Open Access

Asian Australas. J. Anim. Sci.
Vol. 27, No. 9 : 1345-1354 September 2014
http://dx.doi.org/10.5713/ajas.2013.13857

www.ajas.info
pISSN 1011-2367 eISSN 1976-5517

Copy Number Deletion Has Little Impact on Gene Expression Levels in Racehorses

Kyung-Do Park[1,a], Hyeongmin Kim[2,a], Jae Yeon Hwang[2], Chang-Kyu Lee[2], Kyoung-Tag Do[1], Heui-Soo Kim[3], Young-Mok Yang[4], Young-jun Kwon[5], Jaemin Kim[5], Hyeon Jeong Kim[6], Ki-Duk Song[1], Jae-Don Oh[1], Heebal Kim[2,6], Byung-Wook Cho[7], Seoae Cho[6], and Hak-Kyo Lee[1,]*

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

ABSTRACT: Copy number variations (CNVs), important genetic factors for study of human diseases, may have as large of an effect on phenotype as do single nucleotide polymorphisms. Indeed, it is widely accepted that CNVs are associated with differential disease susceptibility. However, the relationships between CNVs and gene expression have not been characterized in the horse. In this study, we investigated the effects of copy number deletion in the blood and muscle transcriptomes of Thoroughbred racing horses. We identified a total of 1,246 CNVs of deletion polymorphisms using DNA re-sequencing data from 18 Thoroughbred racing horses. To discover the tendencies between CNV status and gene expression levels, we extracted CNVs of four Thoroughbred racing horses of which RNA sequencing was available. We found that 252 pairs of CNVs and genes were associated in the four horse samples. We did not observe a clear and consistent relationship between the deletion status of CNVs and gene expression levels before and after exercise in blood and muscle. However, we found some pairs of CNVs and associated genes that indicated relationships with gene expression levels: a positive relationship with genes responsible for membrane structure or cytoskeleton and a negative relationship with genes involved in disease. This study will lead to conceptual advances in understanding the relationship between CNVs and global gene expression in the horse. (**Key Words:** Copy Number Variation, Genome-wide Expression, Horse, Thoroughbred)

INTRODUCTION

Three types of genetic variations exist in the genome. The largest type is microscopic structural variation, which can be detected by optical microscopy; these include aneuploidies and variation in chromosome size (Reich et al. 2002). The most common and smallest type is single nucleotide polymorphisms (SNPs) (Sachidanandam et al., 2001), detected at the single nucleotide level. Between microscopic structural variation and SNPs, the mid-size variations are submicroscopic variants, such as deletions, duplications and insertions, which are termed copy number variations (CNVs) (Scherer et al., 2007). Usually, CNVs are defined as DNA segments of >1 kb that show copy number differences in comparison with a reference genome, and they have been proposed to exclude insertion/deletions of <1 kb ('indels') (Scherer et al., 2007; Henrichsen et al., 2009a; Alvarez and Akey, 2012), although recent studies have suggested that an indel/CNV cutoff of 50 bp is preferable (Alkan et al., 2011; Mills et al., 2011). These CNVs, so called unbalanced structural variants (SVs) (Mills et al., 2011) are important genetic factors for studying

* Corresponding Author: Hak-Kyo Lee. Tel: +82-31-670-5332, E-mail: breedlee@empal.com

[2] Department of Agricultural Biotechnology, Animal Biotechnology Major, and Research Institute of Agriculture and Life Science, Seoul National University, Seoul 151-921, Korea.

[3] Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea.

[4] Department of Pathology, School of Medicine, and Institute of Biomedical Science and Technology, Konkuk University, Seoul 143-701, Korea.

[5] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea.

[6] CHO & KIM genomics, Seoul 151-919, Korea.

[7] Department of Animal Science, College of Life Sciences, Pusan National University, Miryang 627-702, Korea.

[a] The first two authors should be regarded as joint first authors.

Submitted Dec. 30, 2013; Revised Mar. 5, 2014; Accepted May 12, 2014

human diseases and may have as large of an effect on phenotype as SNPs (Feuk et al., 2006; Sharp et al., 2006). It has been widely accepted that CNVs are involved in differential disease susceptibility (Hollox et al., 2003; Aldred et al., 2005; Gonzalez et al., 2005; Aitman et al., 2006).

At the genome-wide scale, the functional impact of CNVs has been studied in mouse and rat (Guryev et al., 2008; Henrichsen et al., 2009b; Orozco et al., 2009). In mouse, one study reported that the expression of genes within CNVs show moderate correlations with copy number changes (Henrichsen et al., 2009b), and another study reported that 83% of genes within CNVs were differentially expressed (Orozco et al., 2009). In rat, CNVs show functional relationships with 22 expression quantitative loci (Guryev et al., 2008).

A recent study reported 2,368 CNVs in horse (Doan et al., 2012). However, the relationships between CNVs and gene expression levels in horse have not yet been examined. In the current study, we investigated the effects of copy number deletion in the blood and muscle transcriptomes of Thoroughbred racing horses.

MATERIALS AND METHODS

RNA-seq data before and after exercise

We used RNA-seq data from four horses before and after exercise, as described elsewhere (Park et al., 2012; Kim et al., 2013); the data were submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE37870. Briefly, samples of skeletal muscle and blood were taken from Thoroughbred horses before and after exercise. Then, using the Iluumina HiSeq2000, 20 sets of transcriptome data were generated for muscle and blood from four horses both before and after exercise.

The sequences were mapped to a horse reference genome using TopHat (ver.1.4.1) and annotated using the EquCab2 database (http://hgdownload.cse.ucsc.edu/downloads.html-horse) (Supplementary Table 3).

Mapped reads were assembled using Cufflinks (ver.1.3.0) (Trapnell et al., 2010) to estimate the abundance of genes. Fragments per kilobase of exon per million fragments (FPKM) of each sample were calculated to estimate the expression levels of the genes. The FPKM value set for all genes in the muscle and blood tissue samples was normalized using Quantile normalization (Bolstad).

Analysis of horse DNA re-sequencing data

Whole-blood samples were collected from the same four Thoroughbred racing horses used in RNA-seq analysis. Blood (10 mL) was drawn from the carotid artery and treated with heparin to prevent clotting. Manufacturers' instructions were followed to create a paired library. The 90-bp pair-end reads sequence data were generated using HiSeq 2000 (Illumina, Inc., San Diego, CA, USA). Using Bowtie2 (ver. 2.1.0) (Langmead and Salzberg, 2012) with the default settings, pair-end sequence reads were mapped to the reference horse genome (equCab2). We used the following open-source software packages: Picard Tools (ver. 1.75), SAMtools (ver. 0.1.18) (Li et al., 2009), and the genome analysis toolkit (GATK) (version 2.2.4) (McKenna et al., 2010) for downstream processing and variant calling (Supplementary Figure 6).

The DNA re-sequencing data of 14 Thoroughbred racing stallions were downloaded from the NCBI sequence read archive (SRA) database under accession number SRA053569 described elsewhere (Kim et al., 2013). The pair-end sequence reads data were generated by the Korean Racing Authority, and the sampling, sequencing and mapping procedure were same as described above. The sequencing data of all 18 Thoroughbred racing horses are summarized in Supplementary Table 5.

Copy number variation identification for four Thoroughbred horses

We extracted CNVs of the four Thoroughbred racing horses used in RNA sequencing from the aligned re-sequencing data of the combined eighteen horses. The CNV extraction tool, Genome STRUCTURE In Populations (GenomeSTRiP) (ver. 1.04) (Handsaker et al., 2011) was used to acquire deletion calls of CNVs using the program default options. Each variant was genotyped, and the genotype quality was assessed based on the measurement of genotype likelihoods. To consider highly plausible variants, CNVs that passed the genotype quality threshold for all eighteen samples were selected as CNVs of four Thoroughbred horses involved in RNA sequencing. We then excluded two-allele deletion CNVs of which associated gene had certain expression level.

Copy number variations and gene expression analysis

To analyze the relationship between CNVs and gene expression, CNVs and genes were linked according to their start-end position information from GenomeSTRiP and the EquCab2reference annotation database. The linear regression slope between the gene expressions and the deletion status of CNVs was estimated using the 'R' statistics package (Hornik, 2011). The genes that showed consistent tendency in gene expression–CNV slope under all four conditions (blood and skeletal muscle, before and after exercise) were selected to obtain functional information. The equine Ensembl gene IDs were converted to official gene symbols by cross matching with human Ensembl gene IDs and the official gene symbols. The official gene symbols of human homologs of equine genes were used for KEGG pathway annotation and KEGG pathway enrichment analyses using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis Jr et al., 2003). For the genes that show consistent tendency, the representation of functional groups in blood and skeletal muscle relative to the whole genome was investigated using the Expression Analysis Systematic Explorer (EASE) tool (Hosack et al., 2003) within DAVID, of which the EASE is a modified Fisher's exact test used to measure enrichment of gene ontology terms (Alterovitz and Ramoni, 2010).

Copy number variation validation

Genomic DNA samples from each horse were assessed by polymerase chain reaction (PCR) to validate the CNV region selected by GenomeSTRiP. Twenty-six CNV regions of an appropriate size (up to 2.5 kb) for PCR, were selected randomly. All primer pairs listed in Supplementary Table 4 were designed to cover extracted CNV regions. The PCR amplification was carried out using a 2× PCR master mix (iNtRON BioTechnology, Seongnam, Gyeonggi, Korea) containing 1 pM of each primer set. Amplification was performed as follows: 1 cycle of 95℃ for 5 min; 35 cycles of 95℃ for 30 s, annealing at the temperature (Supplementary Table 4) for 20 s, and 72℃ for 1 min or 1 min 40 s (for CNV lengths <1 kb or >1 kb, respectively), followed by 1 cycle of 72℃ for 10 min. All amplicons were separated on a 1% ethidium bromide stained gel for 20 min.

RESULTS

We identified 2,648 possible CNVs from 10,094 CNV candidates using GenomeSTRiP (Handsaker et al., 2011). We retained 1,246 CNVs of deletion calls after applying the genotype quality threshold for further analysis. These plausible CNVs of deletion polymorphisms were linked to the gene expression profiles of four Thoroughbred horses depending on their start-end position. There were four possible position classifications: 'cover' indicates that the genes cover CNV regions, 'inside' indicates that the genes are located in CNV regions, 'front' indicates that the genes span the front part of the CNV regions, and 'rear' indicates that the genes span the rear part of the CNV regions (Figure 2a). Genes not expressed in all four conditions (*i.e.*, blood and skeletal muscle, before and after exercise) were removed from the expression profiles of the four Thoroughbred horses. Finally, we obtained a total of 252 pairs of CNVs and associated genes, including 229 pairs in the 'cover' group, 14 in the 'inside' group, 5 in the 'front' group, and 4 in the 'rear' group.

For each pair of CNVs and associated genes in the skeletal Box-and-Whisker plot was used to display the relationship between CNV status and its position-linked gene expression levels. Figure 1 and Supplementary Figure 2 to 4 depict the relationship between CNVs and gene expression in the 'cover', 'inside', 'front' and 'rear' groups. The deletion status of CNVs was denoted by 0/0 for no deletion at the population- or individual-based CNV region, 0/1 for one allele deletion at the CNV region, or 1/1 for two allele deletions at the CNV region. From the plots, we observed only a reduced gene expression pattern according to allele deletion in the blood of the 'inside' group. For the other groups, we observed no common tendency of positive or negative correlation patterns between the deletion status of CNVs and gene expression levels in blood and muscle before and after exercise.

The linear regression slope was estimated for each pair of CNVs and associated genes to examine the relationship between gene expression level and CNV deletion status individually. Supplementary Figure 1a displays the *PAK7* gene expressions of the four horses (sampled in blood before exercise) and their linked CNV region deletion status, showing a pattern of decreased expression with an increasing number of allele deletions. In contrast, Supplementary Figure 1b presents the *HLA-DQB1* gene (*DQB* gene in horse) expression levels of the four horses (sampled in blood before exercise) and their linked CNV region deletion status, showing a pattern of increased expression with an increasing number of allele deletions.

For the entire trend of increased or decreased gene expression level according to CNV deletion status, the linear regression lines in all four conditions are shown in Figure 2(c, d, e, and f) and the linear regression results in Supplementary Table 1. In all position-related groups and sampling conditions, the linear regression results did not show a clear consistent positive or negative relationship between CNV deletion status and gene expression levels in blood and muscle before and after exercise.

In Supplementary Table 1, the number of positive and negative slope groups indicted no significant difference between the numbers of the positive and negative slope groups at each condition using a paired sample *t*-test (p = 0.692).

For the genes that showed a consistent tendency in the four sampling conditions, human HGNC (Human genome organisation Gene Nomenclature Committee) symbols were annotated using the Ensembl orthologous information, and using the human HGNC symbols, those involved in KEGG pathways were identified (Supplementary Table 2). Among them, 15 genes showed decreased expression patterns when the allele deletion occurred at CNVs in the 'cover' group and in all four sampling conditions (cover negative), and 12 genes show increased expression patterns when the allele deletion occurred at CNVs in the 'cover' group (cover positive). Only one gene showed decreased expression

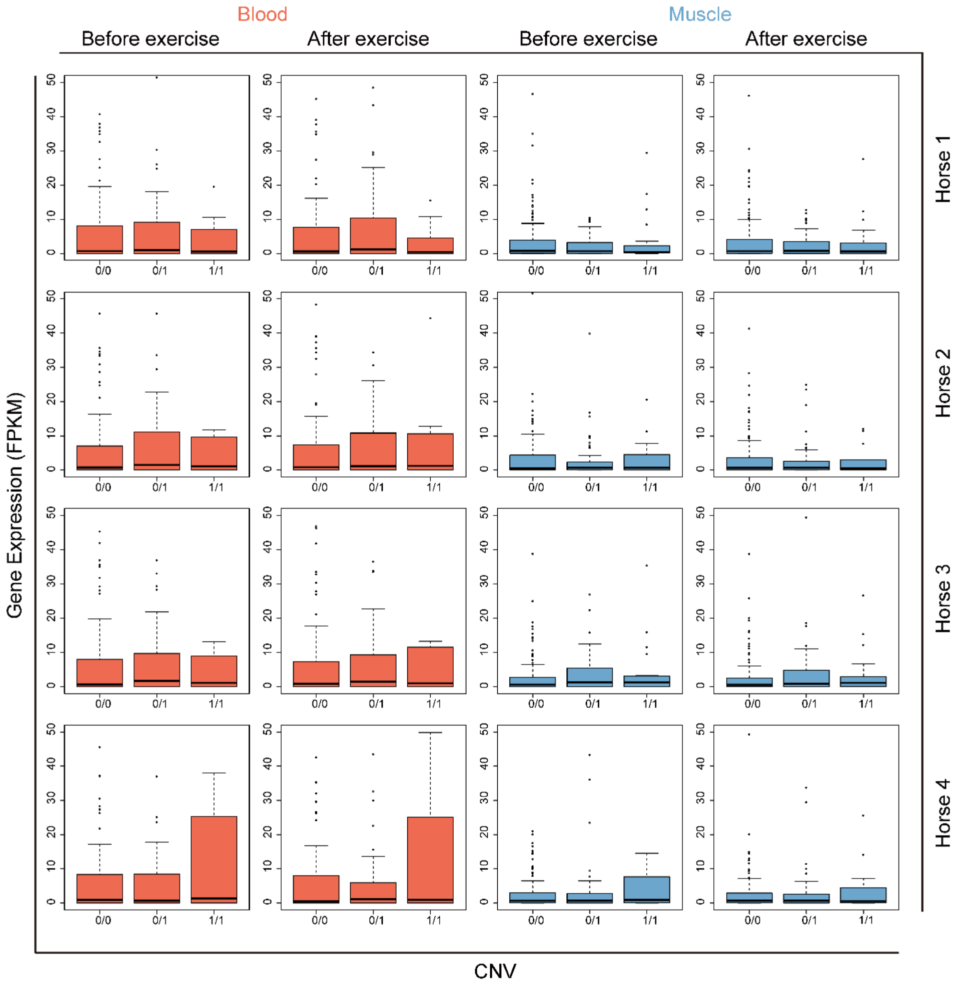

Figure 1. The skeletal Box and Whisker plots show the relationships between copy number variations (CNVs) and 'cover' position linked gene expression in blood and muscle before and after exercise of four Thoroughbred horses. The 'cover' means that the genes cover CNV regions. In each box plot, x-axis shows the deletion status of CNVs and y-axis shows the gene expression level (FPKM). The deletion status of CNVs are denoted by 0/0, 0/1, and 1/1. The 0/0 indicates that the individual or population is homozygous for the wild-type, and the 0/1 indicates that the individual or population is heterozygous for the CNV, and the 0/1 indicates that the individual or population is homozygous for the CNV. The sampling conditions in four horses are shown on the top of the figure, which are blood (red rectangles in the box plot) and muscle (blue rectangles in the box plot) before and after exercise. On the right side of the panel individual horse number was shown.

when the allele deletion occurred at a CNV region in the 'inside' group (inside negative). In 'cover negative', the KEGG pathway term 'cell adhesion molecules'—which includes *CNTN1* and *PTPRM* genes—was enriched.

To confirm that the extracted CNV region is genuine, we performed genomic DNA PCR with CNV region-spanning primer sets (Figure 3a and Supplementary Table 4). Among the extracted CNV regions that were of appropriate size for PCR (<2.5 kb), 26 CNV regions were selected randomly and the CNV deletion type and length in each was examined (Supplementary Figure 5). The amplicons in the deleted allele in the CNV region were of three sizes in comparison to the prediction by GenomeSTRiP: i) similar size of amplicon in the deleted allele in the CNV region (Figure 3b, lane 1), ii) 100 to 200 bp difference in the deleted allele (Figure 3b, lanes 2 and 3), and iii) absence of amplicon (Figure 3b, lane 4). The amplicon size differences might be due to the genuine CNV region length being shorter (second case) or longer (third case) than that predicted by GenomeSTRiP. Then, the distribution of CNV-deleted or -non-deleted alleles in each individual by PCR was compared to the expected model. By combining the criteria of 'deleted amplicon size' and 'distribution of CNV in each individual', we evaluated the extracted CNV region as proper or non-proper according to five classifications (Figure 3c): same pattern and CNV length difference <100 bp (green box) or 100 to 200 bp difference (yellow-green box); different pattern but <100 bp CNV length difference (yellow box); same pattern but the CNV was larger than the amplifying region (orange box,

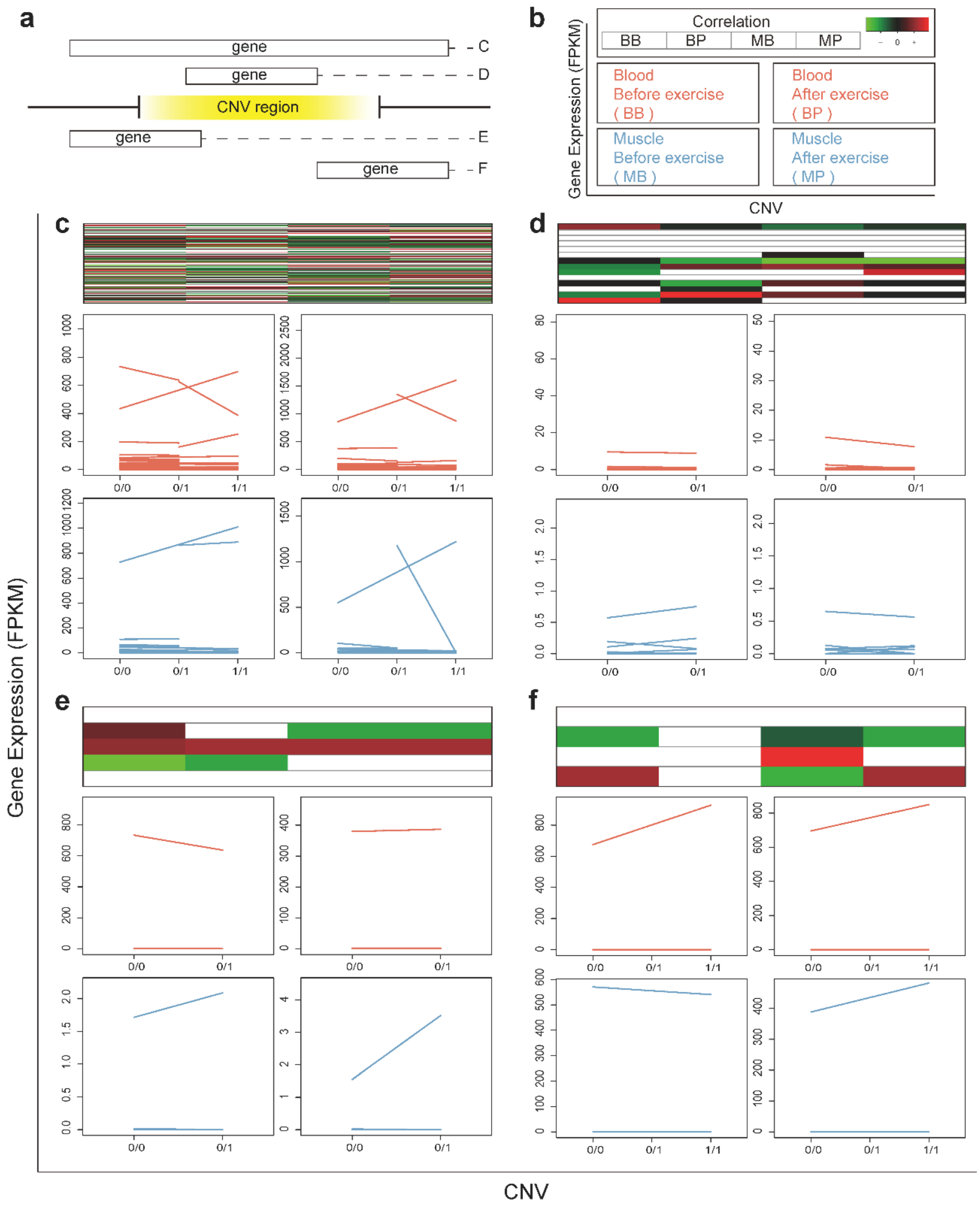

Figure 2. Plot of linear regression and correlation. Plot of linear regression and correlation of gene expression and deletion status of copy number variations (CNVs) in all four conditions (blood and skeletal muscle, before and after exercise) is shown. The position relationships between CNVs and linked genes are shown in (a). The 'cover' indicates that the genes cover CNV regions, 'inside' indicates that the genes are located in CNV regions, 'front' indicates that the genes span the front part of the CNV regions, and 'rear' indicates that the genes span the rear part of the CNV regions. The characters C, D, E, and F on the right side of the position relationships matches (c), (d), (e), and (f). To indicate each sampling condition of (c-f), a simple sampling condition diagram using the same color scheme is shown in (b) (red for blood and blue for muscle).

CNV8), or both the CNV distribution and amplicon size did not match those expected (red box). We considered that the PCR result showing the same CNV distribution pattern and <200 bp difference in the amplicon of the deleted allele (green and yellow-green boxes) was the proper CNV region. According to this standard, the validation process indicated that the extracting CNV region had 76.92% accuracy (Figure 3c, 20 proper CNV regions of 26 CNV regions examined).

DISCUSSION

In a recent study, 2,368 CNVs were identified using array-based methods (Doan et al., 2012). In the current study, 1,246 CNVs of deletion calls were identified using DNA re-sequencing data from 18 Thoroughbred racing horses. Because we used a different approach to identify CNVs based on re-sequencing data, there could be a difference by platform. Also, the caveat of this approach

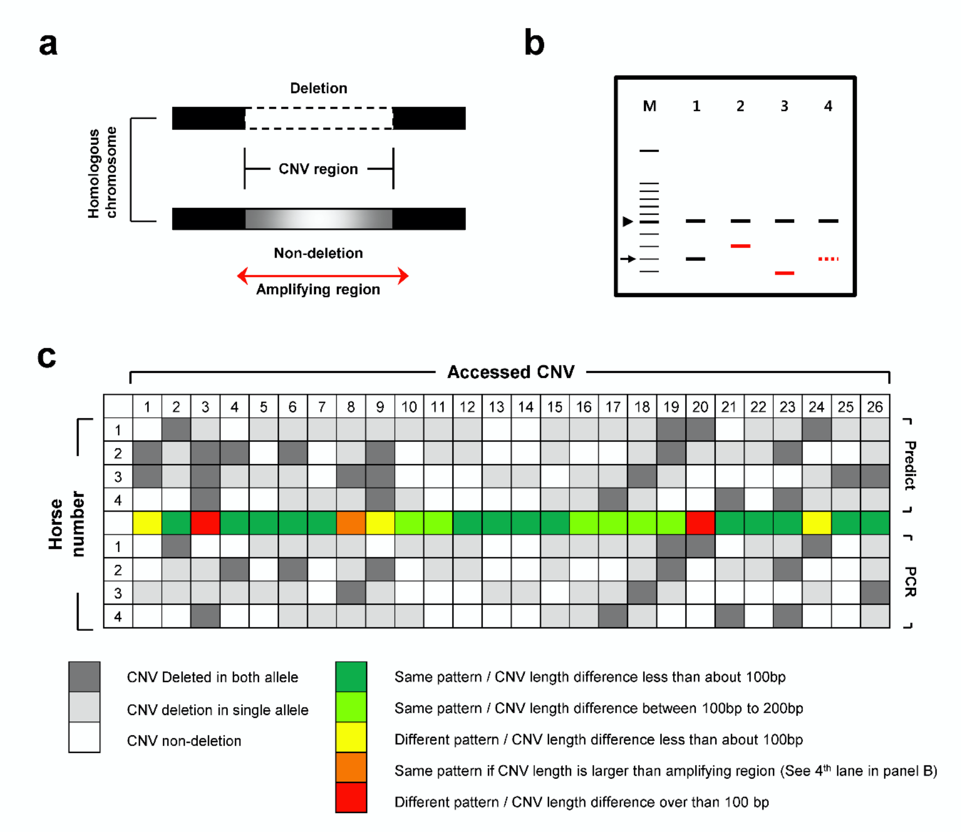
a
Deletion
Homologous chromosome
CNV region
Non-deletion
Amplifying region
b
M
1
2
3
4
c
Accessed CNV
Horse number
Predict
PCR
CNV Deleted in both allele
CNV deletion in single allele
CNV non-deletion
Same pattern / CNV length difference less than about 100bp
Same pattern / CNV length difference between 100bp to 200bp
Different pattern / CNV length difference less than about 100bp
Same pattern if CNV length is larger than amplifying region (See 4th lane in panel B)
Different pattern / CNV length difference over than 100 bp

Figure 3. Copy number variation (CNV) region validation by genomic polymerase chain reaction (PCR) analysis. (a) Genomic DNA amplification model. To validate the reliability of predicted CNV regions using the CNV extraction tool, we performed PCR amplification using the model represented. The primer pairs were designed to contain the CNV region (empty box in CNV del-model and graded box in CNV in-model), and the amplicon is 150 to 350 bp longer than the CNV region. The CNV region deletion type in each individual was confirmed by comparison of the amplicon size. Diagram is not to scale. (b) Amplicon patterns of CNV-deleted alleles by PCR analysis. The PCR amplicons of individuals with CNV deleted and non-deleted alleles are shown (solid triangle for expected non-deleted and arrow for expected deleted). Compared to the predicted size, the following four categories of CNV-deleted allele amplicon size were identified (lane 1-4, red line): similar to predicted size (lane 1), large (lane 2), or small (lane 3) amplicon size, and absence of amplicon from the CNV deleted allele (lane 4, dashed red line). M indicates size marker. (c) CNV patterns of each individual summarized as a heat map. Randomly selected CNV regions obtained using the extraction tool were assessed by PCR analysis. Dark gray, gray, and white boxes indicate two, one, and non-CNV deletions, respectively. By comparing the CNV allele distribution in each individual and amplicon size in CNV-deleted alleles between the predicted and experimental results, we evaluated the suitability of each extracted CNV region by classifying the PCR results into five categories, as described in the Results section (green box to red box).

with the GenomeSTRiP algorithm is that it detects only CNVs of deletions, excluding detection of CNV duplication calls.

We might expect that transcription start sites are located near the start position of genes. In cases of 'cover' (the genes cover CNV regions) and 'front' (the genes span the front part of the CNV regions), these CNVs will not affect the transcription start site of the position-linked genes, but will affect the gene contents. Thus, cover and front CNVs are expected to affect only a small portion of position-linked gene expression variation. However, in the cases of 'inside' (the genes are located in CNV regions) and 'rear' (the genes span the rear part of the CNV regions), these CNVs will affect the transcription start site of the position-linked gene. Thus, inside and rear CNVs are expected to affect gene expression levels significantly. In other words, in the former two cases the expression of genes linked with CNVs may show moderate correlation with copy number changes, whereas in the latter two cases the expression of genes linked with CNVs may exhibit strong correlation with copy number changes. However, in this study, the correlation between the gene expression levels and copy number changes was not marked. Gene expression patterns before and after exercise were analyzed by box plot (Figure 1 and Supplementary Figures 2 to 4) and showed no consistent trend based on copy number changes.

Moreover, a set of individual linear regression lines calculated for single pairs of CNVs and associated genes did not reveal a consistent trend, and the gene expression slope did not show a significant difference in terms of an increasing or decreasing pattern according to copy number changes (Figure 2 and Supplementary Table 1). These results indicate that the CNVs of deletion calls have little relationship with genome-wide expression in the racing horse. This little relationship of the CNVs to gene expression does not agree with a previous study in mouse (Henrichsen et al., 2009b), which reported a moderate correlation between gene copy number and expression. However, our study did not show a significant correlation between gene copy number and expression in the racing horse. This may be due to the CNV identification method, as the copy number was measured in a relative way using a population-based method instead of the individual-based method used in the mouse study. However, in our study, GenomeSTRiP uses both population-level concepts (*e.g.*, average read depth) and the technical features of sequence data of each individual sample such as breakpoint-spanning reads that reflect structural variation (Handsaker et al., 2011; Mills et al., 2011). This unique approach may have contributed to the difference. It is also possible that the difference is due to the gene expression estimate based on RNA sequencing being more sensitive than that of an expression array. Further, the results may differ because the gene copy number and expression data set was only half that of mouse because we only have CNVs of deletion calls, and because our CNV status was clearly presented by allele deletion information. Moreover, we used expression data sets from four horses, whereas the mouse study was based on expression data in only one strain with extreme homozygosity, which may have made identification of common tendency in horse difficult. For example, there is a correlation between gene copy number status and expression if only CNVs of the 'cover' group in the skeletal muscle tissue of horse 1 are considered (Figure 1), or CNVs of the 'inside' group in blood are considered (Supplementary Figure 2). Also, the simple relationship between CNVs and gene expression could be due to strong tissue-specific transcriptional regulation (Subramanian et al., 2005) and strong individual-specific transcriptional regulation due to differences in individual physical activity and nutritional status (Cobb et al., 2005) or by epigenetic regulation by differential genomic imprinting (Jaenisch and Bird, 2003).

From the linear regression results in Supplementary Table 1, it appears that gene expression levels increased or decreased according to allele deletion at CNV regions by chance. One third of CNV linked genes showed increased expression, and one-third showed increased expression, according to copy number. Such observations are in line with findings that genes can show both increased (Somerville et al., 2005; McCarroll and Altshuler, 2007) or decreased (Lee et al., 2006; Guryev et al., 2008) expression with increasing copy number. These patterns of partial or no correlation observed overall suggest either dosage compensation mechanisms or the incomplete inclusion of regulatory elements in the deletion events (Henrichsen et al., 2009a).

The genes showing consistent tendency in the four sampling conditions (blood, skeletal muscle, before and after exercise) were identified and are listed in Supplementary Table 2. In the 'cover negative' group, in which the genes cover the CNV region and show decreased gene expression patterns according to allele deletions at the CNV region, it appears that those genes are involved mainly in membrane structure or cytoskeleton. In the DAVID enrichment test, the KEGG pathway term 'cell adhesion molecules' was enriched, which includes *CNTN1* and *PTPRM* genes. *DIAPH3* and *PAK7* genes (Supplementary Figure 1a) are involved in the 'regulation of actin cytoskeleton' term of the KEGG pathway. The *ANK1* gene encodes Ankyrins, which are a family of proteins that link membrane proteins to the cytoskeleton (Chorzalska et al., 2010), and the *APBB1IP* gene functions in signal transduction from Ras activation to cytoskeleton remodeling (Inagaki et al., 2003). These cellular-structure-related genes were not likely under tight transcriptional regulation, which could be due to the buffering effect of those genes on dosage alteration.

In the 'cover positive' group, in which the genes cover the CNV region and show increased expression patterns according to allele deletion at the CNV region, there were two disease related KEGG pathway terms including two genes, *PARK2* for Parkinson's disease and *HLA-DQB1* (*DQB* gene in horse) (Supplementary Figure 1b) for immune disease. *PARK2* encodes the Parkin protein which plays a role in the cellular machinery that degrades unneeded proteins by ubiquitin tagging (Dawson and Dawson, 2010). *PARK2* haploinsufficiency is a risk factor for Parkinson's disease, and deletions within *PARK2* are associated with its development (Pankratz et al., 2011). *HLA-DQB1*, a gene in the human leukocyte antigen (HLA) region, has an important association with autoimmune disease, and a *HLA-DQB1* polymorphism was reported to be associated with susceptibility to systemic lupus erythematosus (Castaño-Rodríguez et al., 2008). A similar pattern was observed in *DEFA3*, the human α-defensin gene that plays an important role in the innate immune system; *DEFA3* expression was not correlated with genomic copy number (Aldred et al., 2005). Higher expression levels may compensate for lower specific activity. Also, dosage mutation by CNVs could trigger the overexpression of specific genes. However, overexpression according to deletions in CNV regions has not been investigated extensively.

In our CNV validation, we found 77% concordance between the GenomeSTRiP and PCR results. Previous validation reports of GenomeSTRiP prediction showed 83% concordance (5,833 of 7,015 CNVs validated by PCR, array data or breakpoint assembly) (Handsaker et al., 2011) and 87% concordance (14 of 16 CNVs validated by quantitative PCR) (Xi et al., 2011). In comparison, the accuracy of CNV prediction by GenomeSTRiP was moderate.

Due to our limited sample size, the CNV set may include a minor fraction of false positive error. Thus, a large sample size is required, which was not possible in this study. Besides, a more complete set of CNVs—including duplication or insertion call sets—is needed for a more accurate assessment of the impact of CNVs on gene expression. However, we believe that our analyses represent a foundation for further studies of CNVs that affect gene expression.

In summary, we did not observe a clear and consistent relationship between the deletion status of CNVs and gene expression levels before and after exercise in blood and muscle. However, we found some pairs of CNVs and associated genes that indicated relationships with gene expression levels: a positive relationship with genes responsible for membrane structure or cytoskeleton and a negative relationship with genes involved in disease. These observations indicate that copy number deletion has little impact on gene expression levels in racehorses. Our study provides new information regarding the relationship between CNVs and global gene expression in horse. It also motivates further work, such as unraveling the molecular basis of the tight gene expression regulation in spite of allele deletion at CNV regions and the evolutionary mechanism involving ancestral and recent CNVs.

ACKNOWLEDGMENTS

The DNA re-sequencing data of four Thoroughbred horses from this study have been submitted to the NCBI SRA database under accession number SRP017702. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2013R1A1A2012586).

ETHICS STATEMENT

This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of Pusan National University. All experimental procedures used in this study were approved by the Institutional Animal Care and Use Committee of the Pusan National University (PNU-2013-0417). The owners of the Thoroughbred horses gave permission for their animals to be used in this study.

AUTHOR CONTRIBUTIONS

K.-D. Park, C.-K. Lee, K.-T. Do, H.-S. Kim, Y.-M. Yang, K.-D. Song, J.-D. Oh, H.-K. Lee, and S. Cho helped to draft the manuscript. H. Kim participated in the design of the study, carried out the structural variation analysis, RNA-seq expression analysis and drafted the manuscript. J. Y. Hwang carried out the CNV validation. Y. Kwon participated in re-sequencing data analysis. J. Kim and H. J. Kim participated in drafting the manuscript. H. Kim participated in the design of the study and helped to draft the manuscript. B.-W. Cho generated raw data (RNA-seq and DNA re-sequencing of 4 Thoroughbred horses). All authors read and approved the final manuscript.

CONFLICT OF INTEREST

We certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

REFERENCES

Aitman, T. J., R. Dong, T. J. Vyse, P. J. Norsworthy, M. D. Johnson, J. Smith, J. Mangion, C. Roberton-Lowe, A. J. Marshall, E. Petretto et al. 2006. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439:851-855.

Aldred, P. M. R., E. J. Hollox, and J. A. L. Armour. 2005. Copy number polymorphism and expression level variation of the human α-defensin genes *DEFA1* and *DEFA3*. Hum. Mol. Genet. 14:2045-2052.

Alkan, C., B. P. Coe, and E. E. Eichler. 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12:363-376.

Alterovitz, G. and M. F. Ramoni. 2010. Knowledge Based Bioinformatics: From Analysis to Interpretation. John Wiley & Sons, Ltd, Chichester, UK.

Alvarez, C. E. and J. M. Akey. 2012. Copy number variation in the domestic dog. Mamm. Genome 23:144-163.

Bolstad, B. preprocessCore: A collection of pre-processing functions. R package version 1.

Castaño-Rodrı́guez, N., L. M. Diaz-Gallo, R. Pineda-Tamayo, A. Rojas-Villarraga, and J. M. Anaya. 2008. Meta-analysis of *HLA-DRB1* and *HLA-DQB1* polymorphisms in Latin American patients with systemic lupus erythematosus. Autoimmun. Rev. 7:322-330.

Chorzalska, A., A. Łach, T. Borowik, M. Wolny, A. Hryniewicz-Jankowska, A. Kolondra, M. Langner, and A. F. Sikorski. 2010. The effect of the lipid-binding site of the ankyrin-binding domain of erythroid β-spectrin on the properties of natural membranes and skeletal structures. Cell. Mol. Biol. Lett. 15:406-423.

Cobb, J. P., M. N. Mindrinos, C. Miller-Graziano, S. E. Calvano,

H. V. Baker, W. Xiao, K. Laudanski, B. H. Brownstein, C. M. Elson, D. L. Hayden et al. 2005. Application of genome-wide expression analysis to human health and disease. Proc. Natl. Acad. Sci. USA. 102:4801-4806.

Dawson, T. M. and V. L. Dawson. 2010. The role of parkin in familial and sporadic Parkinson's disease. Mov. Disord. 25:S32-S39.

Dennis Jr, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. 2003. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 4:R60.

Doan, R., N. Cohen, J. Harrington, K. Veazy, R. Juras, G. Cothran, M. E. McCue, L. Skow, and S. V. Dindot. 2012. Identification of copy number variants in horses. Genome Res. 22:899-907.

Feuk, L., A. R. Carson, and S. W. Scherer. 2006. Structural variation in the human genome. Nat. Rev. Genet. 7:85-97.

Gonzalez, E., H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P. Quinones, M. J. Bamshad et al. 2005. The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307(5714):1434-1440.

Guryev, V., K. Saar, T. Adamovic, M. Verheul, S. A. A. C. van Heesch, S. Cook, M. Pravenec, T. Aitman, H. Jacob, and J. D. Shull, N. Hubner, and E. Cuppen. 2008. Distribution and functional impact of DNA copy number variation in the rat. Nat. Genet. 40:538-545.

Handsaker, R. E., J. M. Korn, J. Nemesh, and S. A. McCarroll. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet. 43:269-276.

Henrichsen, C. N., E. Chaignat, and A. Reymond. 2009a. Copy number variants, diseases and gene expression. Hum. Mol. Genet. 18:R1-R8.

Henrichsen, C. N., N. Vinckenbosch, S. Zöllner, E. Chaignat, S. Pradervand, F. Schütz, M. Ruedi, H. Kaessmann, and A. Reymond. 2009b. Segmental copy number variation shapes tissue transcriptomes. Nat. Genet. 41:424-429.

Hollox, E. J., J. A. L. Armour, and J. C. K. Barber. 2003. Extensive normal copy number variation of a β-defensin antimicrobial-gene cluster. Am. J. Hum. Genet. 73:591-600.

Hornik, K. The R FAQ [Internet]. ISBN 3-900051-08-9. Available from: http://CRAN.R-project.org/doc/FAQ/R-FAQ.html

Hosack, D. A., G. Dennis Jr, B. T. Sherman, H. C. Lane, and R. A. Lempicki. 2003. Identifying biological themes within lists of genes with EASE. Genome Biol. 4:R70.

Inagaki, T., S. Suzuki, T. Miyamoto, T. Takeda, K. Yamashita, A. Komatsu, K. Yamauchi, and K. Hashizume. 2003. The retinoic acid-responsive proline-rich protein is identified in promyeloleukemic HL-60 cells. J. Biol. Chem. 278:51685-51692.

Jaenisch, R. and A. Bird. 2003. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nat. Genet. 33:245-254.

Kim, H., T. Lee, W. Park, J. W. Lee, J. Kim, B.-Y. Lee, H. Ahn, S. Moon, S. Cho, K.-T. Do et al. 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.

Lee, J. A., R. E. Madrid, K. Sperle, C. M. Ritterson, G. M. Hobson, J. Garbern, J. R. Lupski, and K. Inoue. 2006. Spastic paraplegia type 2 associated with axonal neuropathy and apparent *PLP1* position effect. Ann. Neurol. 59:398-403.

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

McCarroll, S. A. and D. M. Altshuler. 2007. Copy-number variation and association studies of human disease. Nat. Genet. 39:S37-S42.

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.

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

Orozco, L. D., S. J. Cokus, A. Ghazalpour, L. Ingram-Drake, S. Wang, A. Van Nas, N. Che, J. A. Araujo, M. Pellegrini, and A. J. Lusis. 2009. Copy number variation influences gene expression and metabolic traits in mice. Hum. Mol. Genet. 18:4118-4129.

Pankratz, N., A. Dumitriu, K. N. Hetrick, M. Sun, J. C. Latourelle, J. B. Wilk, C. Halter, K. F. Doheny, J. F. Gusella, W. C. Nichols et al. 2011. Copy number variation in familial Parkinson disease. PloS one 6:e20988.

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 et al. 2012. Whole transcriptome analyses of six thoroughbred horses before and after exercise using RNA-Seq. BMC Genomics 13:473.

Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nature Genet. 32:135-142.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933.

Scherer, S. W., C. Lee, E. Birney, D. M. Altshuler, E. E. Eichler, N. P. Carter, M. E. Hurles, and L. Feuk. 2007. Challenges and standards in integrating surveys of structural variation. Nat. Genet. 39:S7-S15.

Sharp, A. J., Z. Cheng, and E. E. Eichler. 2006. Structural variation of the human genome. Annu. Rev. Genomics Hum. Genet. 7:407-442.

Somerville, M. J., C. B. Mervis, E. J. Young, E. J. Seo, M. del Campo, S. Bamforth, E. Peregrine, W. Loo, M. Lilley, and L. A. Pérez-Jurado. 2005. Severe expressive-language delay related to duplication of the Williams–Beuren locus. N. Engl. J. Med. 353:1694-1701.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA. 102:15545-15550.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28:511-515.

Xi, R., A. G. Hadjipanayis, L. J. Luquette, T.-M. Kim, E. Lee, J. Zhang, M. D. Johnson, D. M. Muzny, D. A. Wheeler, R. A. Gibbs et al. 2011. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc. Natl. Acad. Sci. 108:E1128-E1136.