

Research Article

Analysis of the Genetic Variation of Horse Gap Junction Protein Alpha 4 in Thoroughbreds

Jae-Young Choi¹, Donghyun Shin², Jeong-Woong Park³, Kyung Hwan Kim⁴, Jae-Don Oh^{2*}, Byung-Wook Cho^{3*}¹Subtropical Livestock Research Institute, National Institute of Animal Science, RDA, Jeju 63242, Korea²Department of Animal Biotechnology, College of Agricultural and Life Sciences, Chonbuk National University, Jeonju 54896, Republic of Korea³Department of Animal Science and Biotechnology, Kyungpook National University, SangJu, Republic of Korea⁴Department of Animal Science, College of Natural Resources and Life Sciences, Pusan National University, Miryang 627-702, Republic of Korea

*Corresponding Author: Jae-Don Oh, Tel : +82-63-270-5931, Fax : +82-63-270-5937, E-mail : oh5ow@naver.com

Byung-Wook Cho, Tel : +82-55-350-5515, Fax : +82-55-350-5519, E-mail : bwcho@pusan.ac.kr

Abstract

The purpose of this study was to analyze the novel single nucleotide polymorphisms (SNPs) of the gap junction protein alpha 4 gene (GJA4) identified in horse muscle RNA-seq and to predict structural changes of proteins by SNPs. In our previous study, we observed differentially expressed genes (DEGs) in Thoroughbreds before and after exercise through RNA-seq analysis. In addition, we conducted an evolutionary analysis using Thoroughbred and Jeju horse re-sequencing data. As a result, we discovered a novel SNP present in GJA4 (LOC22385534 C>G) in the evolutionarily selected gene in the Thoroughbred horse. Transcription factor (TF) binding sites in the 5'-regulatory region of this gene were identified via PROMO. Additionally, bioinformatics tools were used to predict the effect of non-synonymous SNPs (nsSNP) on function and stability. We identified the change of protein structure owing to the amino acid sequence change, which was proline to arginine according to nsSNP data. Our analysis will be useful as a basis for studying genes and SNPs that affect horses.

Keywords: DEG, GJA4, Protein structure, SNP, Thoroughbred

Introduction

The Thoroughbred breed originated in the early 1700s, by the crossbreeding of Arabian stallions and indigenous mares in England. Modern males originate from one of three stallions (Godolphin Arabian, Byerley Turk, and Darley Arabian) and modern females originate from one of approximately 70 foundation mares (Willett 1970). The aerobic capacity of the Thoroughbred horse is excellent unlike that of other species of similar size because Thoroughbreds have a high aerobic capacity and maximum oxygen uptake (Jones et al. 1989; Young et al. 2002). These specific characteristics have been enhanced by the intense artificial selection of sequence variations that contribute to special racing performance (Gu et al. 2009).

The key achievements of the Horse Genome Project are the complete sequencing of the Thoroughbred horse (EquCab 2.0) and the result of 1,162,753 single nucleotide polymorphisms (SNPs) across different breeds (Wade et al. 2009). The SNP of the myostatin gene (MSTN) (g.66493737C>T) has been associated with muscle hypertrophy in mammals and racing performance in horses (Hill E et al. 2010; Hill EW et al. 2010).

SNPs located in the gene coding region are classified into two types: synonymous SNPs that cause no change to amino acid metabolites and non-synonymous SNPs (nsSNPs) that affects the protein structure and molecular function. However, approximately 90% of nsSNPs cause a change in amino acid metabolites and protein function (Wang and Moulton 2001). SNPs located in the coding region cause mutations in the amino acid sequence. Several studies show that the effects of amino acid allelic variants on protein structure and function can be predicted by the analysis of multiple sequence alignments and protein 3D structures (Chasman and Adams 2001; Ng and Henikoff 2001; Sunyaev et al. 2001).

We estimated the evolutionary relationships of horses using the dielectric data of six types of mammals (horse, human, mouse, dog, pig, and cow) and horses based on the whole transcriptome analyses of Thoroughbred horses before and after exercise (Park et al. 2012; Kim et al. 2013). As a result, evolutionary selected breed-specific genes, including gap junction protein alpha 4 (GJA4) and a novel SNP, were discovered. GJA4 delivers an electrical signal between cells directly through the gap junction composed of connexins (White and Paul 1999).

In this study, we focused on the SNP structure characterization of GJA4 in Thoroughbreds. The transcription factor (TF)-binding sites in the regulatory region of GJA4 were predicted. In addition, we discovered a novel nsSNP in this gene and predicted its influence on protein structure.

Material and Methods

Experimental samples

Blood samples were collected from 87 domestic Thoroughbred racehorses that had run a race at the Seoul Lets Run Park. To extract genomic DNA, 900 mL of red blood cell (RBC) lysis solution was added to 300 mL of blood, processed for 3 min, and centrifuged at 15,000 rpm for 30 s. The supernatant was removed, 300 mL of cell lysis solution and 100 mL of protein precipitation solution were added, and the solution was processed and mixed thoroughly. The DNA solution layer was collected by centrifuging the solution at 15,000 rpm for 5 min, and the supernatant was added to 300 mL of isopropanol and shaken slowly. The resulting solution was centrifuged at 15,000 rpm for 10 min and the supernatant was removed. Ethanol (500 mL) was added to the supernatant, the solution was shaken until it became clear, and centrifuged at 15,000 rpm for 3 min. DNA was extracted by volatilizing and removing ethanol (protocol number: PNU-2017-1553).

Polymerase chain reaction (PCR) analysis

NCBI (<http://www.ncbi.nlm.nih.gov>) and Ensembl Genome Browser (www.ensembl.org) were used to retrieve gene sequence information. The primers used to detect SNPs were synthesized using PRIMER3 software (<http://bioinfo.ut.ee/primer3-0.4.0/>), and the synthesized primers included GJA4 Primer F (3'-CCGGCTTCTAGCACACTCTT-5') and R (3'-TACCTGGGCCACGTCATTTA-5'). To determine the genotype of GJA4 SNPs, PCR was used on the genomic DNA of racehorses using the following conditions: initial denaturation at 94°C for 10 min; 40 cycles of denaturation at 94 °C for 30 s, annealing at 58 °C for 30 s, extension at 72°C for 30 s, and final extension at 72°C for 10 min. PCR products were separated in a 1.5% SeaKem® LE agarose gel (Lonza, Rockland, USA), detected under UV light and subjected to Sanger sequencing for confirmation after cloning. Cloning of PCR products was carried out using a pGEM®-T Easy Cloning Vector System (Promega), and each gene sequence was confirmed by Sanger sequencing. SNPs were examined by comparing the gene sequence obtained after sequencing using those obtained from a BLAST search (National Center for Biotechnology Information, Bethesda, MD, USA).

TF-binding site prediction

TF binding sites were predicted with ALGGEN PROMO v8.3 (<http://alggen.lsi.upc.es>).

Protein structure prediction

To investigate the amino acid sequence changes of GJA4 according to SNP, protein structure prediction analysis was performed and the following tools were used: Sorting Intolerant from Tolerant (SIFT) (<http://sift.jcvi.org/>), Polymorphism Phenotyping v2 (Polyphen-2) (<http://genetics.bwh.harvard.edu/pph2/>), Protein Variation Effect Analyzer (PROVEAN) (<http://provean.jcvi.org/>), and I-Mutant 2.0 (<http://folding.biofold.org/i-mutant/i--mutant2.0.html>). Each tool was used to predict the effect of amino acid sequence substitution on the biological function of the protein and the change in the structure and stability of the protein.

3D structure analysis

Phyre2 (<http://www.sbg.bio.ic.ac.uk/~phyre2>) was used to predict the tertiary structure of proteins. The tridimensional structure was analyzed, and the change of structure elements was validated by referring to the structural classification method of proteins database and data in the protein data bank. Chimera v1.11 was used to confirm the sequence and comparison between protein structures according to SNP sequences.

Results and Discussion

Expression pattern of horse GJA4 and GJA4 associated genes

In our previous studies, we identified GJA4 as a differentially expressed gene in response to exercise and a selection signature gene related to athletic adaptation of Thoroughbred muscle (Park et al. 2012; Kim et al. 2013). From these results, the expression of GJA4 in the muscle was 2.14-fold higher after exercise than before (Figure 1). To explain the expression patterns, 164 TF binding sites were predicted in the 600 bp upstream region of the GJA4 using PROMO (data not shown). When the predicted TF genes were matched to the differentially expressed genes in response to exercise, early growth response 1 (EGR1), interferon regulatory factor 1 (IRF1), JUNB, nuclear factor kappa B (NFkB), and X-box binding protein 1 (XBP1) genes were up-regulated, and the D site of albumin promoter (DBP) gene was down-regulated after exercise (Figure 2 and Table 1).

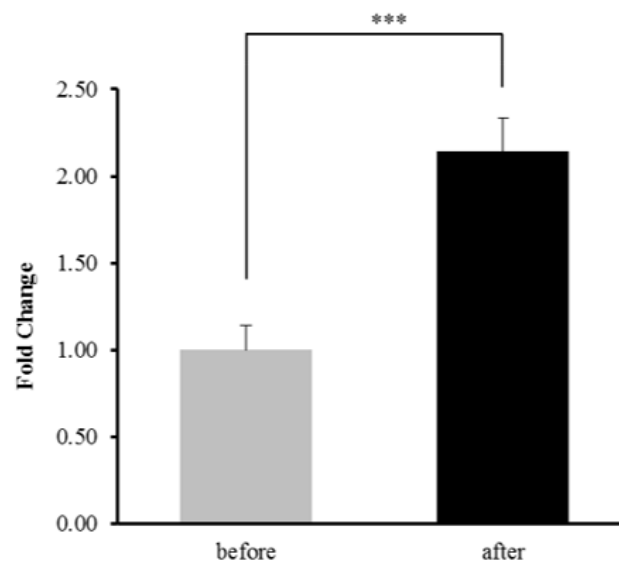


Figure 1. The gene expression of GJA4 by RNA-seq before and after exercise. *** $p < 0.001$.

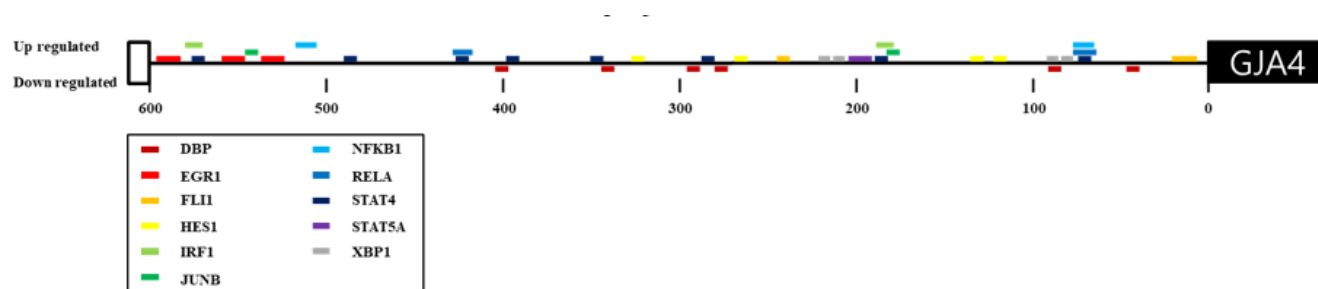


Figure 2. Prediction of TF-binding sites in the 600 bp upstream region of GJA4. Upper side-elements indicate binding sites of the upregulated TFs and downside-elements indicate binding sites of the downregulated TFs after exercise.

Table 1. List of predicted TF-binding site locations and fold change values of DEGs

Transcription factor	Binding sequence	Start position	End position	Average log2 ratio
DBP	CGGATTA	41	47	-1.71
	TAATGAG	82	88	
	TAATCCT	272	278	
	AGCAAGT	257	263	
	AACTGCT	317	323	
	AGCAGGC	372	378	
EGR1	ATCCCCGCCCCC	490	502	5.71
	GCGCCCGCCGCC	511	523	
	CTCCCCGCCCTT	544	556	
IRF1	TCAAGGAAA	170	178	2.47
	TTTCCTGCC	533	541	
JUNB	AGGTCA	167	172	5.11
	TGACCA	505	510	
NFKB1	TGTGATTCCCC	65	75	1.47
	GGGGAGCCTCA	473	483	
XBP1	ATGTCC	76	81	2.10
	ATGAGC	84	89	
	CCGCAT	196	201	
	CCTCAT	204	209	

DEGs were extracted by RNA-seq analysis of Thoroughbred horses before and after exercise

Interferon regulatory factor 1 (IRF1) is involved in most pattern recognition receptor (PRR) signaling events, including the immune activation capacity that links innate and adaptive immunity. PRR is a receptor that recognizes the pathogenic molecule patterns (PAMPs) that cause congenital immune responses (Akira et al. 2006; Tamura et al. 2008). The transcription factor nuclear factor kappa-light-chain-enhancer of activated B cells (NF- κ B) plays an important role in many cellular processes and is involved in anti-inflammation (Pereira and Oakley 2008). NF- κ B is activated by proinflammatory cytokines (Moynagh et al. 1993). Cytokines are intercellular signaling proteins that affect target cells and are involved in immunologically dependent inflammatory reactions. Exercise-induced skeletal muscle damage causes acute inflammatory responses and activates the muscle fiber regeneration processes. Various cytokines regulate in vivo muscle inflammation and muscle fiber

regeneration (Cannon and Pierre 1998; Oppenheim 2001). Expression of zinc finger protein early growth response 1 (EGR1) induces the expression of various environmental signals including growth factors, hormones, and neurotransmitters. EGR1 is involved in the regulation of growth and differentiation (O'Donovan et al. 1999; Thiel et al. 2000), is an important regulator of pathological cardiac growth, and plays a pivotal role in the coordinated transcription of several inflammatory and coagulation factor genes, including those related to the pathogenesis of atherosclerosis and restenosis after vascular injury (Khachigian et al. 1996; Yan et al. 2000; Buitrago et al. 2005).

The GJA4 polymorphism (C1019T) is closely related to human atherosclerosis and myocardial infarction (Wong et al. 2006). The aerobic capacity of Thoroughbred horses is superior, and maximum oxygen uptake is associated with athletic performance in horses (Jones et al. 1989; Harkins et al. 1993; Gauvreau et al. 1995; Young et al. 2002). GJA4 expression is increased at the heart muscle of the Goto-Kakizaki type 2 diabetic rat after exercise (Salem et al. 2013). Therefore, we concluded that GJA4 could be an important regulator of cardiac metabolism and inflammatory responses that were caused by exercise.

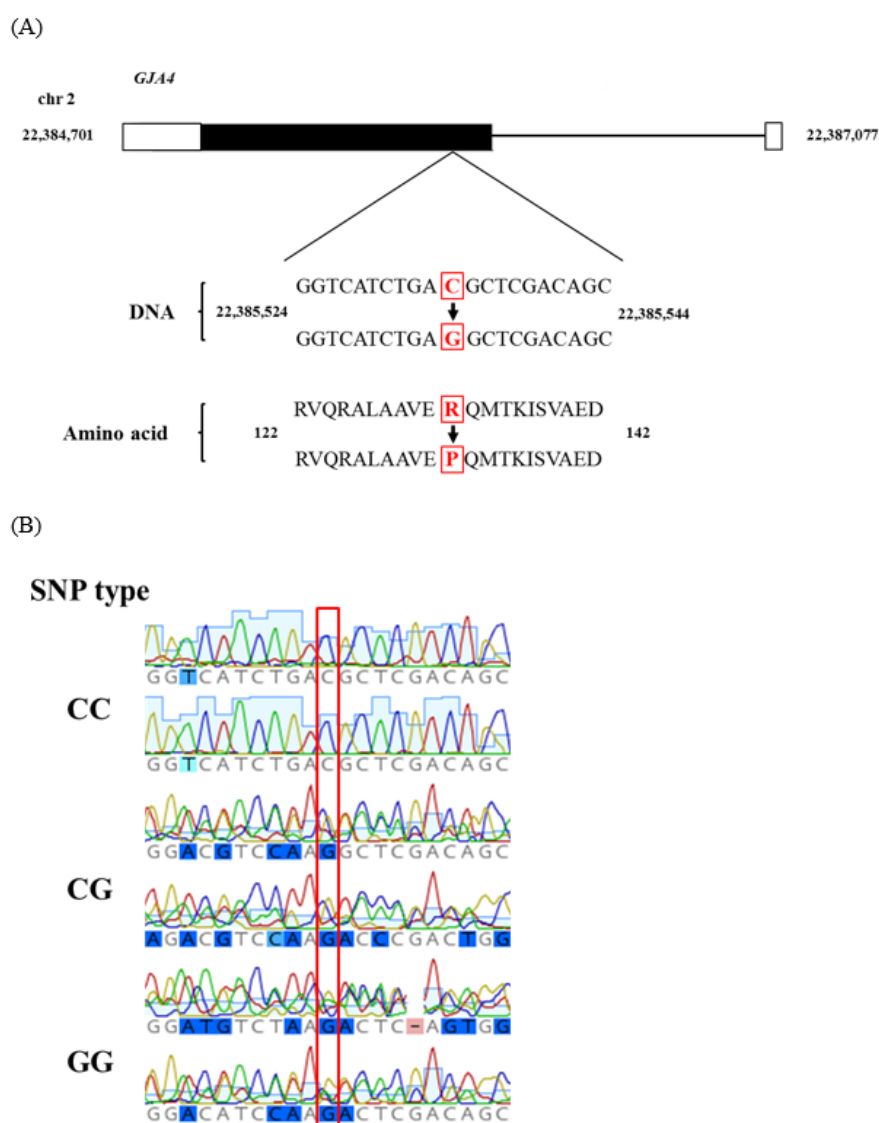


Figure 3. (A) Structure and location of an nsSNP (LOC22385534 C to G) in GJA4. Filled boxes represent GJA4 coding regions. (B) Sequence analysis of the novel nsSNP in GJA4.

Protein prediction analysis between SNPs

In GJA4, the existence of an SNP that changed C to G on the 22385534th sequence of chromosome 2, and three genotype alleles (CC, CG, and GG) were observed. Genotypes were analyzed in 87 Thoroughbred racehorses. The frequency was 82.76% for CC type, 14.94% for CG type, and 2.30% for GG type. As a result of predicting the structure of proteins created according to amino acid replacement owing to SNPs, it was demonstrated that the final product varies according to the allele (Figure 3). In GJA4, amino acid replacement from arginine to proline occurred in amino acid sequence no. 132. To predict whether the mutation of nsSNP influences the function of the protein, protein structure prediction analyses were carried out. Four tools were used to analyze the amino acid sequences substituted by the reference amino acid sequence and the SNP type. SIFT is a tool that predicts whether SNPs affect protein function by using amino acid sequence homology. PolyPhen-2 is a tool that predicts the structure and function of an SNP based on sequence information of amino acid substitutions of nsSNPs. PROVEAN predicts all protein sequence changes such as indel and amino acid sequence substitutions. Finally, I-mutant is a tool that predicts the protein stability free-energy change of nsSNPs. Only Polyphen-2 showed that SNPs affect the protein (Table 2). After 3D-tertiary structure analysis, we observed that a beta-strand changed to an alpha-helix and protein structure changed owing to the influence of SNPs (Figure 4A and B). This indicated that nsSNPs influenced the structure of proteins.

In conclusion, we proposed to correlate the induction of GJA4 and the regulation of inflammatory responses caused by muscle damages and cardiac metabolism in horse exercise. We discovered a novel nsSNP of horse GJA4 that affected protein structure. This study contributes to future research into the exercise ability of horses.

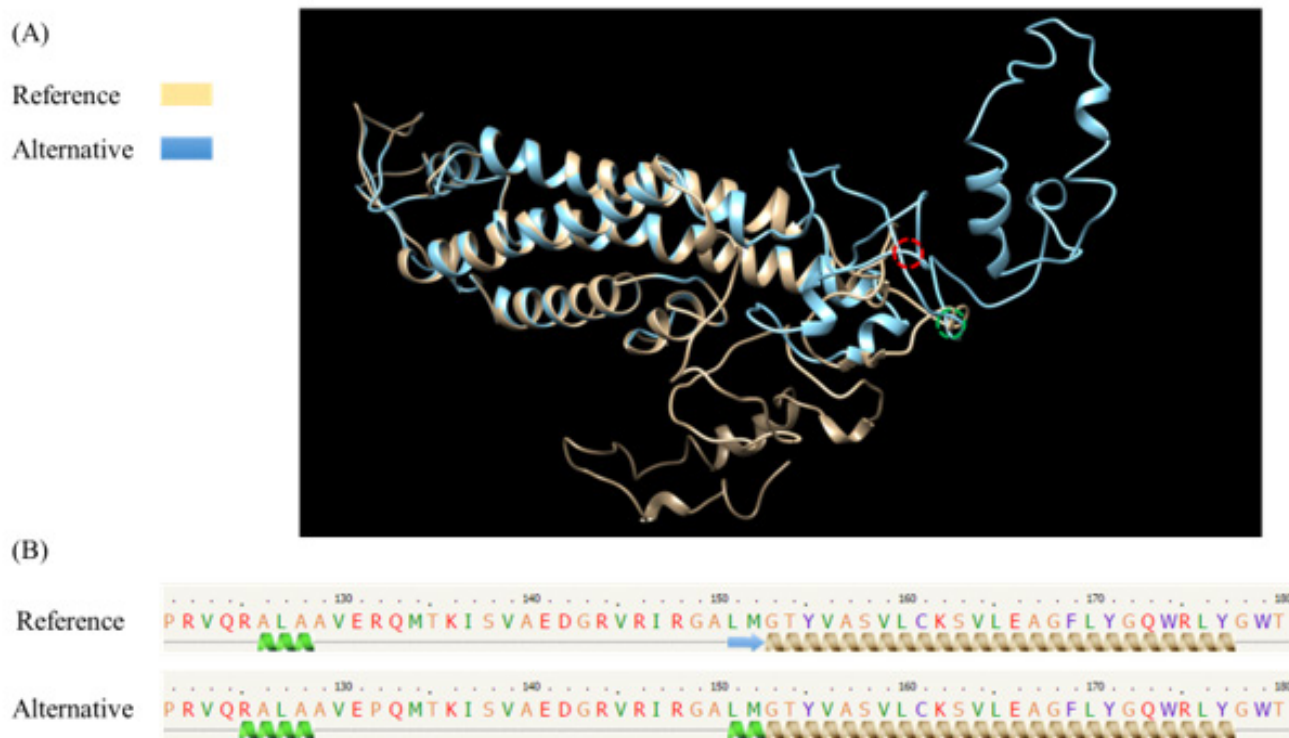


Figure 4. (A) 3D-tertiary structure analysis of SNPs in GJA4. (B) Secondary structure of the GJA4 variation region (circle in Figure A), green circle means reference sequence and red circle means alternative sequence. Secondary structure shows the beta-strand change to an alpha-helix.

Table 2. The prediction results that nsSNP is a mutation that influences protein function

SYMBOL	SIFT		PolyPhen-2			PROVEAN			I-Mutant				
	Prediction	Score	Prediction	Score	Sensitivity	Specificity	Prediction	Score	DDG Value	SVM2	RI	SVM3	RI
GJA4	Tolerated	0.26	Possibly damaging	0.794	0.84	0.93	Neutral	-0.523	-0.41	Decrease	5	Neutral	1

Acknowledgments

This work was supported by a 2-Year Research Grant of Pusan National University

REFERENCES

- Akira, S., Uematsu, S. and Takeuchi, O. 2006. Pathogen recognition and innate immunity. *Cell*. 124(4):783-801.
- Buitrago, M., Lorenz, K., Maass, A.H., Oberdorf-Maass, S., Keller, U., Schmitteckert, E.M., Ivashchenko, Y., Lohse, M.J. and Engelhardt, S. 2005. The transcriptional repressor Nab1 is a specific regulator of pathological cardiac hypertrophy. *Nat. Med.* 11(8):837.
- Cannon, J.G. and St. Pierre, B.A. 1998. Cytokines in exertion-induced skeletal muscle injury. *MOL. CELL. BIOCHEM.* 179(1-2):159-168.
- Chasman, D. and Adams, R.M. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. MOL. BIOL.* 307(2):683-706.
- Gauvreau, G.M., Staempfli, H., McCutcheon, L., Young, S.S. and McDONELL, W.N. 1995. Comparison of aerobic capacity between racing standardbred horses. *J. APPL. PHYSIOL.* 78(4):1447-1451.
- Gu, J., Orr, N., Park, S.D., Katz, L.M., Sulimova, G., MacHugh, D.E. and Hill, E.W., 2009. A genome scan for positive selection in thoroughbred horses. *PLoS One*. 4(6):e5767.
- Harkins, J., Beadle, R. and Kamerling, S. 1993. The correlation of running ability and physiological variables in Thoroughbred racehorses. *EQUINE. VET. J.* 25(1):53-60.
- Hill, E., Gu, J., McGivney, B. and MacHugh, D. 2010. Targets of selection in the Thoroughbred genome contain exercise - relevant gene SNPs associated with elite racecourse performance. *Anim Genet.* 41:56-63.
- Hill, E.W., Gu, J.J., Eivers, S.S., Fonseca, R.G., McGivney, B.A., Govindarajan, P., Orr, N., Katz, L.M. and MacHugh, D. 2010. A Sequence Polymorphism in MSTN Predicts Sprinting Ability and Racing Stamina in Thoroughbred Horses. *Plos One*. 5(1). English.
- Jones, J.H., Longworth, K., Lindholm, A., Conley, K., Karas, R., Kayar, S. and Taylor, C. 1989. Oxygen transport during exercise in large mammals. I. Adaptive variation in oxygen demand. *J. APPL. PHYSIOL.* 67(2):862-870.
- Khachigian, L.M., Lindner, V., Williams, A.J. and Collins, T. 1996. Egr-1-induced endothelial gene expression: a common theme in vascular injury. *Science*. 271(5254):1427-1431.
- Kim, H., Lee, T., Park, W., Lee, J.W., Kim, J., Lee, B.Y., Ahn, H., Moon, S., Cho, S. and Do, K.T. 2013. Peeling back the evolutionary layers of molecular mechanisms responsive to exercise-stress in the skeletal muscle of the racing horse. *DNA. RES.* 20(3):287-298.
- Moynagh, P.N., Williams, D.C. and O'Neill, L.A. 1993. Interleukin-1 activates transcription factor NF κ B in glial cells. *BIOCHEM. J.* 294(2):343-347.
- Ng, P.C. and Henikoff, S. 2001. Predicting deleterious amino acid substitutions. *GENOME. RES.* 11(5):863-874.
- O'Donovan, K.J., Tourtellotte, W.G., Millbrandt, J. and Baraban, J.M. 1999. The EGR family of transcription-regulatory factors: progress at the interface of molecular and systems neuroscience. *TRENDS. NEUROSCI.* 22(4):167-173.
- Oppenheim, J.J. 2001. Cytokines: past, present, and future. *INT. J. HEMATOL.* 74(1):3-8.
- Park, K.D., Park, J., Ko, J., Kim, B.C., Kim, H.S., Ahn, K., Do, K.T., Choi, H., Kim, H.M. and Song, S. 2012. Whole transcriptome analyses of six thoroughbred horses before and after exercise using RNA-Seq. *BMC genomics*. 13(1):473.
- Pereira, S.G. and Oakley, F. 2008. Nuclear factor- κ B1: regulation and function. *Int. J. Biochem. Cell Biol.* 40(8):1425-1430.

- Salem, K., Qureshi, M., Sydorenko, V., Parekh, K., Jayaprakash, P., Iqbal, T., Singh, J., Oz, M., Adrian, T. and Howarth, F. 2013. Effects of exercise training on excitation–contraction coupling and related mRNA expression in hearts of Goto-Kakizaki type 2 diabetic rats. *MOL. CELL. BIOCHEM.* 380(1-2):83-96.
- Sunyaev, S., Ramensky, V., Koch, I., Lathe, III.W., Kondrashov, A.S., Bork, P. 2001. Prediction of deleterious human alleles. *HUM. MOL. GENET.* 10(6):591-597.
- Tamura, T., Yanai, H., Savitsky, D. and Taniguchi, T. 2008. The IRF family transcription factors in immunity and oncogenesis. *Annu. Rev. Immunol.* 26:535-584.
- Thiel, G., Kaufmann, K., Magin, A., Lietz, M., Bach, K. and Cramer, M. 2000. The human transcriptional repressor protein NAB1: expression and biological activity. *BBA-GENE. STRUCT. EXPR.* 1493(3):289-301.
- Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blöcker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N., Penedo, M.C.T., Raison, J.M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D.F., Biagi, T., Binns, M.M., Chowdhary, B.P., Coleman, S.J., Della Valle, G., Fryc, S., Guérin, G., Hasegawa, T., Hill, E.W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J.R., Murray, J., Nergadze, S.G., Onofrio, R., Pedroni, S., Piras, M.F., Raudsepp, T., Rocchi, M., Røed, K.H., Ryder, O.A., Searle, S., Skow, L., Swinburne, J.E., Syvänen, A.C., Tozaki, T., Valberg, S.J., Vaudin, M., White, J.R., Zody, M.C., Broad Institute Genome Sequencing Platform., Broad Institute Whole Genome Assembly Team., Lander, E.S., and Lindblad-Toh, K. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science.* 326(5954):865-867.
- Wang, Z. and Moulton, J. 2001. SNPs, protein structure, and disease. *HUM. MUTAT.* 17(4):263-270.
- White, T.W. and Paul, D.L. 1999. Genetic diseases and gene knockouts reveal diverse connexin functions. *ANNU. REV. PHYSIOL.* 61(1):283-310.
- Willett, P. 1970. The Thoroughbred. *Putnam.*
- Wong, C.W., Christen, T., Roth, I., Chadjichristos, C.E., Derouette, J.P., Foglia, B.F., Chanson, M., Goodenough, D.A. and Kwak, B.R. 2006. Connexin37 protects against atherosclerosis by regulating monocyte adhesion. *NAT. MED.* 12(8):950.
- Yan, S.F., Fujita, T., Lu, J., Okada, K., Zou, Y.S., Mackman, N., Pinsky, D.J. and Stern, D.M. 2000. Egr-1, a master switch coordinating upregulation of divergent gene families underlying ischemic stress. *NAT. MED.* 6(12):1355.
- Young, L., Marlin, D., Deaton, C., Brown - Feltner, H., Roberts, C. and Wood, J. 2002. Heart size estimated by echocardiography correlates with maximal oxygen uptake. *EQUINE. VET. J.* 34(S34):467-471.