Identification of copy number variants in horses

Ryan Doan,[1] Noah Cohen,[2] Jessica Harrington,[2] Kylee Veazy,[2] Rytis Juras,[3] Gus Cothran,[3] Molly E. McCue,[4] Loren Skow,[3] and Scott V. Dindot[1,5,6]

[1]Department of Veterinary Pathobiology, [2]Department of Large Animal Clinical Sciences, [3]Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, Texas 77843, USA; [4]Department of Veterinary Population Medicine, University of Minnesota College of Veterinary Medicine, St. Paul, Minnesota 55108, USA; [5]Department of Molecular and Cellular Medicine, Texas A&M Health Science Center College of Medicine, College Station, Texas 77843, USA

Copy number variants (CNVs) represent a substantial source of genetic variation in mammals. However, the occurrence of CNVs in horses and their subsequent impact on phenotypic variation is unknown. We performed a study to identify CNVs in 16 horses representing 15 distinct breeds (*Equus caballus*) and an individual gray donkey (*Equus asinus*) using a whole-exome tiling array and the array comparative genomic hybridization methodology. We identified 2368 CNVs ranging in size from 197 bp to 3.5 Mb. Merging identical CNVs from each animal yielded 775 CNV regions (CNVRs), involving 1707 protein- and RNA-coding genes. The number of CNVs per animal ranged from 55 to 347, with median and mean sizes of CNVs of 5.3 kb and 99.4 kb, respectively. Approximately 6% of the genes investigated were affected by a CNV. Biological process enrichment analysis indicated CNVs primarily affected genes involved in sensory perception, signal transduction, and metabolism. CNVs also were identified in genes regulating blood group antigens, coat color, fecundity, lactation, keratin formation, neuronal homeostasis, and height in other species. Collectively, these data are the first report of copy number variation in horses and suggest that CNVs are common in the horse genome and may modulate biological processes underlying different traits observed among horses and horse breeds.

[Supplemental material is available for this article.]

Genetic variation, including single nucleotide polymorphisms (SNPs), insertion/deletion polymorphisms (INDELS), and copy number variants (CNVs), represent the major source of diversification of phenotypes in animals and humans (Freeman et al. 2006). Recent studies have shown that CNVs involving duplications and deletions represent a major form of genomic variation in humans, animals, and plants (Guryev et al. 2008; Chen et al. 2009; Nicholas et al. 2009; Zhang et al. 2009; Conrad et al. 2010; DeBolt 2010; Fadista et al. 2010; Mills et al. 2011; Nicholas et al. 2011). CNVs are typically defined as segments of DNA at least 1000 base pairs (1000 bp or 1 kb) in length that vary in copy number between two individuals of a given species. However, recent studies in humans have shown that CNVs can be much smaller, revealing that there is not a clearly defined size of a CNV (Zhang et al. 2009; Boone et al. 2010). Recent high-resolution human array comparative genomic hybridization (CGH) and next generation sequencing studies indicate that the median length of CNVs in the human genome is ~800 bp in length (Redon et al. 2006; Korbel et al. 2007; Levy et al. 2007; Kidd et al. 2008; Wheeler et al. 2008; Zhang et al. 2009; Mills et al. 2011).

The release of the first complete genome sequence of a Thoroughbred horse represented a great scientific achievement and has changed the landscape of genomic research in horses. The horse reference genome assembly has provided a catalog of over 1 million SNPs, a comparative map to other species, and a segmental duplication map of the horse genome (Wade et al. 2009). To date, however, there have been no reports of genome-wide CNV discovery in the horse genome.

In this study, we describe the first comprehensive analysis of copy number variation in horses using a custom-designed whole-exome tiling array. To maximize the detection of CNVs in the horse genome, we examined 16 horses representing 15 diverse breeds and an individual gray donkey. We determined biological processes enriched for CNVs as well as those present in genes associated with Mendelian traits in animals and humans.

Results

Development of a horse exome array for comparative genomic hybridization

To identify CNVs affecting RNA- and protein-coding genes in the horse genome, we developed a high-density tiling array involving 418,576 unique probes targeted specifically to the 5′ and 3′ untranslated regions (UTR) and coding exons of 20,882 Ensembl-annotated genes (Methods). By targeting only exons, we were able to achieve an average resolution of one oligonucleotide per 98 bp (Supplemental Table S1).

Identification of CNVs in the horse genome

We performed CGH with 16 horses (*Equus caballus*) representing 15 diverse breeds, including Andalusian, Arabian, Curly, Gypsy Vanner, Hungarian, Lusitano, Miniature, Paso Fino, Peruvian, Welsh-Arabian Pony, Quarter Horse (two horses), Shagya Arabian, Shire, Thoroughbred, and Hanoverian, and an individual gray donkey (*Equus asinus*) as an evolutionary outlier, for a total of 17 animals. To identify CNVs, a single Thoroughbred mare was used as the reference sample (referred to hereafter as reference Thoroughbred). A self-self hybridization of the reference Thoroughbred

[6]Corresponding author.
E-mail sdindot@cvm.tamu.edu.

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.128991.111.

was performed to establish the parameters for calling variants and to determine the false discovery rate (FDR) of the array. The FDR was ~1%, and probes showing losses and gains in the self-self hybridization were subsequently removed from the analysis. To assess gains and losses in the reference Thoroughbred relative to other horses, a dye-swap experiment was performed using the reference Thoroughbred as the experimental sample and the Thoroughbred mare used for the equine genome sequencing project (referred to hereafter as Twilight) as the reference sample. Array CGH was also performed on a second Quarter Horse using Twilight as the reference sample.

PCR and Sanger sequencing were then used to determine the minimum size of a CNV that could be confirmed by a second method and that reflected the lowest FDR. We identified CNVs as small as 100 bp, but we were unable to confirm CNVs smaller than 197 bp by quantitative PCR or Sanger sequencing. The smallest confirmed CNV was in an exon of the ENSECAG00000022453 gene (*LLRC37B*), which showed a 197-bp gain in the Lusitano and a loss in the Miniature, whereas no CNV was detected in the Gypsy Vanner (Fig. 1A–C). PCR amplification of the CNV in the three horses including the reference Thoroughbred revealed that the Thoroughbred and Gypsy Vanner horses had two amplicons of different sizes—one of the predicted size (742 bp) and one of a larger size (907 bp). The Lusitano had a single amplicon of the larger size and the Miniature had a single amplicon of the predicted size (Fig. 1D). Sanger sequencing of the two amplicons revealed that the predicted size amplicon was identical to the equCab2/Twilight sequence, whereas the larger amplicon contained a 165-bp duplicated region (Fig. 1E). Thus, the Gypsy Vanner and reference Thoroughbred carried a heterozygous duplication (+/Dup), the Miniature carried homozygous wild-type alleles (+/+), and the Lusitano carried a homozygous duplication (Dup/Dup). Quantitative PCR of randomly selected CNVs >197 bp confirmed 23 of 24 (4% FDR) genes within 18 of 19 (5% FDR) CNVs (Supplemental Figs. S1–S17; Fig. 5, see below). Therefore, 197 bp was used as the minimum size cutoff for CNV detection.

Analysis of the horses and the gray donkey identified 3343 CNVs; however, after filtering probes with signal intensities greater than three standard deviations of the mean signal intensity, this number was reduced to 2368. The CNVs ranged in size from 197 bp to 3.5 Mb, with median and mean sizes of 5.3 kb and 99.4 kb, respectively. The CNVs included 1509 gains and 859 losses (Table 1). The number of CNVs per animal ranged from 55 to 347. Merging identical CNVs yielded 775 CNV regions (CNVRs), including 398 gains and 315 losses (Supplemental Table S2). The increased number of gains appeared to reflect the large number of losses in the reference Thoroughbred (41 losses relative to Twilight) (see Table 1). We identified 438 individual CNVs (i.e., a CNV detected in a single horse and not shared with another animal). CNVs affected 1707 genes by either encompassing a group of genes, an individual gene, or a portion of a gene (e.g., duplicated or deleted exon) (Supplemental Table S3). The number of CNV genes per animal ranged from 117 to 671. Sixty-two CNVRs were complex, as they were present as both gains and losses in different horses. Of the CNVs identified, 214 were homozygous deletions and ranged in sizes from 205 bp to 357.3 kb. Merging identical homozygous deletions yielded 36 deletion regions involving 64 genes. Thirty-one of the homozygous deletions were present only in an individual horse and not shared among the other horses. CNVs were identified on each of the 31 autosomes and the X chromosome (Fig. 2; Supplemental Table S4). Chromosomes 12 (15.1%; $P = 2.1 \times 10^{-9}$), 17 (9.1%; $P = 4.1 \times 10^{-3}$), and 23 (8.2%; $P = 1.6 \times 10^{-2}$) were enriched ($P < 0.05$) for CNVs (Supplemental Table S5). We found that 142 CNVRs (18%), including 35 of the complex CNVs, were present in segmentally duplicated regions where larger CNVs tended to be present (Supplemental Fig. S19). The sex of each horse was accurately determined using probes located on the X chromosome (Table 1; Supplemental Fig. S18). We found that 559 of the genes affected by a CNV were known to exist as CNVs in humans (http://projects.tcag.ca/variation/). Collectively the CNVRs totaled 86,373,976 bp of DNA, or ~3.6% of the assembled horse genome. Our findings indicate that CNVs are located throughout the horse genome and represent a substantial source of genetic variation among the horses and the donkey investigated.

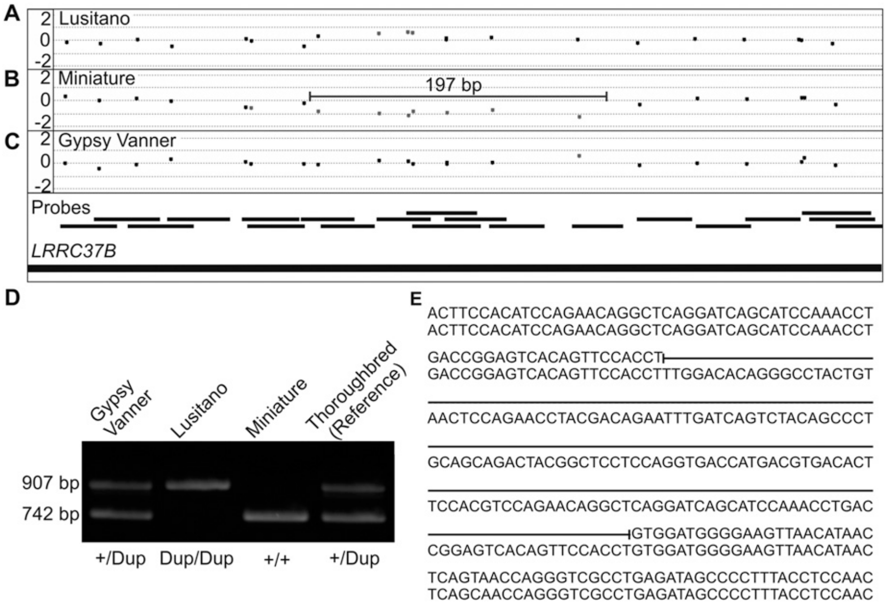

Figure 1. Confirmation of smallest CNV identified by array CGH. Log_2 ratio plots of a Lusitano (*A*), a Miniature (*B*), and a Gypsy Vanner horse (*C*). (*D*) PCR amplification of the CNV region showing predicted (742-bp) and larger (907-bp) amplicons where Dup and + indicate duplicated and wild-type alleles, respectively. (*E*) Sanger sequencing of the predicted and larger amplicon showing 165-bp duplicated region.

CNV sharing among horses

Hierarchical clustering of $\log_2$ ratios of probes within the CNVs grouped some of the horses based on their relative ancestry (e.g., Arabian, Lusitano, and Andalusian) (McCue et al. 2012), with the Donkey being separated from the horse lineage (Fig. 3A; Wade et al. 2009). As a group, the majority of CNVs, including the homozygous deletions, were present only in an individual horse, with sharing among three or more horses occurring in <22% of the animals (Fig. 3B,C). However, the large number of individual CNVs was the

Table 1. Horse breed and number of detected CNVs

Breed	Sex	CNVs	Gains	Losses	# Genes
Andalusian	M	81 (5)	48 (2)	33 (3)	209 (7)
Arabian	F	97 (30)	55 (19)	42 (11)	163 (42)
Curly	M	55 (7)	28 (4)	27 (3)	117 (10)
Gypsy Vanner	F	85 (7)	41 (3)	44 (4)	252 (66)
Hanoverian[a]	F	121 (10)	92 (4)	29 (6)	240 (20)
Hungarian[a]	F	108 (2)	74 (1)	34 (1)	274 (7)
Lusitano	F	86 (6)	49 (3)	37 (3)	185 (17)
Miniature[a]	F	347 (108)	272 (74)	75 (34)	671 (158)
Paso Fino[a]	F	133 (4)	107 (2)	26 (2)	267 (8)
Peruvian[a]	F	137 (5)	102 (4)	35 (1)	271 (20)
Quarter Horse (#1)[a]	F	222 (11)	176 (6)	46 (5)	489 (33)
Shagya Arabian[a]	F	196 (12)	165 (10)	31 (2)	367 (31)
Shire	F	184 (37)	87 (4)	97 (33)	378 (82)
Welsh-Arabian Pony	M	88 (24)	54 (15)	34 (9)	211 (32)
Donkey[a]	M	300 (157)	114 (42)	186 (115)	533 (280)
Quarter Horse (#2)[a,b]	F	58 (12)	16 (4)	42 (8)	222 (9)
Thoroughbred[b]	F	70 (12)	29 (5)	41 (7)	177 (22)

Numbers in parentheses indicate individual CNV.
[a]Significant difference between gain and loss.
[b]Twilight reference.

result of those identified in the Donkey and the Miniature, which represented >50% of the total individual CNVs (Table 1). Examined individually, each horse shared most of their CNVs with at least another animal (Fig. 3D), except for the Donkey, which shared less than half (41.3%). As expected, this reflected the sharing of genes affected by CNVs (Fig. 3E). We estimated that the reference Thoroughbred contained six individual CNVs, as these regions were shared among all of the horses. The CNVs identified in the Donkey did not appear to be due to sequence variation since the number of gains and losses was proportional to that observed in the other horses and as some of the CNVs were confirmed by a second method, particularly gains relative to the Thoroughbred (Table 1). We also examined correlations between CNV length and sharing among the horses and found that smaller CNVs were less likely to be shared, in contrast to larger CNVs (Fig. 3F), although this finding could be because larger CNVs were easier to detect. Collectively, these data suggest that CNVs are shared among horses of closely related breeds and that individual CNVs are more common in divergent breeds relative to the Thoroughbred (e.g., Donkey, Miniature, and Arabian).

Functional analysis of CNVs

We performed functional analysis clustering of the genes affected by a CNV to understand the potential effects of CNVs on gene biotypes and biological processes in horses. We found that 8.2% of the protein-coding genes, 6.7% of the small nucleolar (snoRNA) genes, 3.4% of the microRNA (miRNA) genes, 3.8% of the small nuclear (snRNA) genes, 6.3% of the miscellaneous RNA (miscRNA) genes, and 11.9% of the ribosomal RNA (rRNA) genes analyzed were affected by a CNV (Fig. 4A). Overall, we found that the majority (96.4%) of the genes with CNVs were protein-coding (Fig. 4B).

Next, we performed a functional annotation clustering analysis of genes affected by a CNV to identify biological processes (BP) enriched for these variants. After determining that only a small fraction (0.7%) of the horse Ensembl genes with CNVs had associated BP terms (Huang et al. 2009a,b), we performed the analysis using the human orthologs of the horse genes. We found that CNVs were primarily enriched ($P < 0.05$) in processes involved in sensory perception ($P = 5.4 \times 10^{-79}$), signal transduction ($P = 7.3 \times 10^{-7}$), and metabolism ($P = 0.01$) (Fig. 4C). We also found that CNVs detected only in an individual horse were primarily enriched in processes involved in sensory perception ($P = 3.5 \times 10^{-14}$) (Fig. 4D). Genes affected by homozygous deletions were also primarily enriched in sensory perception ($P = 8.5 \times 10^{-11}$) and signal transduction ($P = 0.01$), with no other processes being affected (Fig. 4E). Collectively, these data indicate that CNVs are present in protein- and RNA-coding genes in the horse genome and that biological processes regulating sensory perception, signal transduction, and metabolism are the primary processes affected by CNVs in the horses investigated.

Mendelian genes affected by CNVs

Next, we examined whether CNVs were present at genes underlying Mendelian traits in animals and humans. Cross-reference of the genes affected by a CNV with the Online Mendelian Inheritance in Animals (OMIA) and Online Mendelian Inheritance in Man (OMIM) databases returned 13 genes, including those associated with blood group antigens, coat color, fecundity, lactation, keratin formation, and height (Table 2). Each gene was differentially affected by a CNV, by either involving a duplicated or deleted portion of an exon, an entire exon or entire exons, a UTR, a gene, or multiple genes. We further investigated two of these CNVs (*PMEL* and *BMPR1B*), because of the association of these genes with known phenotypes (Brunberg et al. 2006; Chu et al. 2007).

The premelanosome protein (*PMEL*) gene, which causes the Silver coat color in horses, was identified as having a loss of exon 6 in the Miniature (Fig. 5A,B) and Shagya Arabian (data not shown). Exon 6 consists of an array of imperfect tandem repeats (Fig. 5C,D) and represents the repeat domain of the PMEL protein (Hoashi et al. 2006). An autosomal dominant C>T missense mutation in exon 11 changes the second amino acid in the cytoplasmic region of *PMEL* in exon 11 (Arg618Cys) and shows 100% concordance with the Silver coat color in horses (Brunberg et al. 2006). Thus, we investigated whether the CNV in exon 6 of the *PMEL* gene corre-

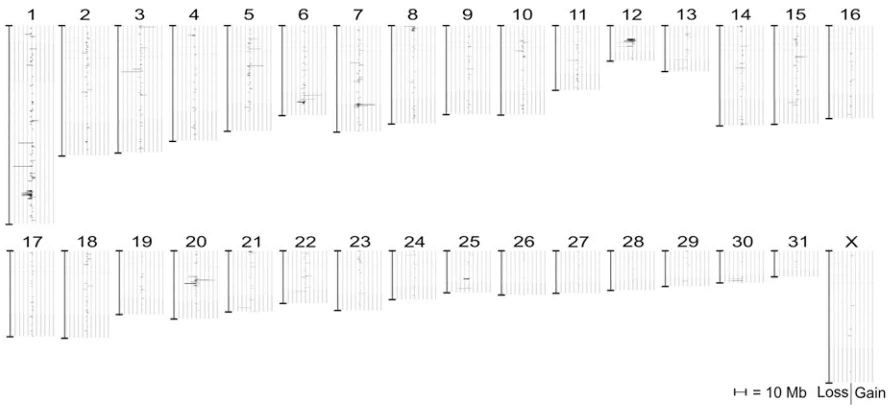

Figure 2. Distribution of CNVs in the horse genome. The bars on *left* indicate losses relative to the reference Thoroughbred, while those on the *right* indicate gains. The increase in bar length indicates an increase in number of samples sharing the CNV.

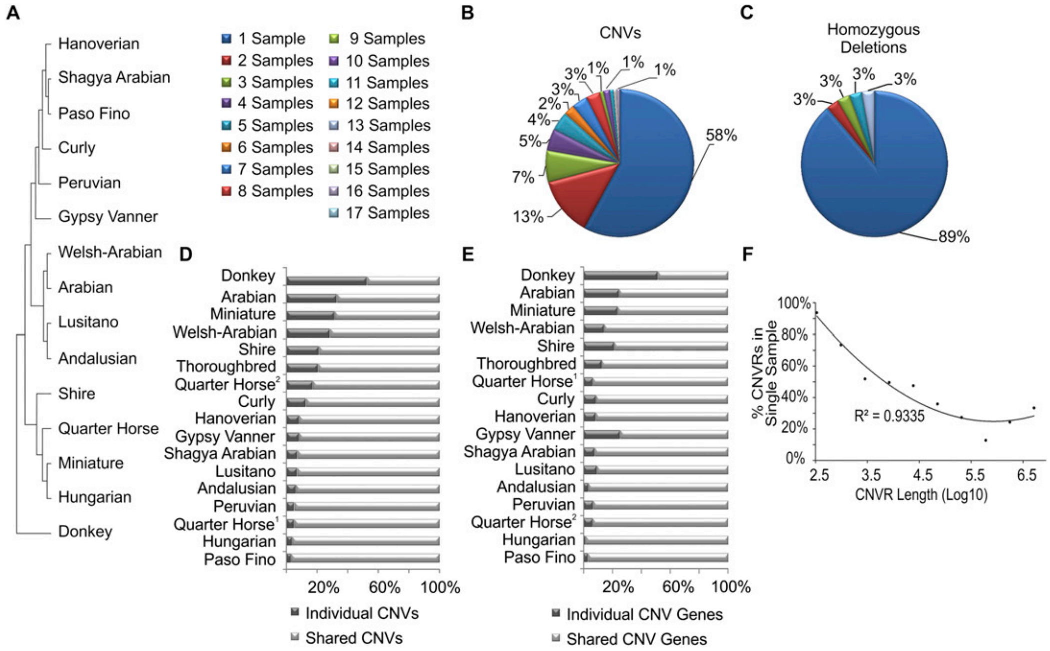

Figure 3. Analysis of CNVs among horse breeds and donkey. (*A*) Hierarchical clustering analysis of CNVs. (*B*) Percentage of CNVs shared among samples. (*C*) Percentage of homozygous deletions shared among samples. (*D*) Comparison of individual and shared CNVs among horse breeds and donkey. (*E*) Comparison of individual CNV genes in all animals. (*F*) Correlation between the size and percentage of individual CNVRs.

lated with the associated mutation and the Silver coat color. The Miniature horse had a Silver coat color and was genotyped as being heterozygous for the Silver mutation (C/T); however, the Shagya Arabian horse had a "flea-bitten" gray coat color and was homozygous for the nonSilver, wild-type alleles (C/C). Given this, we analyzed an additional eight Silver and three nonSilver horses that had been previously genotyped for the associated mutation (Supplemental Table S6). We found that two of the nonSilver (C/C) and four of the Silver horses (C/T and T/T) had the *PMEL* CNV. While genotyping the CNV by qPCR, we realized that horses homozygous for the C/C and T/T alleles continued to yield amplicons of the correct size, suggesting that the exon was not deleted but rather a loss in copy number relative to the WT allele. Although our efforts to define the breakpoints of the CNV by PCR were unsuccessful, we consistently observed a reduction in DNA content by qPCR. A similar CNV spanning exons 6 through 9 has also been detected in the human *PMEL* gene (Kim et al. 2009). We did not have sufficient probe coverage of intron 6 through exon 9, which may reflect our inability to define the breakpoints by PCR. Collectively, these data suggest that the imperfect tandem repeat of exon 6 varies in copy number in horses. Furthermore, these data indicate that the CNV is not associated with the Silver coat color; the CNV is not in linkage disequilibrium with the associated mutation, and the CNV is common among horses.

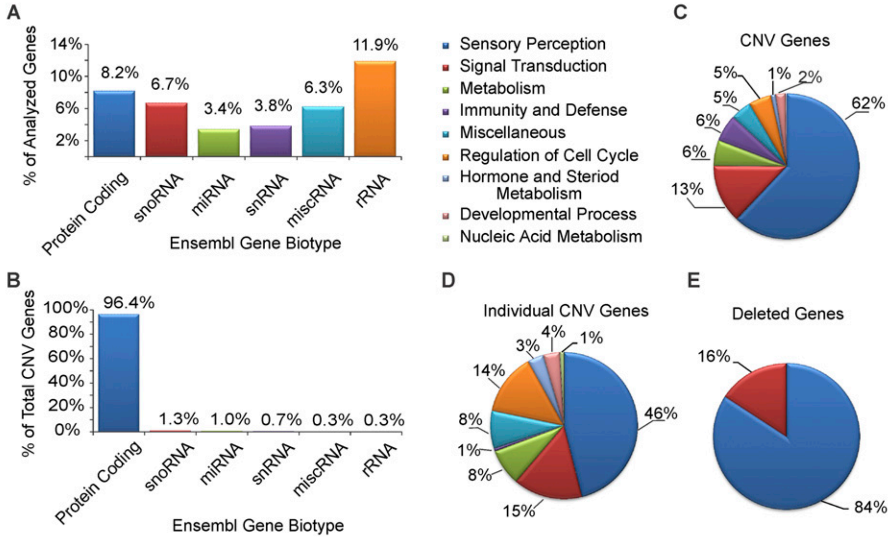

Figure 4. Functional analysis enrichment of genes affected by CNVs. (*A*) The percentage of gene biotypes affected by a CNV. (*B*) The percentage of CNV genes with each biotype. (*C*) Enrichment of biological processes in genes affected by a CNV. (*D*) Enrichment of biological processes in individual CNV genes. (*E*) Enrichment of biological processes in genes affected by homozygous deletions.

The bone morphogenic receptor type-1B (*BMPR1B*) gene, which is associated with increased ovulation rate in sheep (Chu et al. 2007), and the adjacent netrin receptor (*UNC5C*) gene, which is involved in dorsal guidance of hindbrain axons (Kim and Ackerman 2011), were originally identified as having a loss of the 3′ ends of *BMPR1B* and *UNC5C* genes in all of the animals (Fig. 5E). However, analysis of the CNV by qPCR, using primers flanking the CNV and within the *GAPDH* gene as controls, determined that the reference Thoroughbred mare had a duplication of the region

Table 2. CNV genes with predicted effects on phenotypes

Gene	OMIA / OMIM ID	Phenotype	Gene Region	Type	Genomic Location	Size (bp)
ABO	2660	Blood group system ABO	Gene/UTR	Loss	chr25:35463660-35517632	53,972
CSN2	2801	Casein, beta, null allele	Exon	Gain	chr3:64944052-64944249	197
ASIP	415	Coat color, agouti	Genes	Loss	chr22:25168412-25257554	89,142
AHCY	415	Coat color, agouti	Genes	Loss	chr22:25168412-25257554	89,142
PMEL	2833	Coat color, silver in horse	Exon	Loss	chr6:73668082-73668698	616
BMPR1B	676	Fecundity in sheep	Multiple exons	Loss	chr3:43547922-43618069	70,147
GDF9	677	Fecundity in sheep	Genes	Gain	chr14:42733400-42752318	18,918
KRT1	2618	Hyperkeratosis, palmoplantar	Genes	Complex	chr6:69736134-69793428	57,294
PITX3	1236	Microphthalmia in sheep	Genes	Gain	chr1:27894730-27947386	52,656
NAGLU	2869	Mucopolysaccharidosis IIIB in cow	UTR	Loss	chr11:20471939-20472326	387
UROS	2941	Porphyria, congenital erythropoietic	UTR	Loss	chr1:7305481-7305818	337
ZBTB38	%612221	Stature Quantitative Trait Locus 10	UTR	Loss	chr16:75595867-75596224	357
WRN	#277700	Werners Syndrome	Gene	Gain	chr27:12977576-13422891	445,315

(Fig. 5F–H). The Thoroughbred mare had a history of double ovulation, which had been confirmed by ultrasonography of multiple ovulatory follicles and had been pregnant with twins. Thus, this CNV was considered a candidate variant for increased or altered ovulation rates in horses. We investigated the *BMPR1B* CNV in a cohort of horses with altered ovulatory cycles (double ovulation [*n* = 2], anovulatory hemmorhagic follicle [*n* = 1], and 0 ovulation [*n* = 3]) and a cohort of Thoroughbreds with an unknown reproductive history (*n* = 55). We found that none of the horses with altered ovulatory rates had the CNV; however, we found that 18 Thoroughbreds, including Twilight, had the duplication (Supplemental Table S7). Furthermore, these horses were found to be heterozygous (*n* = 14) and homozygous (*n* = 4) for the duplication. Therefore, an association between the *BMPR1B* CNV and altered ovulation rate could not be conclusively determined. This may, in part, reflect the limited number of samples in which phenotypic information regarding ovulatory rates was available.

Despite the fact that there were less clear genotype/phenotype correlations with the other CNVs identified, some may, in fact, cause or modulate phenotypes associated with each gene, or they may be benign variants in horses similar to the ones investigated above. However, future studies with large cohorts of horses and well-defined phenotypes will be required to demonstrate the causal association of each CNV with its proposed phenotype(s).

Population analysis of CNVs

We examined seven CNVRs in 125 horses representing the Arabian, Curly, Peruvian, Quarter Horse, and Thoroughbred breeds (25 horses/breed). These seven CNVs were chosen based on potential functional effects (*BMPR1B*, *PMEL*, *WRN*, and *PTPRC*), ease of genotyping (*LRRC37B*), or because they were suspected to be complex rearrangements (*ABO*, *PTPRC*, and ENSECAG00000006791). Allele and genotype frequencies for each CNV and 95% confidence intervals for allele frequencies are reported in Supplemental Table S8**.** Differences in allele frequen-

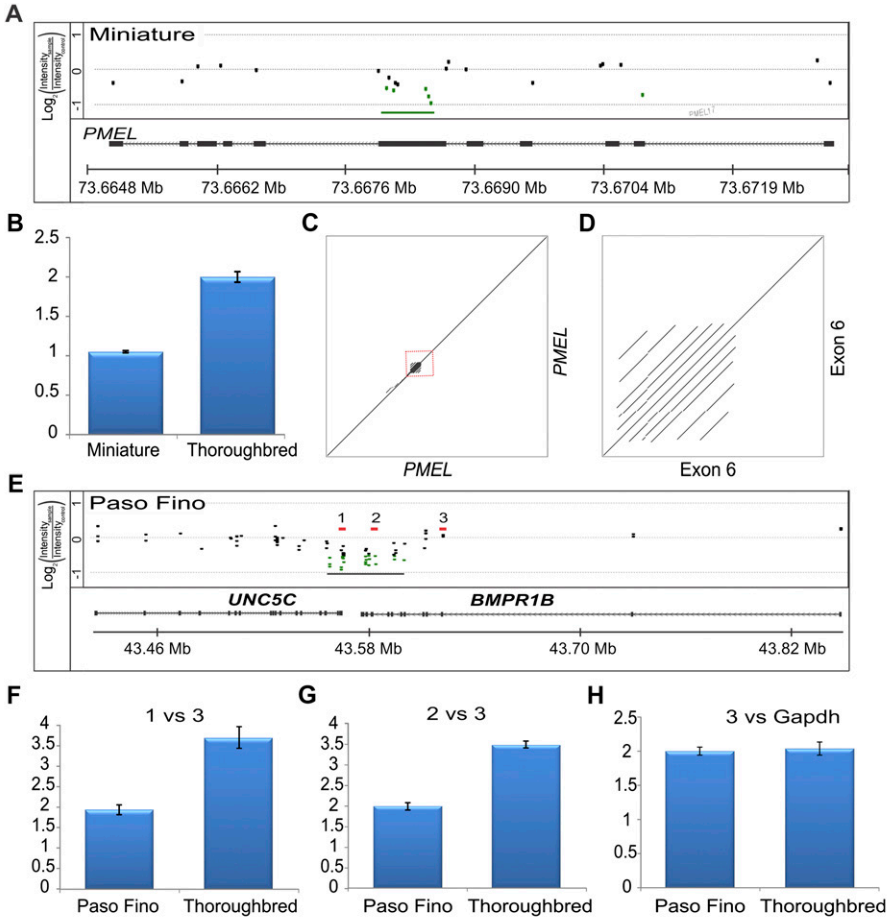

Figure 5. Identification and confirmation of CNVs in *PMEL* and *BMPR1B* genes. (*A*) Log$_2$ ratio plot of *PMEL* gene showing a loss in a Miniature relative to a Thoroughbred horse. (*B*) qPCR confirmation of loss in a Miniature. (*C*) Percent identity plot of *PMEL* gene sequence with exon 6 highlighted by red box. (*D*) Percent identity plot of *PMEL* exon 6 sequence. (*E*) Log$_2$ ratio plot of a *BMPR1B* gene showing a loss in a Paso Fino relative to a Thoroughbred horse. (*F*) qPCR confirmation of exon in *UNC5C*. (*G*) qPCR confirmation of exon in *BMPR1B*. (*H*) qPCR confirmation of normal exon in *BMPR1B*.

cies across breeds were further evaluated by calculation of F_{ST} for each locus. The distribution of F_{ST} across the CNVRs ranged from −0.001 to 0.141 (Supplemental Table S9). Negative values should be considered as zero. Overall, the mean F_{ST} values of the seven CNVRs were similar, although slightly lower, to what has been seen for other types of loci (such as microsatellites) in the horse (Canon et al. 2000; Leroy et al. 2009). Collectively, 2.4% of the diversity for CNVs was found within breeds compared to the total. For a set of 15 microsatellite loci typed for the same five breeds, the mean F_{ST} was 0.095 (data not shown). The low values of F_{ST} for the CNVs are, at least, partly due to the small number of breeds tested and the fact that these breeds are not highly differentiated. However, for the CNVs, almost all of the differentiation was due to *LRRC37B*, where the Thoroughbred and Peruvian breeds were divergent from the other breeds, mainly due to a higher frequency of the WT allele in the Thoroughbred and to a higher frequency of the duplication allele in Peruvians.

Discussion

The development of different breeds of horses is a reflection of human selection and environmental adaptation that has occurred over the past 6000 yr (Vila et al. 2001). Horses have played an instrumental role in transportation, agriculture, and warfare and have been faithful companions to humans since their domestication. The genetic variants that underlie the phenotypic diversification of horse breeds are poorly understood. Moreover, the occurrence of CNVs in horses and their subsequent impact on phenotypic variation has not been investigated to date. In the present study, we describe the first analysis of copy number variation in the horse genome, focusing on 17 horses representing 15 phenotypically and ancestrally divergent breeds, including a gray donkey. Using a whole-exome array and the array comparative genomic hybridization methodology, we identified 2368 CNVs, representing 775 CNV regions and totaling ∼86.37 Mb of the horse genome. These data are consistent with estimates of copy number variation in humans and mice, suggesting that 5%–10% of the genome has undergone recent genomic rearrangements leading to copy number variation among individuals (Sharp et al. 2006; Stankiewicz and Lupski 2010). Collectively, these data represent the largest source of genetic variation identified in the horse genome to date.

By focusing our analysis toward the coding portions of the genome, we were able to increase our resolution for identifying smaller CNVs (one probe per 98 bp), and we were able to identify CNVs potentially having large effects on biological processes. An equally spaced whole-genome array consisting of a similar number of probes and using the calling criteria described in this study would increase the median size of a CNV to ∼18 kb. At this resolution, a conservative estimation of the false negative rate would be ∼50%. Recent studies in human populations indicate that CNVs have median and mean sizes of 729 bp and 8 kb, respectively (Mills et al. 2011), suggesting that smaller, unbalanced rearrangements are more common than previously estimated (Conrad et al. 2010). Array CGH and SNP arrays have been used to identify CNVs in cattle and dogs, and the average CNV size has ranged from 33 kb to 275 kb. However, close inspection of the probe spacing of the arrays used in these studies (200 bp to 50 kb) suggests that CNVs of the expected median size and smaller would be missed (Chen et al. 2009; Nicholas et al. 2009; Fadista et al. 2010; Liu et al. 2010; Hou et al. 2011; Kijas et al. 2011; Nicholas et al. 2011). Although we failed to identify CNVs in intergenic regions because of the genecentric design of the array, we felt that the benefit of identifying smaller genic CNVs outweighed the identification of larger genic and intergenic CNVs.

To identify CNVs potentially having large effects on phenotypes, we designed the study to examine an individual horse representing a phenotypically diverse and evolutionary divergent breed. As anticipated, we identified a large number of CNV regions given the rather small number of animals investigated, suggesting potential roles for these CNVs in shaping the phenotypic diversity displayed by the horses investigated. As observed in human populations and dog breeds (Mills et al. 2011; Nicholas et al. 2011), we anticipate that a small percentage of the CNVs will be unique to their respective breeds and will be ideal candidates for genotype/phenotype studies in horses. Nonetheless, a more comprehensive study will be needed to determine the population genetics of these CNVs in horses.

Our biological processes enrichment analysis indicated that CNVs primarily affected genes involved in sensory perception, signal transduction, and metabolism. We observed that many CNVs consisted of homozygous deletions, mostly of genes involved in sensory perception processes. This is consistent with human studies and reflects the location of these genes in segmentally duplicated regions (Mills et al. 2011). The complete loss of genes in healthy horses is not a novel observation, as it has been shown in several studies involving humans and mice (Freeman et al. 2006). The enrichment of CNVs in sensory genes could have a large effect on the diversification of horse traits linked to flight response and temperament, although this is merely speculative at this point. CNVs were also identified in several Mendelian genes; however, our analysis of the *PMEL* and *BMPR1B* suggests that these rearrangements are benign with respect to the phenotypes investigated. Nonetheless, some of the identified CNVs appear to be strong candidates for monogenic and quantitative traits in horses and will provide the foundation for future genotype/phenotype studies in larger cohorts of horses with well-defined phenotypes.

Collectively, this study represents the first investigation of CNVs in horses and identifies a substantial amount of genetic variation due to genomic rearrangements in the horse genome. Furthermore, some of the CNVs identified in this study may represent causal genetic variants for traits distinguishing the horse breeds. These data will provide the foundation for future studies into the phenotypic effects of specific CNVs in the horse genome, providing a better understanding of the role of genetic variation in influencing traits in horses such as athletic performance, reproduction, and disease.

Methods

DNA samples

Whole blood was collected from 15 different horse breeds and a gray donkey following signed informed consent of the horse owners. Genomic DNA was isolated from whole blood using a standard phenol-chloroform extraction including two phenol-chloroform-isoamyl (PCI) steps, followed by rinses with chloroform, isopropanol, and 70% ethanol. The samples were suspended in Qiagen EB buffer (Qiagen Sciences). This study was approved by the Clinical Research Review Committee, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University.

Array design

The exome array was designed to contain unique oligonucleotides that cover the majority of the exons of the annotated horse genes in the Ensembl gene database. DNA sequences of exons, including

60 bp of flanking intron sequence, were downloaded using the Ensembl Biomart database. Unique oligonucleotides (oligos) were selected from the sequences with OligoWiz2 (Nielsen et al. 2003; Wernersson and Nielsen 2005) using the following parameters: aim length = 60 bp; max oligo length = 60 bp; cross-hybridization minimum homology = 75%; cross-hybridization length = 15 bp; cross-hybridization maximum homology = 98%; cross hybridization length = 80%; minimum distance between oligos = 28 bp. To ensure the best filtering, oligos for protein-coding genes were selected separately from RNA genes. Also, pseudogenes, transposons, and retrotransposons were not included for oligo selection. All oligos were further filtered to create a final list of 418,576 unique oligos using the following criteria: melting temperature = 72°C–84°C; oligo length = 45–60 bp; average cross-hybridization score = 0.69; average folding score = 0.94; average low complexity score = 0.83. The final list of probes was randomly checked for uniqueness in the equine genome by BLAT analysis against the equCab2 reference genome. The final oligo set was densely tiled across 20,882 Ensembl protein-coding and RNA genes with an average exon tiling density of 98 bp (Supplemental Table S1). The custom array design was uploaded to Agilent's eArray web service and printed with Agilent's 60-mer SurePrint technology (Agilent Technologies Inc.).

Array comparative genomic hybridization

We performed CGH to identify copy number variants against the reference Thoroughbred genome through the use of the whole-genome exon array. A Sonic Dismembrator 500 (Fisher Scientific) was used to shear 6 μg of genomic DNA, followed by purification with an Invitrogen PureLink PCR Kit (Invitrogen). The sheared genomic DNA from the reference Thoroughbred was labeled with Alexa Fluor 555, and all other samples were labeled with Alexa Fluor 647 fluorescent dyes using the BioPrime Plus Labeling module (Invitrogen). The reference and experimental DNAs were mixed and denatured with 25 μl Cot-1 DNA (Invitrogen), 26 μl Agilent 10× Blocking Buffer, and 130 μl 2× High-RPM hybridization buffer prior to hybridization at 65°C for 20 h. Array slides were washed in Agilent wash buffer 1 for 5 min, followed by 5 min in prewarmed (37°C) wash buffer 2, and finally for 1 min in acetonitrile. After the washing procedure, arrays were scanned at a 2-μm resolution with 0.05 extended dynamic range (XDR) using an Agilent High Resolution Microarray Scanner 62505C (Agilent Technologies Inc.). The array data were extracted from the images using Feature Extraction 10.5 software. All arrays passed quality control (QC) checks performed by the Feature extraction software to ensure uniform signals in spots and low background-to-signal ratios (Agilent Technologies Inc.). We imported the data into Agilent's Genomics Workbench 5 and, using ADM-2 algorithm, threshold of 6, bin of 10, and a centralization threshold of 6, identified CNVs with respect to the reference sample. Additionally, we applied a custom probe filter to remove probes with signal intensities greater than three standard deviations from the mean. CNVs were required to have an average $\log_2$ ratio of 0.5 across at least three consecutive probes. Homozygous deletions were identified as regions with an average $\log_2$ ratio > 2.5 across three consecutive probes. The inability to amplify deletions by PCR was used as a second confirmation method. CNVs identified in the donkey sample for chromosome 1, which represents a translocation between chromosomes 4 and 31 in the horse, were mapped to the horse chromosome (Yang et al. 2004).

Polymerase chain reaction

We designed unique primer sets for PCR and quantitative PCR using Primer3Plus (http://www.primer3plus.com/) within 24 CNV genes identified by array CGH (Supplemental Table S10). All primer sets were tested to ensure that only the desired products were amplified.

Quantitative PCR

We performed qPCR on samples with and without the called CNVs and used control primers in the GAPDH gene and in exons flanking small CNVs in order to determine fold-changes in copy number. The qPCR was performed using SYBR GreenER regents for ABI Prism following the manufacturer's protocol (Invitrogen). The fold-changes were determined using a standard ΔΔCT method that compares CT values of a reference gene to the gene of interest (Livak and Schmittgen 2001). The fold-changes were normalized to a diploid number for a better comparison of copy number in all qPCR plots.

Confirmation of the smallest CNV

Standard PCR was used to amplify a CNV called to have both gains and losses in copy number. The PCR products from samples that were homozygous for the predicted size and heterozygous or homozygous for the deletion were inserted into a pCR2.1-TOPO plasmid following the manufacturer's protocol using a TOPO TA Cloning kit (Invitrogen). The selected colonies were grown overnight in Lysogeny broth (LB) with kanamycin. Plasmids were isolated from the cultures using a QIAprep Spin Miniprep kit (Qiagen Sciences). Confirmation of the inserts was performed by PCR and restriction digests with EcoRI. Plasmids containing and lacking the duplication were sequenced by the Texas A&M University DNA Technologies Core Laboratory using Sanger sequencing. The sequences obtained were aligned to each other as well as to the reference genome using ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/).

Chromosome enrichment

Enrichment of the horse chromosomes for CNVs was determined by merging all overlapping CNVs into copy number variant regions. The total length of all CNVRs on each chromosome was divided by the length of the chromosome to calculate the percent enrichment. Then, the total length of all CNVRs was divided by the length of the assembled genome to determine the genome enrichment. Enriched chromosomes were identified if their percent enrichment was greater than the percent enrichment of the entire genome.

Hierarchical clustering analysis of CNVs

Hierarchical clustering analysis of CNVs was performed by first extracting the $\log_2$ ratios from all oligos within the identified CNVs. Then, signals for each sample were imported into the Genesis software (Sturn et al. 2002) for clustering using the following parameters: Pearson correlation, hierarchical clustering, and complete linkage (Graz University of Technology: Institute for Genomics and Bioinformatics). We then compared the CNVs among all samples to identify shared CNVs by overlapping all CNVs from all samples, creating CNVRs. CNV sharing was classified by the number of samples with each CNV where 1 is an individual CNV and 17 is shared in all samples. Next, we compared the sharing of CNV genes among the samples by overlapping the gene lists from each sample and determining the total number of samples with CNVs in each gene.

Biological process enrichment analysis

We extracted all overlapping protein- and RNA-coding genes from the identified CNVs for analysis of gene biotypes, as determined by

Ensembl, and enrichment for biological terms in the gene list. Because the horse genome is poorly annotated in the biological process gene ontology database, we converted all genes to the human Ensembl orthologs. These genes were imported as genes in the DAVID Functional Annotation Tool, using the default settings (Huang et al. 2009a,b). The resulting terms were further grouped by manual inspection based on similarities in function to determine enrichment for specific biological processes in the horse genome (Supplemental Table S11).

All genes affected by a CNV were converted to RefSeq gene symbols and imported into the Online Mendelian Inheritance in Animals (OMIA) and Online Mendelian Inheritance in Man (OMIM) databases (http://www.ncbi.nlm.nih.gov/omia; http://www.ncbi.nlm.nih.gov/omim) using the batch import tool. The results provided a list of genes suspected to be involved in several phenotypes in animals.

Population analysis of CNV allele frequencies

One hundred and twenty-five horses from five different breeds were included in population analysis. Breed samples were evaluated, and individuals, when possible, were excluded to ensure that allele frequencies were not biased by close relationships. CNV Genepop for all individuals were determined for seven CNV loci using qPCR and PCR as described above. For simple loci, where only deletions or only duplications were detected (i.e., *BMPR1B*, *LRRC37B*, *WRN*, and *PMEL*), allele and genotype frequencies were determined directly. For complex loci, in which both duplications and deletions were present (i.e., ENSECAG00000006791, *PTPRC*, and *ABO*), the duplication/deletion and wild type/wild type were indistinguishable by our method of genotyping. Therefore, we estimated the frequencies of the WT/WT and Dup/Del genotypes as follows:

$$\mathrm{Dup/Del}_{\mathrm{estimated}} = \mathrm{WT/WT}_{\mathrm{observed}}\left(\mathrm{Dup}_{\mathrm{frequency}} + \mathrm{Del}_{\mathrm{frequency}}\right)$$

$$\mathrm{WT/WT}_{\mathrm{estimated}} = \mathrm{WT/WT}_{\mathrm{observed}} - \mathrm{Dup/Del}_{\mathrm{estimated}}.$$

Due to the limited number of individuals in each breed, the 95% confidence intervals for true allele frequency were calculated using the likelihood-ratio confidence interval method in the R statistical environment (www.r-project.org). The statistical analysis of all resulting genotypic frequencies was performed using Genepop software (http://genepop.curtin.edu.au). The breed and overall population allele frequencies for each CNVR were calculated using the following parameters: Option 5; Allele frequencies; and Diploid. The inter-population differentiation and global F_{ST} were calculated using the following parameters: Option 6; Allele identity (F-statistics) for all populations (global F_{ST}) and for all population pairs (inter-population differentiation F_{ST}); Fit to Ln(distance); Convert F-statistics to F/(1-F)-statistics; Minimum distance between samples, 0.0001; Number of permutations for Mantel test, 1000; Diploid.

Data access

The microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE32702.

Acknowledgments

We thank Dr. James Mickelson for critical reading of the manuscript, Dr. Charlie Love (Texas A&M University College of Veterinary Medicine and Biomedical Sciences) for DNA samples, Ted Sharpe (Broad Institute) for the horse segmental duplication map, and Drs. David Fry and Olivia Rudolphi for assistance with collecting samples. AgriLife Research, the Department of Veterinary Pathobiology, College of Veterinary Medicine and Biomedical Sciences, and the Link Equine Research Endowment, Texas A&M University provided funding for this study. This publication is based in part on work supported by King Abdullah University of Science and Technology (KAUST), award no. KUS-C1-016-04.

References

Boone PM, Bacino CA, Shaw CA, Eng PA, Hixson PM, Pursley AN, Kang SH, Yang Y, Wiszniewska J, Nowakowska BA, et al. 2010. Detection of clinically relevant exonic copy-number changes by array CGH. *Hum Mutat* **31:** 1326–1342.

Brunberg E, Andersson L, Cothran G, Sandberg K, Mikko S, Lindgren G. 2006. A missense mutation in PMEL17 is associated with the Silver coat color in the horse. *BMC Genet* **7:** 46. doi: 10.1186/1471-2156-7-46.

Canon J, Checa ML, Carleos C, Vega-Pla JL, Vallejo M, Dunner S. 2000. The genetic structure of Spanish Celtic horse breeds inferred from microsatellite data. *Anim Genet* **31:** 39–48.

Chen WK, Swartz JD, Rush LJ, Alvarez CE. 2009. Mapping DNA structural variation in dogs. *Genome Res* **19:** 500–509.

Chu MX, Liu ZH, Jiao CL, He YQ, Fang L, Ye SC, Chen GH, Wang JY. 2007. Mutations in BMPR-IB and BMP-15 genes are associated with litter size in Small Tailed Han sheep (*Ovis aries*). *J Anim Sci* **85:** 598–603.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464:** 704–712.

DeBolt S. 2010. Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* **2:** 441–453.

Fadista J, Thomsen B, Holm LE, Bendixen C. 2010. Copy number variation in the bovine genome. *BMC Genomics* **11:** 284. doi: 10.1186/1471-2164-11-284.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, et al. 2006. Copy number variation: New insights in genome diversity. *Genome Res* **16:** 949–961.

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, et al. 2008. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* **40:** 538–545.

Hoashi T, Muller J, Vieira WD, Rouzaud F, Kikuchi K, Tamaki K, Hearing VJ. 2006. The repeat domain of the melanosomal matrix protein PMEL17/GP100 is required for the formation of organellar fibers. *J Biol Chem* **281:** 21198–21208.

Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim ES, Matukumalli LK, Ventura M, Song J, VanRaden PM, et al. 2011. Genomic characteristics of cattle copy number variations. *BMC Genomics* **12:** 127. doi: 10.1186/1471-2164-12-127.

Huang D- W, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37:** 1–13.

Huang D- W, Sherman BT, Lempicki RA. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64.

Kijas JW, Barendse W, Barris W, Harrison B, McCulloch R, McWilliam S, Whan V. 2011. Analysis of copy number variants in the cattle genome. *Gene* **482:** 73–77.

Kim D, Ackerman SL. 2011. The UNC5C netrin receptor regulates dorsal guidance of mouse hindbrain axons. *J Neurosci* **31:** 2167–2179.

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, et al. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460:** 1011–1015.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.

Leroy G, Callede L, Verrier E, Meriaux JC, Ricard A, Danchin-Burge C, Rognon X. 2009. Genetic diversity of a large set of horse breeds raised in France assessed by microsatellite polymorphism. *Genet Sel Evol* **41:** 5. doi: 10.1186/1297-9686-41-5.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5:** e254. doi: 10.1371/journal.pbio.0050254.

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME, et al. 2010. Analysis of copy number variations among diverse cattle breeds. *Genome Res* **20:** 693–703.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-ΔΔ C(T)) method. *Methods* **25:** 402–408.

McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, Distl O, Guerin G, Hasegawa T, Hill EW, et al. 2012. A high density SNP array for the domestic horse and extant perissodactyla: Utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet* **8:** e1002451. doi: 10.1371/journal.pgen.1002451.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470:** 59–65.

Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM. 2009. The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res* **19:** 491–499.

Nicholas TJ, Baker C, Eichler EE, Akey JM. 2011. A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. *BMC Genomics* **12:** 414. doi: 10.1186/1471-2164-12-414.

Nielsen HB, Wernersson R, Knudsen S. 2003. Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Res* **31:** 3491–3496.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.

Sharp AJ, Cheng Z, Eichler EE. 2006. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* **7:** 407–442.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61:** 437–455.

Sturn A, Quackenbush J, Trajanoski Z. 2002. Genesis: Cluster analysis of microarray data. *Bioinformatics* **18:** 207–208.

Vila C, Leonard JA, Gotherstrom A, Marklund S, Sandberg K, Liden K, Wayne RK, Ellegren H. 2001. Widespread origins of domestic horse lineages. *Science* **291:** 474–477.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, et al. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326:** 865–867.

Wernersson R, Nielsen HB. 2005. OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res* **33:** W611–W615.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452:** 872–876.

Yang F, Fu B, O'Brien PC, Nie W, Ryder OA, Ferguson-Smith MA. 2004. Refined genome-wide comparative map of the domestic horse, donkey and human based on cross-species chromosome painting: Insight into the occasional fertility of mules. *Chromosome Res* **12:** 65–76.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10:** 451–481.

Received July 12, 2011; accepted in revised form February 29, 2012.