Whole genome sequence and analysis of the Marwari horse breed and its genetic origin

JeHoon Jun1†, Yun Sung Cho1,2†, Haejin Hu1, Hak-Min Kim1, Sungwoong Jho1, Priyvat Radhvi1, Kyung Mi Park3, Jeongheui Lim4, Woon Kee Paek4, Kyudong Han5,6, Andrea Manica7, Jeremy S Edwards8, Jong Bhak1,2,3*

From Asia Pacific Bioinformatics Network (APBioNet) Thirteenth International Conference on Bioinformatics (InCoB2014)
Sydney, Australia. 31 July - 2 August 2014

Abstract
Background: The horse (Equus ferus caballus) is one of the earliest domesticated species and has played an important role in the development of human societies over the past 5,000 years. In this study, we characterized the genome of the Marwari horse, a rare breed with unique phenotypic characteristics, including inwardly turned ear tips. It is thought to have originated from the crossbreeding of local Indian ponies with Arabian horses beginning in the 12th century.

Results: We generated 101 Gb (~30 × coverage) of whole genome sequences from a Marwari horse using the Illumina HiSeq2000 sequencer. The sequences were mapped to the horse reference genome at a mapping rate of ~98% and with ~95% of the genome having at least 10 × coverage. A total of 5.9 million single nucleotide variations, 0.6 million small insertions or deletions, and 2,569 copy number variation blocks were identified. We confirmed a strong Arabian and Mongolian component in the Marwari genome. Novel variants from the Marwari sequences were annotated, and were found to be enriched in olfactory functions. Additionally, we suggest a potential functional genetic variant in the TSHZ1 gene (p.Ala344>Val) associated with the inward-turning ear tip shape of the Marwari horses.

Conclusions: Here, we present an analysis of the Marwari horse genome. This is the first genomic data for an Asian breed, and is an invaluable resource for future studies of genetic variation associated with phenotypes and diseases in horses.

Background
The horse (Equus ferus caballus) was one of the earliest domesticated species and has played numerous important roles in human societies: acting as a source of food, a means of transport, for draught and agricultural work, and for sport, hunting, and warfare [1]. Horse domestication is believed to have started in the western Asian steppes approximately 5,500 years ago, and quickly spread across the Eurasian continent, with herds being augmented by the recruitment of local wild horses [2]. Domestication in the Iberian Peninsula might have represented an additional independent episode, involving horses that survived in a steppe refuge following the reforestation of Central Europe during the Holocene [3].

The horse reference genome has provided fundamental genomic information on the equine lineage and has been used for improving the health and performance of horses [1,4]. Horses exhibit 214 genetic traits and/or diseases that are similar to those of humans [5]. To date, several horse whole genomes have been sequenced and analyzed [4,6]. In 2012, the first whole genome re-sequencing analysis was conducted on the Quarter Horse breed to identify novel genetic variants [4]. In 2013, divergence times among horse fossils, donkey, Przewalski's horse, and several domestic horses were estimated, together with their
demographic history [6]. However, currently available whole genome sequences of modern horses only comprise western Eurasian breeds.

Over the centuries, more than 400 distinct horse breeds have been established by genetic selection for a wide number of desired phenotypic traits [7]. The Marwari (also known as Malani) horse is a rare breed from the Marwar region of India, and is one of six distinct horse breeds of India. They are believed to be descended from native Indian ponies, which were crossed with Arabian horses beginning around the 12th century, possibly with some Mongolian influence [8-10]. The Marwari horses were trained to perform complex prancing and leaping movements for ceremonial purposes [11,12]. The Marwari population in India deteriorated in the early 1900s due to improper management of the breeding stock, and only a few thousand purebred Marwari horses remain [12].

Here, we report the first whole genome sequence of a male Marwari horse as one of the Asian breeds and characterize its genetic variations, including single nucleotide variations (SNVs), small insertions/deletions (indels), and copy number variations (CNVs). To investigate relationships among different horse breeds, we carried out a genome-wide comparative analysis using previously reported whole genome sequences of six western Eurasian breeds [4,6], and single nucleotide polymorphism (SNP) array data of 729 horses from 32 worldwide breeds [13]. Our results provide insights into its genetic background and origin, and identify genotypes associated with the Marwari-specific phenotypes.

Results and discussion

Whole genome sequencing and variation detection

Genomic DNA was obtained from a blood sample of a male Marwari horse (17 years old) and was sequenced using an Illumina HiSeq2000 sequencer. A total of 112 Gb of paired-end sequence data were produced with a read length of 100 bp and insert sizes of 456 and 462 bp from two genomic libraries (Additional file 2: Figure S1, Figure S2). A total of 1,013,642,079 reads remained after filtering, and 993,802,079 reads were mapped to the horse reference genome (EquCab2.0 from the Ensembl database) with a mapping rate of 98.04%. (Additional file 1: Table S3). We assessed the mutational frequency at the single nucleotide level in the Marwari and compared it to estimates from other breeds (Additional file 1 Table S4). Interestingly, we found that the prevalent mutation types were not consistent among horse breeds. The mutation spectrum of the Marwari was dominated by C>T (G>A) transitions; a pattern which was also observed in the Icelandic, Norwegian Fjord, and Quarter horses. Conversely, the genomes of the Arabian, Standardbred, and Thoroughbred horses were dominated by A>G (T>C) transitions. A significant association between the mutation spectrum and horse breed (p-value < 0.001) was found when we applied a chi-square test using SPSS [14] to statistically compare the differences in the mutation spectrums among the breeds.

The Marwari genome consisted of 2,383,702 (40.2%) homozygous and 3,539,864 (59.8%) heterozygous SNVs (Table 1). Among them, 18,195 were found to be non-synonymous SNVs (nsSNVs). When the Marwari variants were compared to those previously reported from the genomes of other breeds [4,6] and the horse SNP database from the Broad Institute, 1,577,725 SNVs and 249,609 indels were novel variants. Of these, 4,716 variants (4,413 nsSNVs and 303 indels in coding regions) represented amino acid changes which were found in 2,770 genes (2,584 genes with nsSNVs, 279 genes with indels in coding regions, and 93 genes with nsSNVs and indels in coding regions simultaneously). To annotate the variants using well-known functional databases, human orthologs were retrieved from the Ensembl BioMart utility. A total of 1,970 of the 2,770 genes had human orthologs, and 1,896 genes were annotated using the DAVID Bioinformatics Resource 6.7 [15]. The genes with nsSNVs and/or indels in coding regions were highly enriched in olfactory functions (Additional file 1: Tables S5 and S6).
Copy number variations (CNVs) were identified using the R library “ReadDepth package” [16]. A total of 2,579 CNVs, including 869 gain and 1,710 loss blocks, were identified in the Marwari genome. The sizes ranged from 3 Kb to 6.43 Mb with an average length of 56 Kb. The CNV region (140 Mb in length) contained 2,504 genes which were duplicated (1,138 genes) or deleted (1,366 genes) (Additional file 1: Table S7). From the functional enrichment analysis, we found that the duplicated genes were enriched in olfactory function, whereas the deleted genes were enriched in immune regulation and metabolic processes (Additional file 1: Table S8, Table S9, Table S10, and Table S11).

Relatedness to other horse breeds

We constructed a phylogenetic tree using SNVs found in the whole genome data of the seven horse breeds (Arabian, Icelandic, Marwari, Norwegian Fjord, Quarter, Standardbred, and Thoroughbred) [4,6]. We identified 11,377,736 nucleotide positions that were commonly found in the seven horse genomes. A total of 25,854 nucleotide positions were used for phylogenetic analysis after filtering for minor allele frequency (MAF), genotyping rate, and linkage disequilibrium (LD). We found that the Marwari horse is most closely related to the Arabian breed (Additional file 2: Figure S6), while the Icelandic horse and Norwegian Fjord were the most distinct from the other breeds, all of which are known to have an Arabian ancestry [17,18]. Additionally, we found that the Marwari horse fell between Arabian and Mongolian horses, indicating their dual genetic influences on the Marwari horse as previously suggested [8-10].

We applied the STRUCTURE program [23,24] to estimate the genetic composition of the Asian horse breeds including the Marwari horse. For K = 2 groups, the Arabian horses were strongly separated from Mongolian horses, and the genetic composition of the Marwari horse was composed of alleles clustering with both the Mongolian horse (65.8%) and the Arabian horse (34.2%) (Figure 2). Other Asian breeds (Akhal Teke, Caspian, Table 1 Variants in the Marwari horse genome.

| Description       | SNVs         | Indels       |
|-------------------|--------------|--------------|
|                   | Homozygous   | Heterozygous | Novel   | Homozygous | Heterozygous | Novel   |
| Total             | 2,383,702    | 3,539,864    | 1,577,725 | 343,789    | 234,266     | 249,609 |
| INTERGENIC        | 1,565,078    | 2,352,370    | 1,060,195 | 215,679    | 153,412     | 164,564 |
| INTRAGENIC        | 3,474        | 5,134        | 1,919     | 556        | 332         | 329     |
| UPSTREAM          | 113,184      | 168,184      | 74,923    | 18,300     | 10,582      | 11,178  |
| DOWNSTREAM        | 111,918      | 166,365      | 75,592    | 17,866     | 11,504      | 12,203  |
| UTR_5_PRIME       | 600          | 684          | 279       | 171        | 27          | 25      |
| UTR_3_PRIME       | 1,188        | 1,660        | 802       | 264        | 129         | 149     |
| INTRON            | 569,725      | 817,541      | 351,960   | 89,594     | 57,856      | 60,614  |
| Noncoding exon variant | 3,259    | 4,368        | 3,433     | 280        | 198         | 244     |
| Synonymous mutation | 8,053     | 12,586      | 4,209     | 0          | 0           | 0       |
| Nonsynonymous mutation | 7,223      | 10,972      | 4,413     | 0          | 0           | 0       |
| Indels in coding region | 0          | 0           | 0         | 1,279      | 226         | 303     |
and Tuva) also showed genetic admixture between Arabian and Mongolian horses. From K = 3 to K = 5, the Marwari had high genetic components of both Arabian and Mongolian horses, whereas Akhal Teke and Caspian horses were mostly assigned to other clusters. These results indicate that the Marwari is genetically closely related to the Arabian and Mongolian horses. It is unclear whether the latter relationship represents direct genetic input from Mongolian horses or whether these horses are the closest population to the Indian ponies from which the Marwari is thought to have descended [8-10]. Further analysis including Indian ponies and Marwari horses will be required to distinguish the relative importance of these two scenarios, which are not mutually exclusive.

Phenotype association of the identified variants

To provide insight into the unique Marwari phenotypes, we investigated amino acid changes specific to this breed compared to those of other breeds (Arabian, Icelandic, Norwegian Fjord, Quarter, Standardbred, and Thoroughbred). A total of 343 amino acid changes in 236 genes were unique to the Marwari horse. Among the 236 genes, 75 genes included one or more amino acid changes predicted by the PolyPhen2 program to alter protein function [25] (Additional file 1: Table S12). Interestingly, the teashirt zinc finger family member 1 (TSHZ1) gene had a homozygous p.Ala344>Val variant (Figure 3). TSHZ1 is involved in transcriptional regulation of developmental processes and is associated with congenital aural atresia in humans, a malformation of the ear occurring in approximately 1 in 10,000 births [26,27]. Additionally, TSHZ1-deficient mice show malformations in the middle ear components [28]. Therefore, the A334V amino acid change in TSHZ1 is a strong candidate for the inward-turning ear tips characteristic of the Marwari breed. A future genomic comparison with the Kathiawari horse, which also has inward-turning ear tips, might support this prediction.

Selection in the equid lineage

We assessed the signatures of selection in the equid lineage using the $d_{NS}/d_S$ (nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site) ratio [56]. A consensus horse (equid) sequence was constructed by integrating all of the available genomes (Arabian, Icelandic, Marwari, Norwegian Fjord, Quarter, Standardbred, and Thoroughbred) in an attempt to remove breed specificity and to include an Asian breed component via the central Asian heritage of the Marwari (in contrast to the western Eurasian breeds for which whole genomes had been previously sequenced). A total of 7,711 out of 22,305 genes in the horse reference genome were substituted by the consensus sequences. Using the protein sequences of seven non-horse genomes (camel, pig, cow, minke whale, dog, mouse, and human), 5,459 orthologous gene families were constructed using OrthoMCL [57]. Using alignments of these gene families to estimate $d_{NS}/d_S$, we identified 188 genes under selection in the horse genome (Additional file 1: Table S13). The selected genes were particularly
enriched in immune response (immune effector process, leukocyte mediated immunity, positive regulation of immune system process, and defense response) and possible motor ability (T-tubule, muscle contraction, and regulation of heart contraction) functions (Additional file 1: Table S14). Over evolutionary time, the horse has developed increased speed and its musculature has become specialized for efficient strides [58,59]. It is therefore possible that the motor activity-associated genes we identified to be under positive selection have contributed to the muscular efficiency seen in modern horses.

**Conclusion**

Our study provides the first whole genome sequences and analyses of the Marwari, an Asian horse breed. Comparing the Marwari genome to the horse reference genome, approximately 5.9 million SNVs and 0.6 million indels, including 4,716 variants that cause amino acid changes, were identified. We found a clear Arabian and Mongolian component in the Marwari genome, although further work is needed to confirm whether modern Marwari horses also descended from Indian ponies. We analyzed the Marwari variants and found a candidate SNV determining its characteristic inward-
turning ear tips. Additionally, we investigated selection in the horse genome through comparisons with other mammalian genomes. By creating a consensus sequence that included information on an Asian breed, we found a number of genetic signatures of selection, providing insights into possible evolutionary and environmental adaptations in the equid lineage. The whole genome sequencing data from the Marwari horse provides a rich and diverse genomic resource that can be used to improve our understanding of animal domestication and will likely be useful in future studies of phenotypes and disease.

**Methods**

**Sample preparation and whole genome sequencing**

Genomic DNA was extracted from the blood of a 17 year old male Marwari horse with the XcelGene Blood gDNA Mini Kit (Xcelris Labs Ltd, Gujarat, India) following the manufacturer’s protocol. Two genomic libraries with insert sizes of 456 and 462 bp were constructed at Theragen BiO Institute (TBI), TheragenEtex, Korea. The genomic DNA was sheared using Covaris S series (Covaris, MS, USA). The sheared DNA was end-repaired, A-tailed, and ligated to paired-end adapters, according to the manufacturer’s protocol (Truseq DNA Sample Prep Kit v2, Illumina, San Diego, CA, USA). Adapter-ligated fragments were then size selected on a 2% Agarose gel, and the 520–620 bp band was extracted. Gel extraction and column purification were performed using the MinElute Gel Extraction Kit (Qiagen, CA, USA) following the manufacturer’s protocol. The ligated DNA fragments containing adapter sequences were enhanced via PCR using adapter-specific primers. Library quality and concentration were determined using an Agilent 2100 BioAnalyzer (Agilent, CA, USA). The libraries were quantified using a KAPA library quantification protocol. Based on the differences in sequencing depths were detected using the MinElute Gel Extraction Kit (Qiagen, CA, USA) following the manufacturer’s protocol. The ligated DNA fragments containing adapter sequences were enhanced via PCR using adapter-specific primers. Library quality and concentration were determined using an Agilent 2100 BioAnalyzer (Agilent, CA, USA). The libraries were quantified using a KAPA library quantification kit (Kapa Biosystems, MA, USA) according to Illumina’s library quantification protocol. Based on the qPCR quantification, the libraries were normalized to 2 nM and denatured using 0.1 N NaOH. Cluster amplification of denatured templates was performed in flow cells according to the manufacturer’s protocol (Illumina, CA, USA). Flow cells were paired-end sequenced (2 × 100 bp) on an Illumina HiSeq2000 using HiSeq Sequencing kits. A base-calling pipeline (Sequencing Control Software, Illumina) was used to process the raw fluorescent images and the called sequences.

**Filtering and mapping processes**

Before the mapping step, raw reads were filtered using NGS QC toolkit version 2.3 (cutoff read length for high quality, 70%; cutoff quality score, 20) [60]. After the filtering step, clean reads were mapped to the horse reference genome (Ensembl EquCab2.0, release 72) [1] with BWA version 0.7.5a [61] with minimum seed length (-k 15) and Mark shorter split hits as secondary (-M). We realigned the reads using the GATK [62] IndelRealigner algorithm to enhance the mapping quality, and marked duplicate reads using MarkDuplicates from picard-tools version 1.92 (http://broadinstitute.github.io/picard/).

**De novo assembly of unmapped reads**

We extracted unmapped sequences from aligned Marwari BAM files. To find Marwari specific genomic regions, we assembled unmapped reads using SOAPdenovo2 [63] with “all” mode and multiple K values (ranged from 23 to 63). A total of 120,159 contigs were obtained, and N50 length was 227bp. To identify whether these contigs are in non-reference regions, we aligned contigs to the horse reference genome. A total of 25,614 contigs were not aligned to the reference genome. The non-reference sequences were further analyzed by BLAST to NCBI protein and DNA sequence databases with the criteria E ≤ 10⁻⁵ and identity ≥ 70%.

**Variant detection and annotation**

Putative variant calls were made using the SAMtools version 0.1.16 [64] mpileup command. In this step, we used the -E option to minimize the noise resulting from pairwise read alignments, and the -A option to use regardless of insert size constraint and/or orientation within pairs. Variants were called using bcftools and then filtered using vcutils varFilter (minimal depth of 8, maximal depth of 100, Phred scores of SNP call ≥ 30, and no indel present within a 2 bp window) as previously reported [6]. SnpEff [65] was used to annotate the variants. To find unique variants for the Marwari horse, SNVs and small indels were further compared with the horse SNP database that was identified by the Horse Genome Project (http://www.broadinstitute.org/mammals/horse/snp), and other previously reported horse breed genomes [4,6]. Copy number variants based on the differences in sequencing depths were detected using R library “ReadDepth package” with default options. The ReadDepth calculated the thresholds for copy number gain (2.642) and loss (1.380).

**Phylogenetic tree construction**

Genotype data were extracted from a total of 11,377,736 single nucleotide positions, which were shared and sufficiently covered regions (> 8 × depth), in the seven horse whole genome data (Arabian, Icelandic, Marwari, Norwegian Fjord, Quarter, Standardbred, and Thoroughbred) [4,6]. The genotyping data were merged, and then filtered to remove those SNP with a genotyping rate of < 0.05 and allele frequency > 0.2 using PLINK [66]. SNPs that were in linkage disequilibrium (LD) were also removed: the merged files were pruned for r²
< 0.1 in PLINK, considering 100 SNP windows and moving 25 SNPs per set (indep-pairwise 100 25). After the filtering and pruning process, 25,854 SNPs remained and were used for the phylogenetic analysis. RAxML version 7.28-ALPHA [67] was used to generate the parsimony starting trees, and RAxML-Light version 1.0.9 [68] was used to carry out tree inference with the GTR+GAMMA model of nucleotide substitutions. A total of 100 bootstrap trees were generated for each phylogeny. The resulting tree was drawn by MEGA6 [69].

MDS and population structure analyses
Equine SNP array data of 729 individuals belonging to 32 horse breeds were obtained from a previous report [13]. The Marwari horse data used in this analysis were selected from 54,330 nucleotide positions that were derived from the SNP array data. The SNP array and Marwari data were filtered and pruned to remove SNPs with the same cutoffs described above, except that the MAF option was set to maf < 0.05. A total of 10,554 single nucleotide positions were used for the following comparative analyses.

The MDS plot was drawn in R [70] using the "MASS" library and "canberra" distance metric. STRUCTURE version 2.3.4 [23,24] was used to cluster Asian breeds based on genetic similarity, investigating K values from 2 to 5. Each run for a given K value consisted of a 15,000 steps burn-in and 35,000 MCMC repetitions. We applied a default admixture model and a default assumption that allele frequencies were correlated. The convergence of STRUCTURE runs was evaluated by the equilibrium of alpha. Individual and population clump files were produced with Structure Harvester [71] and visualized in Distruct1.1 [72].

Orthologous gene family
Protein sequences of cow (Bos taurus), dog (Canis familiaris), human (Homo sapiens), mouse (Mus musculus), and pig (Sus scrofa) were downloaded from the Ensembl database version 72. Protein sequences of minke whale (Balaenoptera acutorostrata) [73] and camel (Camelus bactrianus) [74] were obtained from the original publications. A total of eight species were used to identify orthologous gene clusters with OrthoMCL 2.0.9. Pairwise sequence similarities between all protein sequences were calculated using BLASTP with an e-value cutoff of 1E-05. On the basis of the BLASTP results, OrthoMCL was used to perform a Markov clustering algorithm with inflation value (-I) of 1.5. The OrthoMCL was run with an e-value exponent cutoff of -5 and percent match cutoff of 50%. In total, 5,501 orthologous groups were shared by all eight species. The representative sequences for each gene cluster were selected using the longest horse transcript and the corresponding protein sequences of the other species. BLASTP searches (E-value 1E-5 cutoff) between horse and all the other species were used in this process. Finally, we identified 5,459 1:1:1:1:1:1:1:1 orthologs.

Molecular evolutionary analysis
The phylogenetic tree was constructed from 5,459 single copy orthologous genes. CODEML in PAML 4.5 [75] was used to estimate the \( d_{NS}/d_{S} \) ratio, where \( d_{NS} \) indicates nonsynonymous substitution rate and \( d_{S} \) indicates synonymous substitution rate. The \( d_{NS}/d_{S} \) ratio along the horse branch (free-ratio model) and \( d_{NS}/d_{S} \) ratio for all branches (one-ratio model) were calculated as the branch model. We also applied the branch-site model to further examine potential positive selection [76]. The LRTs (likelihood ratio tests) were applied to assess statistical significance of the branch-site model. We supposed that positively selected genes are that of having a higher \( d_{NS}/d_{S} \) ratio with the free-ratios model than that with the one-ratio model and having p-value < 0.05 from branch-site model.

Availability of supporting data
Whole genome sequence data was deposited in the SRA database at NCBI with Biosample accession numbers SAMN02767683. SRA of whole genome sequencing can be accessed via reference numbers SRX535352. The data can also be accessed through BioProject accession number PRJNA246445 for the whole genome sequence data.

Additional material

**Additional file 1:** Table S1 Sequence generation and mapping information, Table S2 Statistics of de novo assembly using unmapped reads, Table S3 Mutation rate of the Marwari horse and other horses. SNP rate, heterozygosity, and average depth of coverage for autosomal chromosomes (minimum depth > 8, maximum depth < 100), Table S4 The mutation spectrums of horses. The dominant mutation spectrums are shown in bold, Table S5 Functional annotation clustering of putative Marwari horse nsSNVs/indels in coding regions using DAVID, Table S6 Annotation on KEGG pathway of putative Marwari horse nsSNVs/indels in coding regions using DAVID tools, Table S7 CNVs of Marwari horse genome, Table S8 Functional annotation clustering of duplicated genes in Marwari horse genome using DAVID, Table S9 Annotation on KEGG pathway of duplicated genes in Marwari horse genome using DAVID tools, Table S10 Functional annotation clustering of deleted genes in Marwari horse genome using DAVID, Table S11 Annotation on KEGG pathway of deleted genes in Marwari horse genome using DAVID tools, Table S12 Polyphen-2 prediction result of Marwari-specific amino acid changes. A total of 343 amino acid changes in 236 genes and then 78 amino acid changes in 75 genes were predicted to alter protein function in PolyPhen2, Table S13 Positively selected genes (PSGs) in horse. PSGs were calculated by comparing horse and seven mammalians. 188 genes were identified as PSG. (\( \omega \) value of free ratio model > \( \omega \) value of one ratio model, and P-value of branch-site model < 0.05), and Table S14 Functional annotation clustering of positive selection genes in horse using DAVID tools.
Additional file 2: Figure S1 The BioAnalyzer profiles of two libraries. Figures indicating fluorescence unit (FU) and DNA fragment size. Numbers mean the area under the peak within the regions. Peaks of 35bp (green) and 10,380bp (purple) are lower and upper markers. (A) The BioAnalyzer profile of library targeting fragment size of 576 bp (insert size of 456 bp) (B) The BioAnalyzer profile of library targeting fragment size of 582 bp (insert size of 462bp), Figure S2 Base quality distribution. Figure indicating the distribution of phred quality score and number of bases, Figure S3 Mating proportion distribution of selected reads. The mode value of mapping quality is 60 (83.61% of total), and average mapping quality is 53.16, Figure S4 Mapping rate distribution. The Mapping rates were dropped when increasing the mapping quality cutoff. We used unfiltered data for next analysis, Figure S5 A total of 25,614 novel contigs were analyzed using BLAST to a NCBI protein database, and Figure S6 Maximum likelihood tree calculated from 25,854 single nucleotide polymorphisms in whole genome data of seven horses. Maximum likelihood tree created from 25,854 single nucleotide polymorphisms. Percent bootstrap result supporting all branches calculated from 100 replicates is shown.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived and designed the experiments: JB and WKP. Performed the experiments: KMP. Analyzed the data: JHJ, SJho, YSC, HJH, and JL. Study design, subject recruitment, and sample preparation: PG and JB. Data interpretation: JHJ, SJho, YSC, HJH, and JL. Wrote the paper: JHJ, YSC, HMK, HJH, JE, KH, AM, PG, and JB.

Acknowledgements
This work was supported by the National Research Foundation of Korea (2008-2004707 and 2013M3A9A5047052), the Industrial Strategic Technology Development Program, 10040231, ‘Bioinformatics platform development for next generation bioinformation analysis’ funded by the Ministry of Knowledge Economy (MKE, Korea), and the National Science Museum (NMK, Korea). The Equestrian Club of Gujarat and its office-bearer Mr. Virendra Kankaria provided samples of a registered Marwari horse for the study. We also thank Xcelris Genomics, Ahmedabad, India, and Dr Surendra Chikara, Ms. Arpita Ghosh, and the technical team of Xcelris for their work in DNA extraction, library preparation and shipment to GRF, South Korea. Authors thank TheragenEtex for supporting the research by providing GRF with computational and experimental resource for the NGS analyses. Publication costs were supported by the National Science Museum (NMK, Korea).

This article has been published as part of BMC Genomics Volume 15 Supplement 9, 2014: Thirteenth International Conference on Bioinformatics (InCoB2014): Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S9.

Authors' details
1Personal Genomics Institute, Genome Research Foundation, Suwon 443-270, Republic of Korea. 2The Genomics Institute, Biomedical Engineering Department, UNIST, Ulsan, Republic of Korea. 3Theragen BIo Institute, TheragenEtex, Suwon 443-270, Republic of Korea. 4National Science Museum, Daeyeon 305-705, Republic of Korea. 5Department of Nanobiomedical Science & BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook University, Cheonan, 330-714, Republic of Korea. 6DKU-Theragen institute for NGS analysis (DTiNa), Cheonan, 330-714, Republic of Korea. 7Evolutionary Ecology Group, Department of Zoology, University of Cambridge, Cambridge, UK. 8Department of Chemistry and Chemical Biology, Department of Molecular Genetics and Microbiology, Department of Chemical and Nuclear Engineering, Cancer Research and Treatment Center, University of New Mexico, Albuquerque, NM 87106, USA.

Published: 8 December 2014

References
1. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellow RR, Blocker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MC, Pason JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Birns MM, Crowdhary BP, Coleman SJ, De Maeyer G, Fyrc S, Gue´rin G, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 2009, 326:865-867.
2. Warnuth V, Eriksson A, Bowser MA, Barker G, Barrett E, Hanks BK, Li S, Lomtatidvl D, Ochri-Goryaeva M, Sizonov GV, Suyonov V, Manica A. Reconstructing the origin and spread of horse domestication in the Eurasian steppes. Proc Natl Acad Sci USA 2012, 109:8202-8206.
3. Warnuth V, Eriksson A, Bowser MA, Cañon J, Cothan G, Distl O, Gómez-Kulllis ML, Hunt H, Luís C, do Mar Oom M, Yupaquique JT, Zábeek T, Manica A. European Domestic Horses Originated in Two Holocene Refugia. PLoS One 2011, 6:e18194.
4. Doan R, Cohen ND, Sayyaf S, Jayhoff N, Johnson CD, Dindor SD. Whole-Genome Sequencing and Genetic Variant Analysis of a Quarter Horse Mare. BMC Genomics 2012, 13:78.
5. Online Mendelian Inheritance in Animals[http://omia.angis.org.au/].
6. Orlando L, Ginhinac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PL, Fumagalli M, Wistrup JT, Raghavan M, Korneliusen T, Malaspina AS, Vogt J, Sklazicky D, Kelstrap CD, Vinther J, Dolanac A, Stenderup J, Velazquez AM, Cahill J, Rasmussen M, Wang X, Min J, Zazula GD, Seguin-Orlando A, Mortensen C, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 2013, 499:74-81.
7. Hendricks B. International Encyclopedia of Horse Breeds. Norman: University of Oklahoma Press; 1995.
8. Eleon Hartley Edwards: The Encyclopedia of the Horse New York: Dorling Kindersley; 1994.
9. Wendy Doniger: The Hindus: An Alternative History. New Delhi: Penguin Books; 2009.
10. Behl R, Behl J, Gupta N, Gupta SC. Genetic relationships of five Indian horse breeds using microsatellite markers. Animal 2007, 4:483-488.
11. Duton Judith: Storey's Illustrated Guide to 96 Horse Breeds of North America. North amds: Storey Publishing, 2005.
12. Gupta AK, Chauhan M, Tandon SN, Sonia : Genetic diversity and bottleneck studies in the Marwari horse breed. J Genet 2005, 84:295-301.
13. Wendt J, McElroy J, Rendahl AK, Walberg SJ, Anderson LS, Alexson J, Bailey E, Bannasch D, Birns MM, Borges AS, Brana P, da Cama Marinho A, Capomaccio S, Capelli K, Cothan EG, Distl O, Fox-Chlipman L, Graves KT, Guerin G, Haase B, Hasegawa T, Hemmann K, Hill EW, Leeb T, Lindgren G, Loh L, Lopes MS, McGivney BA, Mikko S, Orr N, et al. Genomewide analysis reveals selection for important traits in domestic horse breeds. PLoS Genet 2013, 9:e1003211.
14. IBM Corp. IBM SPSS Statistics for Windows, Version 22.0. NY: IBM; 2013.
15. Huang da W, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R, Wall J, Kondrashov A, Krzywinska G, Ashburner M, Tarkin S, Olsowski W, Keshishian H, Ray M, Tu S, Hupp T, Manica A, Bek T, Manica A, Bek T, Manica A, Bek T, Manica A. A method and server for predicting damaging missense mutations. Nat Methods 2010, 7:248-249.
16. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerassimova E, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010, 7:248-249.
40. Brunberg E, Andersson L, Cothran G, Sandberg K, Mikko S, Lindgren G.  
39. Spirito F, Charlesworth A, Linder K, Ortonne JP, Baird J, Meneguzzi G.  
38. Makvandi-Nejad S, Hoffman GE, Allen JJ, Chu E, Gu E, Chandler AM.  
37. Brooks SA, Lear TL, Adelson DL, Bailey E.  
36. Wagner HJ, Reissmann M.  
35. Aleman M, Riehl J, Aldridge BM, Lecouteur RA, Stott JL, Pessah IN.  
34. Brooks SA, Gabreski N, Miller D, Brisbin A, Brown HE, Streeter C, Mezey J,  
33. Brooks SA, Bailey E.  
32. Brooks SA, Lear TL, Adelson DL, Bailey E.  
31. Wagner HJ, Reissmann M.  
30. Bellone RR, Forsyth G, Leeb T, Archer S, Sigurdsson S, Imsland F, Mauceli E,  
29. Jun et al.  
28. Macfadden BJ.  
27. Yellon RF, Branstetter BF 4th.  
26. ALTMANN F.  
25. Tron RC, White SD, Bannasch DL.  
24. Brooks SA, Baleone RR, Forsyth G, Leeb T, Archer S, Sigurdsson S, Imsland F, Mauceli E,  
23. Jun et al.  
22. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
21. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
20. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
19. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
18. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
17. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
16. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
15. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
14. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
13. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
12. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
11. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
10. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
9. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
8. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
7. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
6. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
5. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
4. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
3. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
2. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
1. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P,  
0. Bolkow L, Scharf K, Li P, Cloud J, Schmidt J, Sullards S, Lypowsky J, Scharf K, Li P.
68. Stamatakis A, Aberer AJ, Goll C, Smith SA, Berger SA, Izquierdo-Carrasco F. RAxML-Light: a tool for computing terabyte phylogenies. Bioinformatics 2012, 28:2064-2066.

69. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol 2013, 30:2725-2729.

70. Ihaka R, Gentleman R: R: A Language for Data Analysis and Graphics. J Comput Graph Stat 1996, 5:299-314.

71. Earl DA, Vonholdt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 2012, 4:359-361.

72. Rosenberg NA. DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 2004, 4:137-138.

73. Yim HS, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, Oh HM, Lee JH, Yang EC, Kwon KK, Kim YJ, Kim TW, Kim W, Jeon JH, Kim SJ, Choi DH, Jho S, Kim HM, Ko J, Kim H, Shin YA, Jung HU, Zheng Y, Wang Z, Chen Y, Chen M, Jiang A, Li E, Zhang S, Hou H, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet 2014, 46:88-92.

74. Ji R, Cui P, Ding F, Gong J, Gao H, Zhang H, Yu J, Hu S, Meng H. Monophyletic origin of domestic bactrian camel (Camelus bactrianus) and its evolutionary relationship with the extant wild camel (Camelus bactrianus ferus). Anim Genet 2009, 40:377-382.

75. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007, 24:1586-1591.

76. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 2005, 22:2472-2479.

doi:10.1186/1471-2164-15-S9-S4

Cite this article as: Jun et al.: Whole genome sequence and analysis of the Marwari horse breed and its genetic origin. BMC Genomics 2014 15(Suppl 9):S4.

Submit your manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit