532.2|	Olfactory receptor 5P3 (Olfactory receptor OR11-94)(Olfactory receptor-like protein JCG1)

ENSECAG00000021446	ENSG00000176884	6	2	2	N-methyl-D-aspartate selective glutamate receptor activity	XP_001917018.1XP_001497085.2|XP_001917547.1|XP_001497085.2|XP_001917547.1|XP_001497085.2|XP_001917547.1|	Glutamate [NMDA] receptor subunit zeta-1 Precursor (N-methyl-D-aspartate receptor subunit NR1)
ENSECAG00000010275	ENSG00000171189	6	2	2	kainate selective glutamate receptor activity	XP_001503964.1|XP_001915295.1|XP_001503964.1|XP_001915295.1|XP_001503964.1|XP_001915295.1|	Glutamate receptor, ionotropic kainate 1 Precursor (Glutamate receptor 5)(GluR-5)(GluR5)(Excitatory amino acid receptor 3)(EAA3)
ENSECAG00000007499	ENSG00000156222	6	2	2		SLC28A1SLC28A2|XP_001916663.1|XP_001489009.2|XP_001489536.1||	Sodium/nucleoside cotransporter 1 (Na(+)/nucleoside cotransporter 1)(
ENSECAG00000023294	ENSG00000133020	8	3	3	ATP binding	XP_001917202.1XP_001504901.1|MYH1_HORSE|XP_001504901.1|MYH1_HORSE|XP_001504901.1|MYH1_HORSE|XP_001504901.1|MYH1_HORSE|	Myosin-8 (Myosin heavy chain 8)(Myosin heavy chain, skeletal muscle, perinatal)(MyHC-perinatal)
ENSECAG00000004322	ENSG00000189332	9	4	3		XP_001504315.1|MCART6|SLC25A21|XP_001504315.1|MCART6|SLC25A21|XP_001504315.1|MCART6|SLC25A21|	Putative mitochondrial carrier protein
ENSECAG00000011054	ENSG00000124006	10	4	4		XP_001493930.2|XP_001495907.1|XP_001501342.1|XP_001493930.2|XP_001495907.1|XP_001501342.1|	Obscurin-like protein 1 Precursor
ENSECAG00000021904	ENSG00000102080	12	5	3		XP_001494371.2TMCC2|TMCC1|TMCC3|TMCC2|TMCC1|TMCC3|TMCC2|TMCC1|TMCC3|	Testis-specific protein TEX28
ENSECAG00000006320	ENSG00000212747	12	4	4	protein binding,	XP_001488032.1XP_001488032.1|LDOC1|LDOC1L||XP_001488032.1|LDOC1|LDOC1L||XP_001488032.1|LDOC1|LDOC1L||	Protein FAM127C (Mammalian retrotransposon derived protein 8B)
ENSECAG00000004825	ENSG00000179695	15	2	7		XP_001490308.2XP_001491314.2|XP_001491394.2|XP_001489664.2|XP_001490884.2|XP_001489623.2|XP_001489426.2|XP_001489736.2|XP_001491527.2|XP_001491562.2|XP_001491449.2|XP_001491476.2|XP_001491262.2|XP_001504765.2|XP_001489091.2|XP_001489967.2|	Olfactory receptor type 6C

ENSECAG00000006668	ENSG00000173272	4	1	1	XP_001488058.1XP_001914845.1||XP_001914845.1||	Putative uncharacterized protein
ENSECAG00000013296	ENSG00000198183	4	1	1	XP_001498856.1|XP_001498787.2|	Secretory protein in upper respiratory tracts
ENSECAG00000015474	ENSG00000163539	3	1	1	CLASP2|CLASP1|	Cytoplasmic linker-associated protein 2

Table S10. SNP Discovery Statistics

Breed Code	Breed	Aligned Reads	Examined Bases	Total SNP	Unique SNP
S255	Arabian	94,904	79,822,819	63,134	52,434
S256	Quarter horse	94,903	79,894,408	61,583	50,727
S257	Andalusian	95,381	82,019,434	66,723	55,658
S258	Icelandic horse	92,207	77,633,110	62,691	52,417
S259	Akhal teke	96,534	82,282,641	67,986	56,298
S260	Standardbred	92,103	78,493,569	64,578	53,509
S261	Thoroughbred	79,582	62,539,497	42,718	34,577
G836	Thoroughbred (Twilight)	6,173,763	2,367,053,447	797,330	746,073
	Multi-breed (not Twilight)				9,803
	Multi-breed (with Twilight)				51,257
	Total Unique SNP				1,162,753

Table S11. Haplotype sharing across breeds is assessed by genotyping SNP within the first 100kb of each region. The haplotype frequency within each breed is shown.

Region	Haplo-type	All breed	Andal-usian	Arabian	Belgian draft	French-Trotter	Hanoverian	Hokkaido	Icelandic	Norwegian Fjord	Quarter horse	Standardbred	Thorough-bred	Twilight	Breeds With This Haplotype
chr18	111322	0.183	0.083	0.045	0.438	0.646	0.208	0.083	0.087	0.312	0.087		0.188		10
chr18	311322	0.181	0.25	0.136	0.125	0.146	0.167	0.083	0.457	0.104	0.174	0.333	0.104		11
chr18	311324	0.173		0.045	0.167		0.125	0.708		0.188	0.109	0.125	0.208		8
chr18	111324	0.141	0.333	0.136	0.104		0.104	0.125		0.25	0.196		0.271		8
chr18	313322	0.079	0.021	0.227		0.021	0.167				0.326	0.104	0.104		7
chr15	2343	0.428	0.354	0.318	0.458	0.333	0.208	0.875	0.196	0.521	0.457	0.521	0.583		11
chr15	2323	0.213	0.167	0.636	0.271	0.167	0.062	0.104	0.5	0.312	0.217		0.021		10
chr15	2341	0.116	0.208	0.023	0.042	0.208	0.208	0.021			0.109	0.396	0.062		9
chr15	3323	0.072	0.146			0.188	0.229				0.196		0.021		5
chr14	?4	0.766	1	0.773	0.958	0.75	0.729	0.542	0.739	0.938	0.804	0.583	0.583	0.5	11
chr14	?3	0.234		0.227	0.042	0.25	0.271	0.458	0.261	0.062	0.196	0.417	0.417	0.5	10
chr29	2441	0.488	0.5	0.477	0.771	0.5	0.312	0.688	0.478	0.604	0.261	0.583	0.188		11
chr29	2431	0.225	0.271	0.159	0.021	0.417	0.146	0.292	0.326	0.104	0.217	0.229	0.333		11
chr29	2444	0.081	0.062	0.182			0.062				0.174	0.042	0.292		6
chr29	2141	0.053	0.083			0.021	0.042	0.021	0.174	0.146	0.022		0.042		8
chr10	3313	0.393	0.062	0.295	0.792	0.188	0.271	0.729	0.283	0.896	0.5	0.229	0.042		11
chr10	2313	0.335	0.188	0.432	0.042	0.583	0.417	0.021	0.261	0.083	0.391	0.312	0.896	1	11
chr10	3333	0.104	0.25	0.068	0.146		0.104	0.083	0.435			0.104	0.021		8
chr10	3233	0.099	0.458	0.205		0.167	0.042	0.083	0.022			0.083	0.021		8
chr6	4?32	0.539	0.604	0.977	0.521	0.729	0.708	0.521	0.391	0.271	0.022	0.771	0.521	0.5	11
chr6	2?12	0.225	0.354		0.458	0.25	0.083	0.479	0.109	0.271	0.087	0.229			9
chr6	4?34	0.146		0.023			0.062		0.457	0.271	0.87				5
chr6	2?14	0.074				0.021	0.104		0.043	0.188	0.022		0.479	0.5	6
chr5	323	0.525	0.417	0.409	0.354	0.708	0.604	0.812	0.413	0.438	0.522	0.646	0.375	0.5	11
chr5	143	0.275	0.354	0.455	0.292	0.25	0.396	0.062			0.435	0.25	0.625	0.5	9
chr5	123	0.114		0.114	0.333			0.062	0.217	0.458	0.043	0.042			7
chr5	321	0.051	0.062					0.042	0.37			0.062			4
chr4	1433134	0.289	0.354	0.227		0.146	0.146	0.417	0.783	0.333	0.457		0.542		9
chr4	3413134	0.092	0.208	0.227			0.062	0.333	0.065	0.042	0.087		0.083		8
chr4	1233134	0.07	0.021	0.25		0.146	0.146		0.043	0.167	0.043		0.021		8
chr4	1234312	0.063	0.042			0.292	0.062			0.021	0.022		0.312		6
chr4	1413134	0.055	0.042	0.023				0.208		0.062					4
chr4	1434312	0.055	0.021			0.292	0.104				0.022				4
chr17	12241	0.519	0.521	0.25	0.729	0.771	0.5	0.229	0.674	0.396	0.435	0.729	0.375	0.5	11
chr17	22211	0.153	0.396	0.295		0.146	0.229	0.479	0.043		0.196	0.021	0.042		9
chr17	12211	0.123		0.159	0.271	0.021	0.062	0.229	0.174	0.125	0.087	0.167	0.083		10
chr17	12442	0.086		0.182			0.042		0.087	0.396	0.043		0.271		6
chr11	212113	0.201		0.5		0.188	0.458				0.065	0.438	0.188		6
chr11	212?13	0.194	0.708						0.913	0.708					3
chr11	211133	0.113		0.205		0.104	0.125				0.087	0.396	0.25		6
chr11	212213	0.083						0.5			0.5				2

chr11	?12113	0.081		0.958					1
chr11	211113	0.063	0.136		0.104	0.167	0.109	0.188	5

† Analysis uses LD samples (Table S1).

Table S12. Haplotype sharing across breeds assessed by resequencing the first 5kb of each genomic region.

Chromosome	Number of Common Haplotypes at locus*	Number of Unique Haplotypes	Sum of all Haplotypes that are Common Haplotypes	Total Haplotypes present	Percentage of haplotypes present that are unique
4	7	8	30	38	21.0%
5	6	22	16	38	57.9%
6	4	23	15	38	60.5%
10	9	10	28	38	26.3%
11	5	0	38	38	0.0%
14	3	30	8	38	78.9%
15	6	5	33	38	13.2%
17	6	5	33	38	13.2%
18	7	21	17	38	55.3%
29	4	17	21	38	44.7%
Sum all loci	57	141	239	380	37.1%

* Common haplotypes occur in more than one horse
† Table uses re-sequencing horses from Table S1.

Table S13. SNPs needed to differentiate horse haplotypes for within-breed gene mapping (by simulation)

Desired mean r^2max	0.7	0.8	0.9	1
High LD breed	30,000	100,000	175,000	245,000
Moderate LD breed	155,000	225,000	300,000	370,000
Low LD breed	250,000	320,000	390,000	460,000

Table S14. High quality differences between Przewalski horse and EquCab2.0

Place	Type of difference	Differences	Aligned Bases
chr14_approx88.9Mb	SNP	1	523
chr5_approx54.8Mb	SNP	1	675
chr11:25106446-25106700	single base insertion in Przewalski	1	206
chr4:44817456-44818287	0	0	660
chr17:62037100-62037435	SNP	1	333
chr18:16037590-16037897	12bp insertion (polymorphism missing in KB4064)	1	320
chr10:19766365-19766990	SNP	1	668
		6	3385

Supplemental References

S1. D. B. Jaffe *et al.*, *Genome Res* **13**, 91 (Jan, 2003).
S2. K. Lindblad-Toh *et al.*, *Nature* **438**, 803 (Dec 8, 2005).
S3. T. Raudsepp *et al.*, *Cytogenet Genome Res* **122**, 28 (2008).
S4. T. L. Lear, J. Lundquist, W. W. Zent, W. D. Fishback, Jr., A. Clark, *Cytogenet Genome Res* **120**, 117 (2008).
S5. T. S. Mikkelsen *et al.*, *Nature* **447**, 167 (May 10, 2007).
S6. R. C. Edgar, E. W. Myers, *Bioinformatics* **21 Suppl 1**, i152 (Jun, 2005).
S7. A. L. Price, N. C. Jones, P. A. Pevzner, *Bioinformatics* **21 Suppl 1**, i351 (Jun, 2005).
S8. O. Kohany, A. J. Gentles, L. Hankus, J. Jurka, *BMC Bioinformatics* **7**, 474 (2006).
S9. W. Gish, in *http://blast.wustl.edu*. (2004).
S10. T. Leeb *et al.*, *Genomics* **87**, 772 (Jun, 2006).
S11. L. Carbone *et al.*, *Genomics* **87**, 777 (Jun, 2006).
S12. S. Trazzi *et al.*, *PLoS One* **4**, e5832 (2009).
S13. M. Bieda, X. Xu, M. A. Singer, R. Green, P. J. Farnham, *Genome Res* **16**, 595 (May, 2006).
S14. M. Garber *et al.*, *Bioinformatics* **25**, i54 (Jun 15, 2009).
S15. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (Sep 1, 1997).
S16. D. M. Carone *et al.*, *Chromosoma* **118**, 113 (Feb, 2009).
S17. G. S. Slater, E. Birney, *BMC Bioinformatics* **6** (Feb 15, 2005).
S18. Z. Ning, A. J. Cox, J. C. Mullikin, *Genome Res* **11**, 1725 (Oct, 2001).
S19. A. J. Viterbi, *IEEE Trans. Inform. Process* **13**, 260 (1967).
S20. P. Scheet, M. Stephens, *Am J Hum Genet* **78**, 629 (Apr, 2006).
S21. J. Felsenstein, *Syst Biol* **46**, 101 (Mar, 1997).
S22. D. Gordon, C. Abajian, P. Green, *Genome Res* **8**, 195 (Mar, 1998).
S23. L. S. Sandmeyer, C. B. Breaux, S. Archer, B. H. Grahn, *Vet Ophthalmol* **10**, 368 (Nov-Dec, 2007).
S24. R. R. Bellone *et al.*, *Genetics* **179**, 1861 (Aug, 2008).
S25. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, *Bioinformatics* **21**, 263 (Jan 15, 2005).
S26. G. Rosengren Pielberg *et al.*, *Nat Genet* **40**, 1004 (Aug, 2008).
S27. E. S. Lander *et al.*, *Nature* **409**, 860 (Feb 15, 2001).
S28. C. G. Elsik *et al.*, *Science* **324**, 522 (Apr 24, 2009).
S29. R. H. Waterston *et al.*, *Nature* **420**, 520 (Dec 5, 2002).
S30. M. Ventura *et al.*, *Science* **316**, 243 (Apr 13, 2007).
S31. P. E. Warburton, *Chromosome Res* **12**, 617 (2004).
S32. D. J. Amor, K. H. Choo, *Am J Hum Genet* **71**, 695 (Oct, 2002).
S33. O. Capozzi *et al.*, *Genome Res* **19**, 778 (May, 2009).
S34. M. Ventura *et al.*, *Genome Res* **14**, 1696 (Sep, 2004).
S35. D. J. Amor *et al.*, *Proc Natl Acad Sci U S A* **101**, 6542 (Apr 27, 2004).
S36. A. C. Chueh, E. L. Northrop, K. H. Brettingham-Moore, K. H. Choo, L. H. Wong, *PLoS Genet* **5**, e1000354 (Jan, 2009).
S37. A. R. Kwon, E. C. Choi, *Arch Pharm Res* **28**, 561 (May, 2005).

S38. Y. Y. Pang, A. Schermer, J. Yu, T. T. Sun, *J Cell Sci* **104 (Pt 3)**, 727 (Mar, 1993).
S39. J. W. Wu *et al.*, *J Eur Acad Dermatol Venereol* **23**, 174 (Feb, 2009).
S40. M. A. Levine, *Equine Vet J Suppl*, 6 (Apr, 1999).
S41. W. Ridgeway, *The origin and influence of the Thoroughbred horse*. A. E. Shipley, Ed., Cambridge Biological Series (Cambridge University Press, London, 1905), pp. 565.
S42. C. Vila *et al.*, *Science* **291**, 474 (Jan 19, 2001).
S43. G. Lindgren *et al.*, *Nat Genet* **36**, 335 (Apr, 2004).
S44. G. T. Marth, E. Czabarka, J. Murvai, S. T. Sherry, *Genetics* **166**, 351 (Jan, 2004).
S45. R. A. Gibbs *et al.*, *Science* **324**, 528 (Apr 24, 2009).
S46. A. N. Lau *et al.*, *Mol Biol Evol* **26**, 199 (Jan, 2009).
S47. E. Oancea *et al.*, *Sci Signal* **2**, ra21 (2009).