Open Access
Asian Australas. J. Anim. Sci.
Vol. 27, No. 9 : 1236-1243 September 2014
http://dx.doi.org/10.5713/ajas.2013.13694
www.ajas.info
pISSN 1011-2367 eISSN 1976-5517

Thoroughbred Horse Single Nucleotide Polymorphism and Expression Database: HSDB

Joon-Ho Lee[1,a], Taeheon Lee[2,a], Hak-Kyo Lee[1], Byung-Wook Cho[3], Dong-Hyun Shin[2], Kyoung-Tag Do[4], Samsun Sung[5], Woori Kwak[5,6], Hyeon Jeong Kim[5], Heebal Kim[2,6], Seoae Cho[5,*], and Kyung-Do Park[1,*]

[1] Genomic Informatics Center, Hankyong National University, Anseong 456-749, Korea

ABSTRACT: Genetics is important for breeding and selection of horses but there is a lack of well-established horse-related browsers or databases. In order to better understand horses, more variants and other integrated information are needed. Thus, we construct a horse genomic variants database including expression and other information. Horse Single Nucleotide Polymorphism and Expression Database (HSDB) (http://snugenome2.snu.ac.kr/HSDB) provides the number of unexplored genomic variants still remaining to be identified in the horse genome including rare variants by using population genome sequences of eighteen horses and RNA-seq of four horses. The identified single nucleotide polymorphisms (SNPs) were confirmed by comparing them with SNP chip data and variants of RNA-seq, which showed a concordance level of 99.02% and 96.6%, respectively. Moreover, the database provides the genomic variants with their corresponding transcriptional profiles from the same individuals to help understand the functional aspects of these variants. The database will contribute to genetic improvement and breeding strategies of Thoroughbreds. (**Key Words:** Database, Thoroughbred, Variants, Expression, Horse)

INTRODUCTION

The Thoroughbred is an important animal and a favorite breed for use in the horse racing industry. The speed and agility of Thoroughbred horses have resulted in the emergence of an industry involved in the breeding, training, and racing of elite racehorses worth many billions of dollars (Gordon, 2001). The genome-wide analysis of the horse has grown rapidly in recent years and is expected to impact the racing capacity of Thoroughbreds (Petersen et al., 2013). In horse genomics studies, SNP is a valuable resource for functional genomics and for annotating functional genes on the horse genome. For example, strong candidate genes, such *MSTN*, related to racing performance can be searched using single nucleotide polymorphism (SNP) chips at the genomic level. There is a great interest in SNPs since a catalog of SNPs is expected to facilitate large-scale studies in association genetics, functional and pharmaco-genomics, population genetic and evolution biology (Sherry et al., 2001).

Genome databases are comprehensive public repositories for genome mapping data from animal species including cow, pig and so forth (Sherry et al., 2001; Hubbard et al., 2002). Within a short span of ten years, information on the structure and organization of the horse genome has grown exponentially (Chowdhary and

* Corresponding Authors: Seoae Cho. Tel: +82-2-876-8820, Fax: +82-2-876-8827, E-mail: seoae@cnkgenomics.com / Kyung-Do Park. Tel: +82-31-670-5332, Fax: +82-31-675-5331, E-mail: doobalo@hknu.ac.kr
[2] Department of Agricultural Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-742, Korea.
[3] Department of Animal Science, College of Life Sciences, Pusan National University, Miryang 627-702, Korea.
[4] Department of Equine Sciences, Sorabol College, Gyeongju 780-711, Korea.
[5] C&K Genomics, Seoul National University Research Park, Seoul 151-919, Korea.
[6] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea.
[a] These two first authors contributed equally to this work.
Submitted Nov. 4, 2013; Revised Feb. 20, 2014; Accepted Jun. 21, 2014

Raudsepp, 2008). A whole genome database of the horse was constructed by the Horse Genome Project at the Broad Institute (Wade et al., 2009). Also, several groups have studied horse genomics by sequencing the horse genome and using SNP chips (Wade et al., 2009; Hill et al., 2010; Park et al., 2012). While the horse reference genome is available, genome-wide linkage maps of the horse as well as quantitative trait mapping were not integrated into the database.

Although the horse is an important animal, there are only a few genome browsers dedicated to the horse such as the Model Organism Database for Horses and University of California Santa Cruz Genome Browser for horse. The known genomic information related to horses is a small portion of the whole horse genome and in order to better understand horses, more variants and other integrated information are needed. Thus, we construct a horse genomic variants database from a Thoroughbred population using next-generation sequencing (NGS) technologies. Although only limited genomic information is available on horses, integration of various NGS data on horses will accelerate the progress of horse genome research. Therefore, we integrated genes based on SNPs from NGS data of Thoroughbreds, expression from RNA-seq results and information on linkage disequilibrium (LD).

This database provides horse genomic variants and transcriptome information, which is easy to apply to other data for both researchers and breeders. (Horse single nucleotide polymorphism and expression database (HSDB) provides the unexplored genomic variants which still remain to be identified in the horse genome including rare variants by using horse population genome sequencing. Moreover, the database provides the genomic variants with their corresponding transcriptional profiles from the same individuals to help understand the functional aspects of these variants.

MATERIALS AND METHODS

Ethics

This study was carried out in strict accordance with recommendations in the Guide for the Care and Use of Laboratory Animals of Pusan National University and the Korean Racing Authority. In addition, all experimental procedures used in this study related to animals were approved by the Institutional Animal Care and Use Committee of the Pusan National University (PNU-2013-0417).

Whole genome re-sequencing

Two set of whole-blood samples were collected from 18 Thoroughbred racing horses from the Korean Racing Authority. Genomic DNA was extracted and the DNA quality was checked with agarose gel electrophoresis and fluorescence-based quantification tests. Constructing and sequencing library were carried out using Illumina's TruSeq DNA Sample Preparation kits and Hiseq2000 protocols.

RNA-seq library preparation and sequencing

Muscle and blood samples were collected before and after exercise from four Thoroughbred racing horses. Total RNA from the resulting 16 samples from four horses were isolated using TRIzol (Invitrogen) and RNeasy RNA purification kits with DNase treatment (Qiagen). Isolated mRNA was reverse transcribed into double strand cDNA. Constructing and sequencing RNA-seq library of each sample were carried out using Illumina Hiseq2000 protocols.

Identification of single nucleotide polymorphism: whole genome re-sequencing

The quality of raw data was checked by FastQC. To reduce false positive SNP, we trimmed the 3'-end of reads to have a minimum phred-scaled quality score of over 20. Paired-end sequence reads were aligned to the reference horse genome (EquCab2) with Bowtie2 (version 2.0.0-beta6) (Langmead and Salzberg, 2012) using very-sensitive and no-mixed mode option. For 4 Horses data encoded Phred Scale Quality Score with Illumina 1.5 with—phred64 option. Picard tools, SAMtools (Li et al., 2009), and genome analysis toolkit (GATK 2.1.8) (McKenna et al., 2010) were used for downstream processing and variant calling. Downstream processing was carried out using typical GATK pipeline. For the base quality score recalibration (BQSR) step of GATK, Dstitution calls were made with GATK UnifiedGenotyper and the variants were discarded if i) recalibrated quality score was less than 30, ii) three SNPs existed within a 10 base pair window, iii) SNPs existed in a detected insertion and deletion (INDEL), iv) the number or proportion of reads, which have mapping quality score of 0 are bigger than 4% or 10%, respectively, v) the number of alternative allele was bigger than one (multi-allele type).

Identification of single nucleotide polymorphisms: RNA-seq

For SNP detection using RNA-seq data, we pooled the raw data of four samples (muscle before exercise, muscle after exercise, blood before exercise, and blood after exercise) of each individual. Pooled paired-end sequences were aligned to the same reference horse genome using Tophat (version 2.0.4) (Trapnell et al., 2009) with very-sensitive option. Downstream processing was carried out with the same procedure as the whole genome re-

sequencing except for the local realignment process because the mapping result of Tophat is not proper for the local realignment step.

Expression from RNA-seq

Twenty-four sets of transcriptome data were generated for muscle and blood from six horses both before and after exercise. TopHat, (used to align RNA sequences to a genome in order to identify exon-exon splice junctions) was used to map the sequences to a horse reference genome, annotated using the EquCab2 and determined gene expression as discrete measurement. The R package edgeR (Robinson et al., 2010) was used to identify differentially expressed genes (DEGs), which is based on a negative binomial model, to examine differential expression of replicated count data.

Single nucleotide polymorphisms chip

The SNP chip data used in this study are from Kim et al. (2013) and comes from 11 Thoroughbred horses genotyped using EquineSNP50 Genotyping BeadChip (Illumina, Inc., San Diego, CA, USA).

Annotation and building database and linkage disequilibrium

Variant files of each individual were merged using VCF tools (Danecek et al., 2011) and then annotated using snpEff 3 (Cingolani et al., 2012). Ensembl general feature format (GTF, gene sets) information was used to build Equcab2.68 snpEff database. The result of snpeff annotation was modified to make SNP database with SQLite3. Linkage disequilibrium was calculated by Haploview (Barrett et al., 2005).

RESULTS

Contents of the horse single nucleotide polymorphism and expression database

HSDB used comprehensive genomic information source from horse with human and mouse to confirm gene predictions that have been integrated with external data. HSDB is available at http://snugenome2.snu.ac.kr/HSDB. This database displays novel SNPs detected using RNA-seq as well as SNPs from re-sequencing data. Also this database displays an expression level and LD map for SNPs in the selected genes. The web interface allows interactive use of the information related to the horse variants. The interface consists of five menus: Introduction, SNP Search, Advanced Search, Expression Search, and Contact (Figure 1). Users can obtain variant information from the SNP Search and Advanced Search menu, and expression information from the Expression Search menu. Genetic variants and expression information can be searched by gene symbol or SNP position and the query will retrieve an array of results including a distribution of variants, gene information, expression and differentially expressed test results.

The SNP Search page provides a simple search function for detailed SNP information. The SNPs were represented in different colors according to the characteristic of the distribution of variants, i.e., if a SNP exists in an intron region its position is marked pink, and non-synonymous SNPs that lead to an amino acids change in the exon is represented in blue. The details of SNPs are reported in a table which includes predicted effects of variants (e.g. splice site donor) by using snpEff and the impact (High, Moderate, Low, Modifier) defined by snpEff. User can select the source of the SNP information: re-sequencing or RNA-seq. Moreover, there is gene description including gene function, human and mouse orthologous genes information and related gene ontology terms. The Advanced Search page is similar to the SNP Search page, but the user can search SNPs with additional options catered to their interests including known SNPs or novel SNPs, the effect, quality options include depth, quality and frequency, etc. The Expression Search page consists of four menus: 'DEG between before and after exercise', 'Basic Search', 'Bulk Query' and 'Advanced Tools'. The user can search the expression information and find information on DEGs between before and after exercise from blood and muscle. The Advanced Tools of Expression Search page has several filtering option, tissue, chromosome, log fold change, p-value, false discovery rate corrected p-value and which state (before and after) is higher average values.

Identification genic single nucleotide polymorphisms and differentially expressed genes

We analyzed both genomic and transcriptomic sequences from Thoroughbred individuals. Users can explore the relationship between identified genomic variants and its corresponding variants from the RNA-seq data. We made it possible that analyses integrate genome and transcriptome data across multiple individuals and reveal extensive variation at both levels.

We identified a total of 3,418,393 non-redundant SNPs from genic region from the re-sequencing of 18 samples (3,356,634 SNPs) and RNA-seq of 4 samples (443,725 SNPs) (Supplementary Tables S1 and S2**)** and 14.8% loci were shared between the two datasets. Approximately 10% of the SNPs overlapped with loci of dbSNP; 346,489 (10.32%) of re-sequencing and 56,931 (12.83%) of RNA-seq.

The SNPs were classified into several categories based on the region of their location (Table 1). A large number of SNPs exist in the intron region for both re-sequencing and RNA-seq data. The large number of RNA-seq SNPs in intronic regions is similar to previous reports of a large

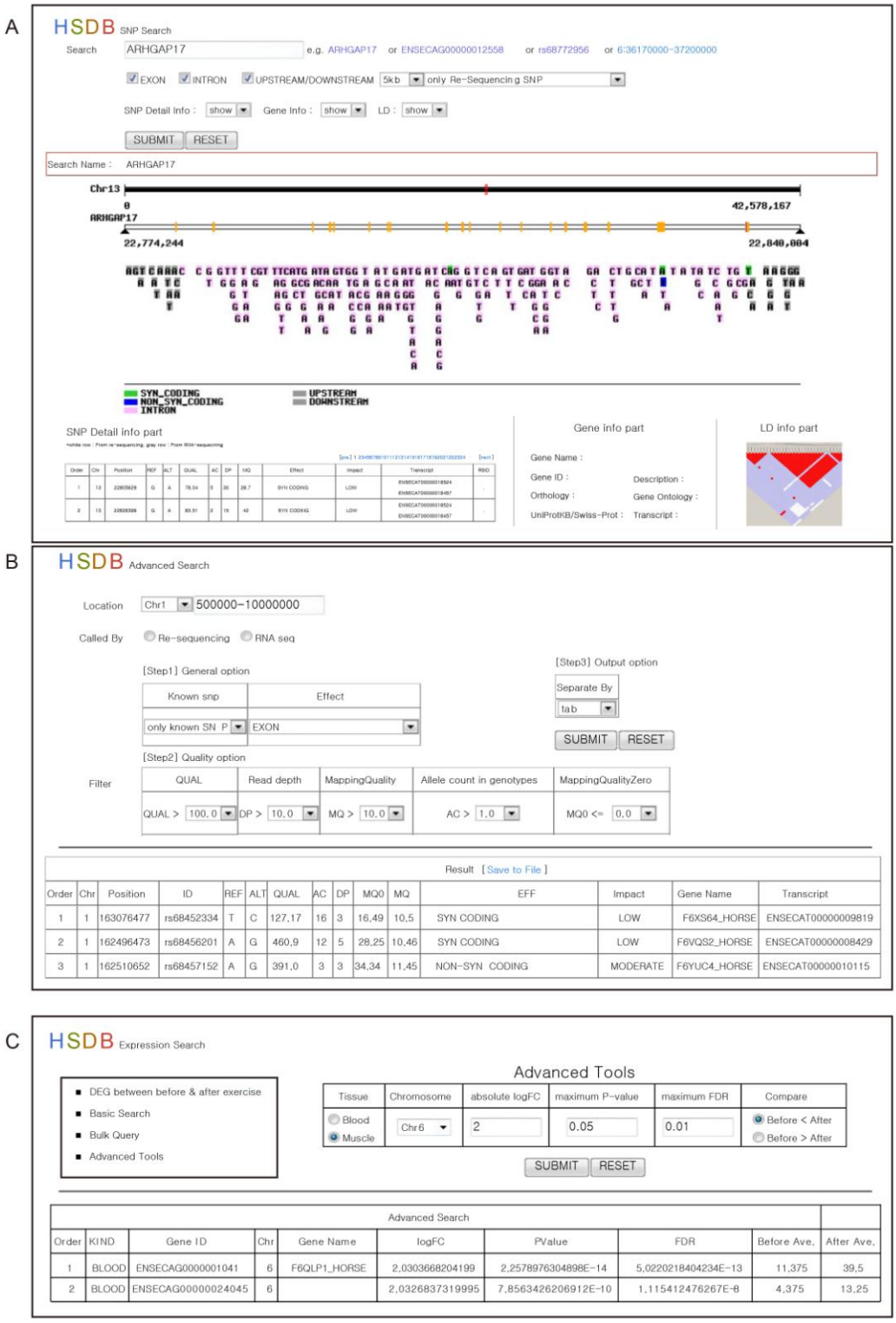

Figure 1. Horse SNP database web page. Horse single nucleotide polymorphism and expression database (HSDB) is available at http://snugenome2.snu.ac.kr/HSDB. (A) Users can search SNP or corresponding gene information using SNP position or RS ID or gene symbol. After searching the SNP, user can see the SNPs within the gene, there are several colors depending their effect. User will get more detail information including reference and alternative allele information, SNP position, SNP quality, depth, and effect impact (high, moderate, low, modifier) which definition follows snpEff. HSDB also enables the user to obtain gene annotation information including transcript ID and gene description, gene ontology term and orthology ID with human and mouse. Users can see the linkage disequilibrium map of interest gene. (B) Advanced Search provides similar results of SNP Search, but user can search the SNPs with many filtering options. (C) HSDB provides expression information of muscle and blood from 4 horses both before and after exercise, especially DEG information. User can search the expression information according to their interest using the Advanced Tools of Expression Search. SNP, single nucleotide polymorphism.

Supplementary Table S1. The number of variants

Chrom	Chom length	Genic region length*	DNA-Seq		RNA-seq	
			Number of SNPs	Average interval of SNP	Number of SNPs	Average interval of SNP
1	185,838,109	85,058,767	265,638	320	32,909	2,585
2	120,857,687	54,020,259	181,513	298	25,093	2,153
3	119,479,920	47,558,748	168,436	282	19,786	2,404
4	108,569,075	48,153,912	157,729	305	15,086	3,192
5	99,680,356	49,305,787	156,993	314	21,656	2,277
6	84,719,076	40,835,311	140,519	291	17,360	2,352
7	98,542,428	46,347,586	154,512	300	19,013	2,438
8	94,057,673	37,008,741	133,993	276	18,050	2,050
9	83,561,422	29,609,146	102,049	290	11,771	2,515
10	83,980,604	38,905,681	132,805	293	21,489	1,810
11	61,308,211	42,313,527	121,290	349	22,887	1,849
12	33,091,231	19,338,230	89,790	215	11,944	1,619
13	42,578,167	25,459,957	91,239	279	16,823	1,513
14	93,904,894	36,769,770	117,602	313	15,874	2,316
15	91,571,448	35,044,790	114,827	305	16,148	2,170
16	87,365,405	41,746,323	143,468	291	19,294	2,164
17	80,757,907	24,276,765	80,866	300	8,364	2,903
18	82,527,541	31,749,417	101,681	312	12,322	2,577
19	59,975,221	24,180,234	87,671	276	9,236	2,618
20	64,166,202	27,844,102	127,436	218	18,589	1,498
21	57,723,302	20,506,447	75,458	272	9,486	2,162
22	49,946,797	23,385,284	74,281	315	11,390	2,053
23	55,726,280	19,324,601	69,413	278	9,026	2,141
24	46,749,900	21,845,462	74,277	294	10,803	2,022
25	39,536,964	22,453,941	71,426	314	11,890	1,888
26	41,866,177	11,026,255	46,915	235	5,391	2,045
27	39,960,074	12,495,321	52,320	239	5,080	2,460
28	46,177,339	22,368,313	78,472	285	10,698	2,091
29	33,672,925	12,545,609	48,395	259	5,476	2,291
30	30,062,385	12,311,357	51,249	240	5,401	2,279
31	24,984,650	10,541,353	44,371	238	5,390	1,956
Total	2,242,939,370	974,330,996	3,356,634	290	443,725	2,196

SNP, single nucleotide polymorphism.
* Genic region length (upstream, downstream 5k).

proportion of intronic reads in many RNA-seq datasets (Kapranov et al., 2010; Van Bakel et al., 2010; Wetterbom et al., 2010; Ameur et al., 2011; St Laurent et al., 2012). We identified regulatory related SNPs such as those in a splice site, upstream and downstream ~20kb region as well as genic SNPs.

The DEGs information from another Thoroughbred study (Kim et al., 2013) which generated 24 RNA-seq datasets from muscle and blood tissue from six horses, taken before and after exercise, identified 1,822 up-regulated genes (URGs) and 930 down-regulated genes (DRGs) in muscle tissue, 222 URGs and 200 DRGs in

Supplementary Table S2. Minor allele frequency of data

	DNA-Seq	RNA-seq
MAF	0.055556	0.125

MAF, minor allele frequency.

Table 1. Identified variants from re-sequencing and RNA-seq

Region	Seq		RNA-seq	
	Count	Percent	Count	Percent
Downstream (5kb)	532,802	12.09	110,047	17.66
Exon	94,502	2.15	40,856	6.56
Intron	3,213,091	72.92	385,484	61.88
Splice site acceptor	362	0.01	4,912	0.79
Splice site donor	485	0.01	10,360	1.66
Upstream (5kb)	555,335	12.60	65,621	10.53
UTR 3’	6,198	0.14	3,641	0.58
UTR 5’	3,552	0.08	2,069	0.33

UTR, untranslated region.

blood tissue after exercise.

Data concordance and validation

We checked the level of concordance and validated the SNPs by comparing the results. We used commercial horse chips to confirm allele type and used RNA-seq data to reduce the false positive rate. The SNPs were validated on 11 horse SNP Chip data by showing that 16,795 (99.96%) of the 16,802 identified loci had same alleles. We also calculated concordance as the fraction of identical genotypes between the SNP chip and re-sequencing data (Figure 2). Eleven samples had 99.02% genotype concordance with ranges from 98.3% to 99.57%. Four out of 18 re-sequenced samples were used as RNA-seq samples, so we calculated the loci and genotype concordance of the two datasets per-individual. Each sample had a similar number of SNPs, at about 2,128K. About 214K out of the total loci (10%) was shared between samples when the loci of SNPs from RNA-seq and re-sequencing were compared. Average genotype concordance rate, which is the proportion of concordant calls with the same allele between the two datasets, is 96.6% out of over ~217K loci. The high genotype concordance rate and percentage of known SNPs indicate that the concordant SNPs were of high quality and accuracy.

DISCUSSION

In this article we described HSDB, an integrated database with genomic variants from whole genome sequencing and RNA-seq, gene annotation, orthologous genes, states-specific expression data and LD maps of horse data sets. Integrating various data, HSDB is a useful tool for researching genomic and biological mechanisms of the horse. HSDB has successfully searched the 3,418K identified SNPs in horse genes including upstream and downstream regions. Validation was conducted by comparing the variants of re-sequencing of four individuals, the variants of RNA-seq of the corresponding four individuals and SNP chip of four individuals. The identified variants were confirmed with high concordance between individuals. The variants from NGS were confirmed with SNPs from SNP chip data of eleven individuals and on average 99.02% genotypes were the same. The 96.6% among 214K shared loci in each sample between RNA-seq and re-sequencing had consistent same genotypes

HSDB provides integrated information to the users in a single database. Users can search by SNP position or gene symbol and get various information: SNPs, expression information, gene structure, function, orthologous information, gene ontology and LD. This allows researchers to use integrated analysis of variants and expression. HSDB is more than a SNP database for the horse genome. Users can research interesting SNPs or genes using this integrated information. Therefore, the user could get information of significant SNPs as their genome-wide association study (GWAS) result, and could get the material for input and fine mapping of a notable region. In addition, it provides the expression data and SNPs that exist ~20kb upstream and downstream that have regulatory components for the corresponding gene. The information can be used to analyze relationships between genomic variant of the regulatory region and the expression level in muscle and blood. Thus, it is a powerful tool for understanding the diversity and

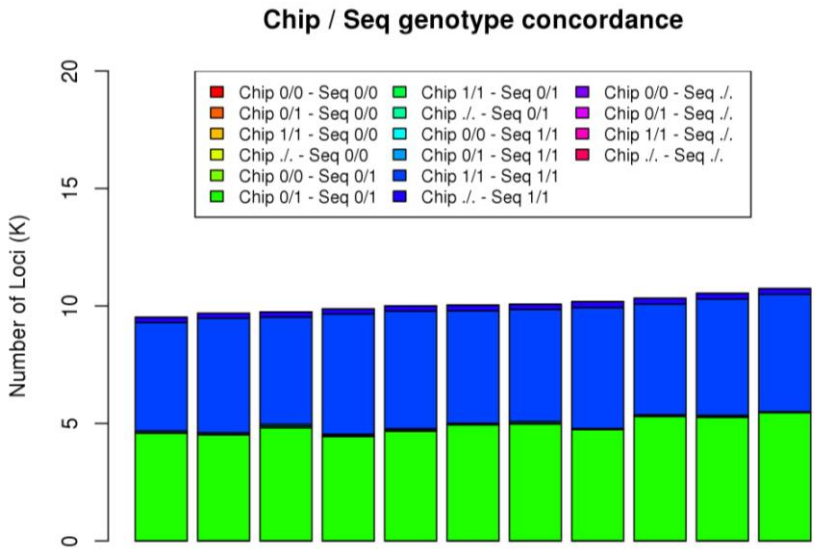

Figure 2. Genotype concordance. Genotype concordance rate which is the proportion of concordant calls having a consistent same genotype between the SNP chip and re-sequencing datasets of 11 samples. Loci of SNP chip were filtered with Hardy-Weinberg equilibrium p-value <0.001, SNP call rate <99% .Type of genotypes are showed as 0/0, 0/1, 1/1 and "./.". 0 and 1 represents reference allele, alternative allele and "." means missing allele. SNP, single nucleotide polymorphism.

expressional aspects of horse genomic variants.

The significance of the study lies in the creation of a well-established horse related database because genetics is an important aspect of breeding and selection of horses. To better understand the genetics underlying horses, information on variants is needed. We believe that the integration of various NGS data on horses will help accelerate horse genome research. In this study, we integrated genes based on SNPs from NGS data of Thoroughbreds, expression from RNA-seq results and annotation information to make a database. Even though HSDB gives a large amount of information, it also showed that a number of unexplored genomic variants still remain to be identified in the horse genome including rare variants by using population genome sequences of eighteen horses and RNA-seq of four horses.

Currently, genome research of animals has been focused on genome-wide information. The database provides information on variant and expression level that is useful for identifying genes associated with economically important traits, studying functional genomics of the horse and researching genetic breeding. The database may contribute to genetic improvements and breeding strategies of horses.

ACKNOWLEDGMENTS

This work was supported by a grant from the Next Generation BioGreen 21 Program (No.PJ008106 and PJ008196), Rural Development Administration, Republic of Korea.

REFERENCES

Ameur, A., A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier, and L. Feuk. 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. 18:1435-1440.

Barrett, J., B. Fry, J. Maller, and M. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Chowdhary, B. P. and T. Raudsepp. 2008. The Horse Genome Derby: racing from map to whole genome sequence. Chromosome Res. 16:109-127.

Cingolani, P., A. Platts, L. Wang, M. Coon, T. Nguyen, S. J. Land, X. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6:80-92.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and Genomes Project Analysis. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Gordon, J. 2001. The Horse Industry – Contributing to the Australian Economy. Rural Industries Research and Development Corporation, Canberra, Australia. 1-58.

Hill, E. W., B. A. McGivney, J. Gu, R. Whiston, and D. E. MacHugh. 2010. A genome-wide SNP-association study confirms a sequence variant (g. 66493737C> T) in the equine myostatin (*MSTN*) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. BMC Genomics 11:552.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. 2002. The Ensembl genome database project. Nucl. Acids Res. 30:38-41.

Kapranov, P., G. St Laurent, T. Raz, F. Ozsolak, C. P. Reynolds, P. H. B. Sorensen, G. Reaman, P. Milos, R. J. Arceci, J. F. Thompson, and T. J. Triche. 2010. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. BMC Biol. 8:149.

Kim, H., T. Lee, W. Park, J. W. Lee, J. Kim, B. Y. Lee, H. Ahn, S. Moon, S. Cho, K. T. Do, H. S. Kim, H. K. Lee, C. K. Lee, H. S. Kong, Y. M. Yang, J. Park, H. M. Kim, B. C. Kim, S. Hwang, J. Bhak, D. Burt, K. D. Park, B. W. Cho, and H. Kim. 2013. Peeling back the evolutionary layers of molecular mechanisms responsive to exercise-stress in the skeletal muscle of the racing horse. DNA Res. 20:287-298.

Langmead, B. and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357-359.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-1303.

Park, K. D., J. Park, J. Ko, B. C. Kim, H. S. Kim, K. Ahn, K. T. Do, H. Choi, H. M. Kim, S. Song, S. Lee, S. Jho, H. S. Kong, Y. M. Yang, B. H. Jhun, C. Kim, T. H. Kim, S. Hwang, J. Bhak, H. K. Lee, and B. W. Cho. 2012. Whole transcriptome analyses of six Thoroughbred horses before and after exercise using RNA-Seq. BMC Genomics 13:473.

Petersen, J. L., J. R. Mickelson, A. K. Rendahl, S. J. Valberg, L. S. Andersson, J. Axelsson, E. Bailey, D. Bannasch, M. M. Binns, A. S. Borges, P. Brama, A. da Camara Machado, S. Capomaccio, K. Cappelli, E. G. Cothran, O. Distl, L. Fox-Clipsham, K. T. Graves, G. Guerin, B. Haase, T. Hasegawa, K. Hemmann, E. W. Hill, T. Leeb, G. Lindgren, H. Lohi, M. S. Lopes, B. A. McGivney, S. Mikko, N. Orr, M. C. Penedo, R. J. Piercy, M. Raekallio, S. Rieder, K. H. Roed, J. Swinburne, T. Tozaki, M. Vaudin, C. M. Wade, and M. E. McCue. 2013. Genome-wide analysis reveals selection for important traits in domestic horse breeds. PLoS Genet. 9:e1003211.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. edgeR:

a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. Nucl. Acids Res. 29:308-311.

St Laurent, G., D. Shtokalo, M. R. Tackett, Z. Yang, T. Eremina, C. Wahlestedt, S. U. Inchima, B. Seilheimer, T. A. McCaffrey, and P. Kapranov. 2012. Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. BMC Genomics 13:504.

Trapnell, C., L. Pachter, and S. L. Salzberg. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105-1111.

Van Bakel, H., C. Nislow, B. J. Blencowe, and T. R. Hughes. 2010. Most "dark matter" transcripts are associated with known genes. PLoS Biol. 8:e1000371.

Wade, C. M., E. Giulotto, S. Sigurdsson, M. Zoli, S. Gnerre, F. Imsland, T. L. Lear, D. L. Adelson, E. Bailey, R. R. Bellone, H. Blocker, O. Distl, R. C. Edgar, M. Garber, T. Leeb, E. Mauceli, J. N. MacLeod, M. C. 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. Guerin, 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. Roed, O. A. Ryder, S. Searle, L. Skow, J. E. Swinburne, A. C. Syvanen, 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, and K. Lindblad-Toh. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326:865-867.

Wetterbom, A., A. Ameur, L. Feuk, U. Gyllensten, and L. Cavelier. 2010. Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing. Genome Biol. 11:R78.