

Investigation of *De Novo* Unique Differentially Expressed Genes Related to Evolution in Exercise Response during Domestication in Thoroughbred Race Horses

Woncheoul Park¹, Jaemin Kim², Hyeon Jeong Kim³, JaeYoung Choi⁴, Jeong-Woong Park⁴, Hyun-Woo Cho⁴, Byeong-Woo Kim⁴, Myung Hum Park⁵, Teak-Soon Shin⁴, Seong-Keun Cho⁴, Jun-Kyu Park⁶, Heebal Kim^{1,3}, Jae Yeon Hwang¹, Chang-Kyu Lee¹, Hak-Kyo Lee⁷, Seoae Cho^{3*}, Byung-Wook Cho^{4*}

1 Department of Agricultural Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea, **2** Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea, **3** C&K genomics, Seoul National University, Seoul, Republic of Korea, **4** Department of Animal Science, College of Life Sciences, Pusan National University, Miryang, Republic of Korea, **5** TNT Research, Anyang, Republic of Korea, **6** Leaders in Industry-university Cooperation, Pusan National University, Miryang, Republic of Korea, **7** Genomic Informatics Center, Hankyong National University, Anseong, Republic of Korea

Abstract

Previous studies of horse RNA-seq were performed by mapping sequence reads to the reference genome during transcriptome analysis. However in this study, we focused on two main ideas. First, differentially expressed genes (DEGs) were identified by *de novo*-based analysis (DBA) in RNA-seq data from six Thoroughbreds before and after exercise, hereafter referred to as “*de novo* unique differentially expressed genes” (DUDEG). Second, by integrating both conventional DEGs and genes identified as being selected for during domestication of Thoroughbred and Jeju pony from whole genome re-sequencing (WGS) data, we give a new concept to the definition of DEG. We identified 1,034 and 567 DUDEGs in skeletal muscle and blood, respectively. DUDEGs in skeletal muscle were significantly related to exercise-induced stress biological process gene ontology (BP-GO) terms: ‘immune system process’; ‘response to stimulus’; and, ‘death’ and a KEGG pathways: ‘JAK-STAT signaling pathway’; ‘MAPK signaling pathway’; ‘regulation of actin cytoskeleton’; and, ‘p53 signaling pathway’. In addition, we found *TIMELESS*, *EIF4A3* and *ZNF592* in blood and *CHMP4C* and *FOXO3* in skeletal muscle, to be in common between DUDEGs and selected genes identified by evolutionary statistics such as F_{ST} and Cross Population Extended Haplotype Homozygosity (XP-EHH). Moreover, in Thoroughbreds, three out of five genes (*CHMP4C*, *EIF4A3* and *FOXO3*) related to exercise response showed relatively low nucleotide diversity compared to the Jeju pony. DUDEGs are not only conceptually new DEGs that cannot be attained from reference-based analysis (RBA) but also supports previous RBA results related to exercise in Thoroughbred. In summary, three exercise related genes which were selected for during domestication in the evolutionary history of Thoroughbred were identified as conceptually new DEGs in this study.

Citation: Park W, Kim J, Kim HJ, Choi J, Park J-W, et al. (2014) Investigation of *De Novo* Unique Differentially Expressed Genes Related to Evolution in Exercise Response during Domestication in Thoroughbred Race Horses. PLoS ONE 9(3): e91418. doi:10.1371/journal.pone.0091418

Editor: Marc Robinson-Rechavi, University of Lausanne, Switzerland

Received: September 8, 2013; **Accepted:** February 11, 2014; **Published:** March 21, 2014

Copyright: © 2014 Park et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by a grant from the Next Generation BioGreen 21 Program (No. PJ008106), Rural Development Administration, Republic of Korea and ‘Leaders Industry-University Cooperation’ project, Ministry of Education (MOE), Republic of Korea. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: MHP has the following financial competing interest: Paid employment in TNT Research Company Limited, Dongan-gu, Anyang-si, Republic of Korea. The rest of the authors have declared that no competing interests exist. This does not alter the authors’ adherence to all the PLoS ONE policies on sharing data and materials. HK has the following financial competing interest: Paid employment in C&K Genomics company INC, C&K genomics Main Bldg. #514, SNU Research Park, Seoul National University Mt.4-2. SC has the following financial competing interest: CEO of C&K Genomics company INC, C&K genomics Main Bldg. #514, SNU Research Park, Seoul National University Mt.4-2.

* E-mail: bwcho@pusan.ac.kr (BWC); seoae@cnkgenomics.com (SC)

Introduction

Since domestication, at around 3500 B.C.E, horses have mainly been used for riding and racing [1]. One domesticated breed of horses, the Thoroughbred, has been specifically bred for speed, endurance, and strength since the 18th century. The extreme selection for these traits has resulted in a highly adapted athlete [2] with very high aerobic capacity [3], and high skeletal muscle mass [4], which comprises over 55% of total body mass [5]. The Thoroughbred is an excellent breed for competitive horse racing and by extension a valuable model for studying exercise response. A previous study has shown that exercise training in Thorough-

bred resulted in coordinated changes in the expression of genes related to metabolism, oxidative phosphorylation and muscle structure [6].

Domestication leads to gradual changes at the genetic level by a process of selection in a population of animals or plants. Most domestic animals were selectively bred for the goal of benefitting human beings. Due to the combined effect of natural selection and human-controlled selective breeding, phenotypic changes, which are related to genetic mutation, accompany the domestication process. Some genetic mutations with beneficial phenotypic effects have been either highly enriched or vanished by selective sweeps [7]. A selective sweep is the reduction of genetic diversity in the

neighboring DNA of a fixed mutation. Selective sweep regions in the genome can potentially be identified by a genome scan, and the low variation interval surrounding the selected gene can be found by fine-scale mapping. Using such genome scans, selective sweeps have been identified in domestic and natural (wild progenitor) populations [8,9].

Previous horse transcriptome studies using RNA-seq were carried out by mapping sequence reads to a reference genome. However, reference genome assembly has been known to have flaws including missing expressed genes [10], hundreds to thousands of miss-assemblies and large genomic deletions [11], and problems in trans-spliced genes [12]. Therefore, the results and success of reference transcriptome assembly depends on both the availability and quality of the reference genome. On the contrary, *de novo* transcript assembly has several advantages. First, it does not depend on a reference genome [13]. This is a key advantage as many organisms do not have a high-quality finished reference genome. For these organisms, *de novo* assembly becomes the first analysis step. Also, it does not depend on the correct alignment of reads to known splice sites [14] or the prediction of novel splicing sites, both of which are required by reference-based assemblers. Trans-spliced transcripts and similar transcripts originating from chromosomal rearrangements can be assembled using the *de novo* approach. In addition, *de novo* transcriptome assembly can help researchers investigate genes that are absent in the reference genome due to the incompleteness of reference sequences [10]. Lastly, it can identify new transcripts and new transcript structures [15,16]. However, reconstruction of full-length transcripts from short reads with considerable sequencing error rates poses substantial computational challenges [17]. Still, *de novo* assembly in RNA-seq is an important approach for carrying out transcriptomic studies.

Recently, many *de novo* assembly software tools have been developed, most of which take the de Bruijn graph approach. This approach usually has two important parameters: k-mer length and coverage cutoff value [16]. Tools such as Trans-ABYSS [15], Trinity [17], ABYSS [18], Oases [19], Rnnotator [20], Multiple-k [21], SOAPdenovo [22] and Velvet [23] follow this approach. Considering these *de novo* assembly software tools, Manfred G Grabherr, *et.al* concluded that Trinity *de novo* assembly software tool is superior to others for a number of reasons. Specifically, Trinity fully reconstructs a large fraction of transcripts, including alternatively spliced isoforms and transcripts from recently duplicated genes. In addition, Trinity resolves ~99% of the initial sequencing errors, determines splice isoforms, and distinguishes transcripts from recently duplicated and identified allelic variants [17].

In this study, we used *de novo*-based analysis (DBA) to identify differentially expressed genes (DEGs) that could not be identified with reference-based analysis (RBA). Furthermore, we integrate DEGs with genes identified as being selected for during domestication to reveal genes that are related to the evolution of exercise response during the domestication process of the Thoroughbred.

Materials and Methods

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.

Analysis of horses RNA-seq data

1. RNA-seq data between before and after exercise. We generated RNA-seq data from six Thoroughbred horses before and after exercise as described in a previous study [24]. Samples of skeletal muscle and blood were taken from six Thoroughbreds before and after exercise. ‘Before exercise’ samples were collected from the triceps brachii of the right leg and from the jugular vein and carotid artery of each horse. After an adequate resting period of several hours, the horses were subjected to a 30-min trot. Then, immediately after this trot, the ‘after exercise’ samples were collected from the same tissues of each individual. Thoroughbreds usually canter for 17–18 min per day. For the purposes of this study, a 30-min trot was taken to be the equivalent to 17–18 min of cantering. Total RNA from the skeletal muscle and blood samples were isolated using TRIzol (Invitrogen) and the RNeasy RNA purification kit with DNase treatment (Qiagen). mRNA was isolated from the total RNA using oligo-dT beads, then reverse transcribed into double-stranded cDNA fragments. Construction and sequencing of an RNA sequence library for each sample was carried out based on Illumina HiSeq2000 protocols in order to generate 90 bp paired-end reads. Twenty-four sets of transcriptome data were generated from muscle and blood samples of six horses obtained before and after exercise.

The RNA-sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) under the accession number GSE37870.

2. *De novo*-based analysis (DBA). We used the Trinity *de novo* assembly software tool [17,25] following the default settings, except for the following options: number of CPU and alignment method: bowtie. First, Trinity tools generated a reference for each obtained sample (a total of 24 samples) to detect DEGs. Second, Trinity tools generated a reference for each individual (a total of 4 samples) to compare with the SNPs from the whole genome sequence, RNA-seq (using reference transcriptome assembly) and RNA-seq (using *de novo* assembly).

The component ID was converted to known transcript ID (ftp://ftp.ensembl.org/pub/release-73/fasta/equus_caballus/cdna/) using Blastall [26], a user-friendly, free open source tool, which is suited for short read alignment. After conversion, we filtered the transcript ID by alignment length of higher than 80%, more detail about the *de novo* assembly method is given in the Supplementary Methods (File S1).

Reference-based analysis (RBA): Most of the RBA used in this study is described in Kim *et al* [27]. TopHat [28] (ver.1.4.1) was used to map the sequences to a horse reference genome and annotated using the EquCab2 database (<http://hgdownload.cse.ucsc.edu/downloads.html#horse>).

3. DEG selection (*de novo* vs reference). We examined the differential expression of replicated count data by applying a method based on negative binomial model as implemented in the R package EdgeR [29]. This package was used because RNA sequence data may exhibit more variability than expected in a Poisson distribution due to wide dispersal in the genome. The method implemented in the EdgeR package automatically takes all known sources of variation into account. Significant DEGs were detected with a cut-off value of FDR<0.01, based on a paired design between ‘before exercise’ and ‘after exercise’.

4. Genotype calling and SNP calling (*de novo* vs reference). Three open source packages were used for downstream processing and variant calling: Picard Tools (<http://picard.sourceforge.net>), SAMtools [30] and the genome Analysis Toolkit

(GATK [31]). Substitution calls were made with the GATK Unified Genotyper [32]. All calls with a Phred-scaled quality of less than 30 were filtered out and VCFtools [33] was used for handling the vcf file format.

5. DAVID analysis (*de novo* vs reference). One simple but extremely widely used systems biology technique for highlighting biological processes is gene category over-representation analysis. In order to perform this analysis, genes are grouped into categories by a common biological property and then tested to find categories that are over represented among the differentially expressed genes. Gene Ontology (GO) categories are commonly used in this technique and there are many tools available for performing GO and KEGG pathway analysis. We used DAVID [34] web tool to convert the equine Ensembl gene IDs to official gene symbols. This was carried out by cross-matching the equine Ensembl gene IDs to the human Ensembl gene IDs and the official gene symbols. 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 [35] within DAVID. The EASE tool is a modified Fisher's exact test used to measure enrichment of gene ontology (GO) terms [36]. To identify enriched GO terms, functionally clustered genes were filtered by an EASE value of <0.01 . In addition, A KEGG pathway enrichment test was performed using EASE, with a cut-off value <0.1 .

6. Quantitative real-time reverse transcript-PCR (qRT-PCR) validation. A blood sample was obtained from a Thoroughbred horse maintained at Ham-an Racing Horse Resort & Training Center before and after exercise. Exercise was performed as a 30-min trotting on a treadmill. Trizol reagent (Invitrogen) was used to extract total RNA from leukocytes after exercise, according to the Invitrogen manual. In order to prevent contamination of genomic DNA, RNase-free DNase kit (Qiagen) was used according to the manufacturer's operating manual. Total RNA quantification was performed by using NanoDrop® ND-1000 Spectrophotometer. cDNAs were synthesized in a reaction with oligo-dT primers, moloney-murine leukemia virus (MMLV) reverse transcriptase (Promega), RNase inhibitor (Promega) and RNase-free ddH₂O, which was incubated at 37°C for 4 h.

To confirm the *de novo* unique differentially expressed genes revealed by RNA-Seq, seven DUDEGs were analyzed by RT-PCR amplification. The primers were designed using the PRIMER3 software (<http://frodo.wi.mit.edu/primer3/>) (Table S9 in File S1).

The RT-PCR conditions were as follows: an initial step of 94°C for 10 min, 35 cycles of 94°C for 30 sec, 60°C for 30 sec, 72°C for 30 sec, and final step of 72°C for 10 min. PCR bands were normalized with *glyceraldehyde-3-phosphate dehydrogenase* (GAPDH) band. RT-PCR products were visualized by gel electrophoresis on a 2.0% SeaKem LE agarose gel.

cDNA was analyzed by BioRad CFX-96. All samples were measured in triplicate to ensure reproducibility, and C_t value was calculated using 2^{-ΔΔC_t} method [37].

Analysis of horse whole genome re-sequencing data

1. Whole genome re-sequencing data of Thoroughbred and Jeju domestic ponies. Whole-blood samples were collected from 18 Thoroughbred racing stallions of the Korean Racing Authority, and from four male and two female Jeju domestic ponies (*Equus caballus*) of the Jeju Provincial Livestock Institute, Korea. A 10 mL sample of blood was drawn from the carotid artery of each horse and was treated with heparin to prevent clotting. Genomic DNA was extracted and a quality check was carried out using fluorescence-based quantification on an

agarose gel, a standard electrophoresis on a 0.6% agarose gel and, via a pulsed-field gel, using 200 ng of DNA. Manufacturers' instructions were followed to create a paired library of 500-bp fragments. This consisted of the following: purified genomic DNA fragments of less than 800 bp, fragments with blunt ends, fragments with 5' phosphorylated ends, fragments with a 3'- dA overhang, some with adaptor-modified ends, purified ligation product, and a genomic DNA library. Following this, we generated sequence data using HiSeq 2000 (Illumina, Inc).

2. Reference genome assembly. Using the Burrows-Wheeler Aligner [38] with the default setting, pair-end sequence reads were mapped to the reference horse genome (ftp://ftp.ensembl.org/pub/release-73/fasta/equus_caballus/dna/) (Table S8 in File S1). The DNA re-sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA) database under the accession numbers SRA053569, SRA054885 and SRP017702.

3. Genotype calling and SNP calling. We used the following open-source software packages; Picard Tools, SAMtools, and the Genome analysis toolkit, for downstream processing and variant calling. Substitution calls were made with GATK UnifiedGenotyper20 and all calls with a Phred-scaled quality of less than 30 were filtered out. For each chromosome, we simultaneously inferred the phased haplotype and inputted the missing alleles for the entire set of Thoroughbred populations using BEAGLE [39].

4. Estimation of Nucleotide diversity, F_{ST} and Cross Population Extended Haplotype Homozygosity (XP-EHH) value. Nucleotide diversity and long run of homozygosity (LROH) of Thoroughbred and Jeju domestic ponies for each chromosome were calculated by VCFtools. Conventional F_{ST} [40] and Reynolds F_{ST} [41] values were calculated for genes using Arlequin 3.5 [42] based on pairwise differences between the haplotypes of Thoroughbred and Jeju domestic ponies. In order to calculate F_{ST}, we used the horse genome to phase the haplotypes of the two populations. Also, to calculate F_{ST} by each gene region, we used the genomic information (Ensembl Genes71, EquCab2), namely the Ensembl reference annotated gene information. We selected the genes of the top 1% of the empirical distribution (empirical p-value <0.01) [43]. The method Cross Population Extended Haplotype Homozygosity (XP-EHH) was used to detect selective sweeps using the software xpehh [44] (<http://hgdp.uchicago.edu/Software/>). For XP-EHH analysis, we used haplotype information for all SNPs of the entire autosome, and we calculated Extended Haplotype Homozygosity (EHH) and the log-ratio integrated EHH (iHH) for the pairwise test of the Thoroughbred and Jeju domestic pony populations. The log ratios were standardized to have a mean of 0 and a variance of 1, and p-values were assigned assuming a normal distribution. We selected SNPs with p-values <0.01 , which are considered to have strong selection signals. Then we apply a cutoff value of XP-EHH values <0 for finding adaptation in the Thoroughbred. We chose genes related with these SNPs by identifying genes located within a 10 kb [45] boundary of these SNPs. Since XP-EHH is not sensitive to allele frequencies, there is no need to stratify the data into frequency bins before determining significance. The p-values are empirical p-values; that is, a low p-value indicates that a locus is an outlier with respect to the rest of the genome. However, we note that loci detected as being under selection using this approach may be an under-representative sample of all truly selected loci; in particular, selection on standing variation and recessive loci are likely to be underrepresented [43].

Results

Differences in the results of reference-based and *de novo*-based assembly and analysis

Transcriptome analysis results of reference-based analysis (RBA) and *de novo*-based analysis (DBA) showed a substantial difference in the number of transcript and differentially expressed genes (DEGs) identified. In RBA, for blood and skeletal muscle, 15,900 and 17,927 transcripts were found, respectively, among which 2,244 and 1,405 were unique transcripts. In DBA, the numbers of transcripts in skeletal muscle and blood were 18,057 and 19,413, respectively with 4,401 and 2,892 unique transcripts. The numbers transcripts in common between RBA and DBA were 13,656 for skeletal muscle tissue and 16,521 for blood (Figure 1a and Figure S1a in File S1). When the sample variance of RBA and DBA in skeletal muscle and blood were compared using multidimensional scaling (MDS) plot, the results for the two analyses were almost identical. The skeletal muscle samples were clustered into two subgroups: before and after exercise, but the blood samples did not show any clustering (Figure 1b, Figure S1b and Figure S2 in File S1). In RBA, the number of DEG in skeletal muscle and blood were 2,818 and 455, respectively with 2,200 and 427 DEGs being unique to RBA. In DBA, the number of DEG in skeletal muscle and blood were 1,652 and 595, respectively with 1,034 and 567 unique DEGs. The number of DEGs identified by both RBA and DBA were 618 and 28 in skeletal muscle and blood, respectively (Figure 1c and Figure S1c in File S1). These DEGs were compared using Heatmap visualization to examine their expression pattern in each analysis. The expression pattern was similar, however, the intensity of the expression was higher with DBA (Figure 1d and Figure S1d in File S1). Overall, in comparison to RBA, DBA identified a higher number of transcripts but a lower number of DEGs.

We detected SNPs from two different next-generation sequencing methods (WGS and RNA-seq) and two different assembly methods (RBA and DBA) for each Thoroughbred sample (F1, F2 and F3 = male, S3 = female). The number and rate of SNPs in DBA of RNA-seq was 108,158 (0.031%), 110,502 (0.031%), 105,920 (0.03%) and 101,887 (0.029%) in F1, F2, F3 and S3 respectively, and the number and rate of SNPs in from RBA of RNA-seq were 284,859 (0.012%), 287,286 (0.012%), 276,241 (0.011%) and 265,729 (0.011%) in F1, F2, F3 and S3, respectively (Table S2 in File S1).

Identification of *de novo* unique differentially expressed genes (DUDEGs) before and after exercise

We identified DUDEGs from RNA-seq data using the expression profiles of genes in skeletal muscle and blood samples taken from six Thoroughbreds before and after exercise. There were a total of 1,034 significant DUDEGs (519 up-regulated, 515 down-regulated) in skeletal muscle and 567 (314 up-regulated, 253 down-regulated) in blood (FDR < 0.01). Among them, 456 (61 up-regulated, 395 down-regulated) in skeletal muscle and 205 (93 up-regulated, 112 down-regulated) in blood were annotated (Table S1 in File S1).

Validation of DUDEGs in horse RNA-seq data using RT-PCR

We performed RT-PCR to validate the DUDEGs detected in horse blood. The seven genes (*TIMELESS*, *EIF4A3*, *PGIW*, *ANK3*, *MSH3*, *SYNRG*, *ASGR2*: 2 up-regulated and 5 down-regulated) were randomly selected with conceptually new DEGs and $\log_{2}FC > 2$ in blood (Table S1, S4 and S5 in File S1). The

expression levels of DUDEGs between RNA-seq and RT-PCR were highly similar (Figure S8 in File S1). The results confirmed that DUDEGs identified in this study were reliable.

Functional annotation of DUDEGs

We summarized the highest biological process gene ontology (BP-GO) of DUDEGs in skeletal muscle sample taken before and after exercise from six Thoroughbred RNA-seq data (Figure 2). The other BP-GO of DUDEGs were summarized separately (Figure S3, Figure S4, Figure S5 and Figure S6 in File S1). The most significantly enriched terms among up-regulated genes in skeletal muscle were 'biological adhesion', 'biological regulation', 'death', 'growth', 'immune system process', 'locomotion', 'multi-organism process', and 'response to stimulus' (P-value = 6.29E-02, P-value = 9.57E-04, P-value = 1.75E-06, P-value = 7.56E-03, P-value = 2.23E-13, P-value = 3.48E-03, P-value = 5.71E-04, and P-value = 5.16E-08, respectively). While, the most significantly enriched terms among down-regulated genes in skeletal muscle were 'cellular component organization', 'cellular process', 'establishment of localization', 'localization', and 'metabolic process' (P-value = 4.94E-04, P-value = 9.12E-04, P-value = 7.09E-02, P-value = 6.87E-02, P-value = 5.28E-02, and P-value = 1.89E-2, respectively). 'Developmental process' was the most significantly enriched term in skeletal muscle among both up and down-regulated genes. However, no terms were highlighted as being significantly enriched in blood. We summarized the cellular components and molecular function gene ontology of DUDEGs in RNA-seq data from skeletal muscle and blood of six Thoroughbreds before and after exercise (Table S3 in File S1).

Enriched KEGG pathways analysis using DUDEGs revealed that up-regulated genes in skeletal muscle and blood are associated with exercise-induced stress. The most significantly enriched terms in skeletal muscle were 'p53 signaling pathway', 'regulation of actin cytoskeleton', 'JAK-STAT signaling pathway', 'MAPK signaling pathway', 'cell adhesion molecules', 'cytokine-cytokine receptor interaction', 'bladder cancer', and 'pathways in cancer'. In addition, the two terms 'colorectal cancer' and 'biosynthesis of unsaturated fatty acids' were significantly enriched in blood (Table 1).

Integration of DUDEG and selected gene associated with nucleotide diversity, F_{ST} and Cross Population Extended Haplotype Homozygosity (XP-EHH)

The genes under selection were further investigated using nucleotide diversity, F_{ST} and XP-EHH from WGS of Thoroughbred and Jeju pony (skeletal muscle = 1,033 and blood = 567). We found 12 genes (*ZNF592*, *CD58*, *C1orf162*, *USP37*, *FOXM1*, *TIMELESS*, *TRMT1*, *CALR*, *ASNA1*, *EIF4A3*, *SYNRG* and *FADS1*) in blood and 14 genes (*HERC2*, *CHD9*, *DDX28*, *CAPZ1*, *TSEN15*, *CHMP4C*, *FOXO3*, *PLD2*, *ANKRD13D*, *UNKL*, *CBFA2T2*, *NECAB3*, *SLC25A29* and *FBLN1*) in skeletal muscle that were both identified as DUDEGs and implicated in F_{ST} analysis as a selected gene (Table S4 in File S1). The F_{ST} distribution histogram for the pair of horse breeds is shown in Figure S7 in File S1.

Between DUDEGs and XP-EHH (selected genes in the Thoroughbred), we found 11 (*ANK3*, *ZNF592*, *TOR1AIP1*, *TIMELESS*, *INSR*, *MED15*, *ZNF567*, *EIF4A3*, *EXOC6B*, *PPP4R2* and *DYRK1A*) and 48 common genes (*URB2*, *NEO1*, *CORO2B*, *CPE*, *FUK*, *SYCE1L*, *TEC*, *RELL1*, *ZNF775*, *RGS4*, *FAM46C*, *PTGFRN*, *DENND2C*, *CTTNBP2NL*, *VAV1*, *BARX2*, *RAB31*,

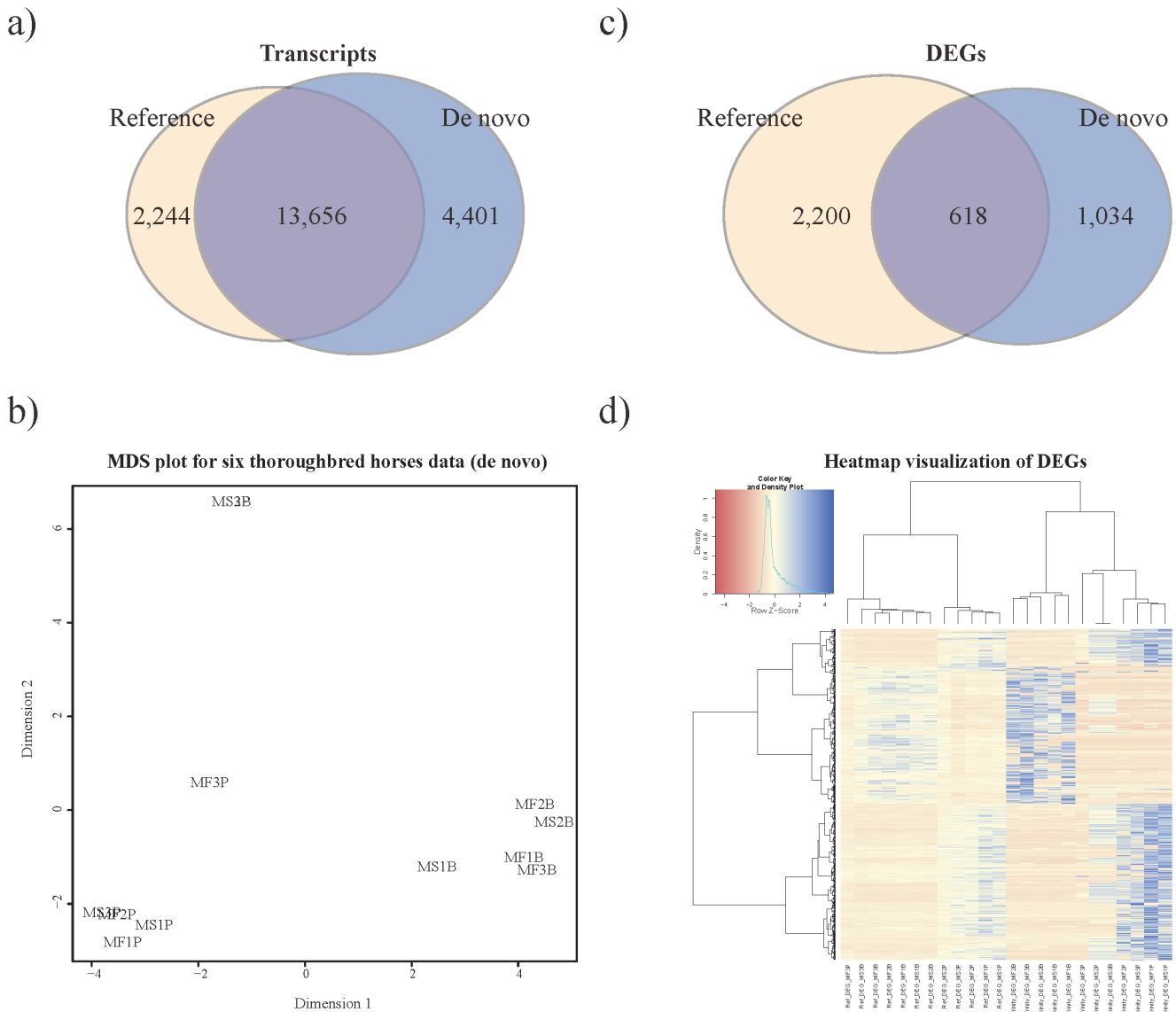


Figure 1. Summary of comparative analysis between *de novo* assembly and reference genome assembly from RNA-seq data of skeletal muscle from six Thoroughbreds before and after exercise (Total 12 samples). a) The number of transcripts in common between *de novo* assembly and reference genome assembly b) MDS plot of six Thoroughbreds before and after exercise using *de novo* assembly. c) The number of DEGs between *de novo* assembly and reference genome assembly: rows represent DEGs from skeletal muscle and columns represent assemble method from 6 horse samples (*First 'B' is for blood and 'M' is for muscle. 'F1', 'F2', 'F3' and 'S3' are horse samples. Last 'B' is for 'before exercise' and 'P' is for 'after exercise'). doi:10.1371/journal.pone.0091418.g001

VPS4B, CHMP4C, DEPTOR, ATG5, FOXO3, OSBPL7, PEX12, PIK3R5, PTTX1, MARCH3, LMNB1, ST8SLA4, RASGRF2, ARSB, INPP4A, RNF144A, PPP4R2, FRMD4B, SPATA13, SLC7A1, CAB39L, B3GALT1, MECOM, PARP14, NPR3, TGM3, DHX35, AUH, C14orf102, COL27A1 and HLCS) in blood and skeletal muscle, respectively (Table S5 in File S1). Among them, three genes, *TIMELESS, EIF4A3* and *ZNFX592*, in blood, and two genes, *CHMP4C* and *FOXO3*, in skeletal muscle were shown to be significant in all three analyses (DUDEG, F_{ST} , and XP-EHH). In comparison to Jeju pony, the Thoroughbred showed relatively low levels of nucleotide diversity at three out of the five identified genes (Table 2 and Figure 3).

Discussion

Differences in result between reference-based and *de novo*-based assembly and analysis

Up to now, many studies of RNA-seq analyses have used reference-based analysis (RBA) when a reference genome for the species is available [27,46,47]. When a species does not have a reference genome, RBA using the reference genome of a closely related species or *de novo*-based analysis (DBA) is used. Several studies of RNA-seq analyses have used the align-then-assemble combined method (align-then-assemble and assemble-then-align) [24]. DBA used in align-then-assemble method assembles the unmapped sequence reads after RBA, which supplements the inherent weakness of RBA. However, we suggest that even if a

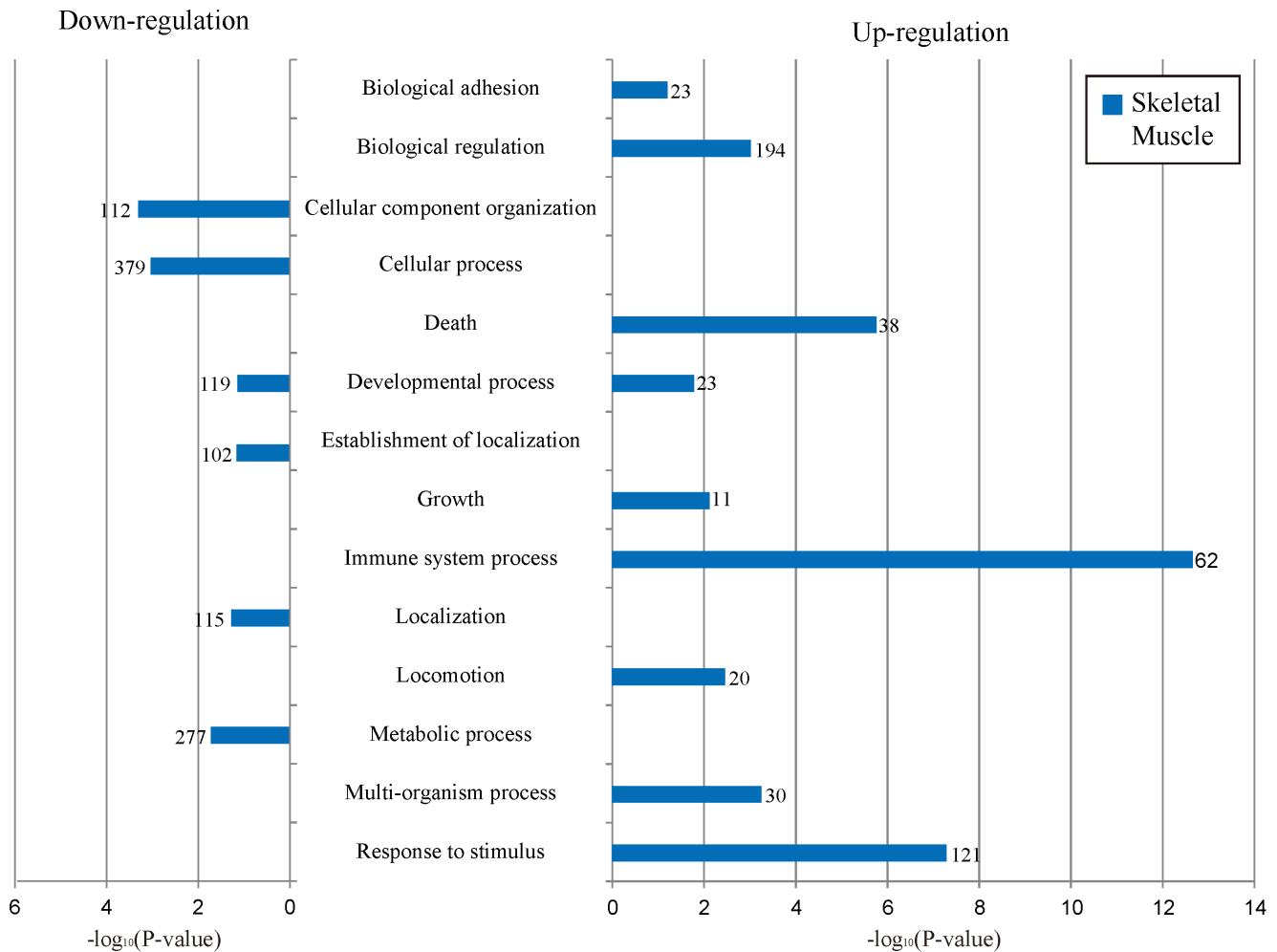


Figure 2. Biological process GO terms of tissues specific DEGs between before and after exercise in Thoroughbred. Up-regulated genes indicate higher activation after exercise than before and down-regulation genes indicate lower activation after exercise than before exercise. doi:10.1371/journal.pone.0091418.g002

species has a reference genome, DBA used in assemble-then-align is necessary for assembly of the total sequence reads including the unmapped sequence reads (Table S6 and S7 in File S1). In this study, we found a significant difference in the number of transcript and differentially expressed genes (DEGs) identified by RBA and DBA. Also, a greater number of unique transcripts were identified by DBA than in RBA (Figure 1a and Figure S1a in File S1). This implies that the horse RNA-seq data in this study includes new transcripts and new transcript structures that are not included in the horse reference genome. Additionally, we found *de novo* unique differentially expressed genes (DUDEGs) which cannot be attained from RBA (Figure 1c and Figure S1c in File S1). Credibility of DBA in RNA-seq has been proven in numerous methods and protocol papers [17,25,48] and in this study. Both multidimensional scaling (MDS) plot (Figure 1b, Figure S1b and Figure S2 in File S1) and expression patterns of common DEGs did not show differences between the results of RBA and DBA. However, the intensity of the expression was different because the newly assembled transcriptome reference represents the individual transcriptome made using de Bruijn graph assemblers during DBA [48] (Figure 1d and Figure S1d in File S1).

RNA-Seq can reveal sequence variations such as SNP in genes [49] as is possible with whole genome sequencing (WGS).

Transcribed SNPs in RNA-seq are needed for accurate measurement of allele-specific expression [50,51] and detection of novel SNPs. [52]. Hence, we compared the number and rate of SNPs identified from the two NGS methods and found differences between the type of references and NGS methods. In summary, we detected more SNPs in *de novo* assembly of RNA-seq than in the reference genome assembly of cDNA (Table S2 in File S1).

Identification and Functional annotation of unique DEGs identified by *de novo* base assembly

We identified DUDEGs to ascertain the important function of DEGs, which cannot be attained from RBA. In the highest biological process gene ontology (BP-GO) of DUDEG result, immune system process had the most significant P-value (P-value = 2.23E-13) in up-regulation of skeletal muscle (Figure 2). Response to stimulus had the second most significant P-value (P-value = 5.16E-08), which is related with immune response caused by exercise-induced stress [53]. Exercise-induced stress is closely related with the regulation of immune response [6,27]. Over-exercise in horses has shown an increase in the expression of alpha-1-antitrypsin protein, which plays an important role in protection of cells from inflammatory enzymes released from neutrophils [54]. Exercise-induced reactive oxygen species was

Table 1. Enriched KEGG pathways associated with DEGs in two tissue such as skeletal muscle and blood.

Highest KEGG	Higher KEGG	KEGG	Blood		Muscle	
			UP	DOWN	UP	DOWN
Cellular Processes	Cell growth and death	Cell cycle				V
		p53 signaling pathway			V	
Environmental Information Processing	Cell motility	Regulation of actin cytoskeleton			V	
	Membrane transport	ABC transporters				V
	Signal transduction	Jak-STAT signaling pathway		V		
		MAPK signaling pathway		V		
		Notch signaling pathway				V
		Phosphatidylinositol signaling system				V
Genetic Information Processing	Signaling molecules and interaction	Cell adhesion molecules (CAMs)		V		
		Cytokine-cytokine receptor interaction		V		
	Folding, sorting and degradation	Ubiquitin mediated proteolysis				V
	Replication and repair	Non-homologous end-joining				V
Human Diseases	Cancers	Colorectal cancer	V			
		Bladder cancer		V		
		Pathways in cancer			V	
		Small cell lung cancer		V		
Metabolism	Carbohydrate metabolism	Butanoate metabolism				V
		Inositol phosphate metabolism				V
	Glycan biosynthesis and metabolism	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis			V	
	Lipid metabolism	Biosynthesis of unsaturated fatty acids	V			
Organismal Systems	Immune system	B cell receptor signaling pathway			V	
		Hematopoietic cell lineage			V	
		Natural killer cell mediated cytotoxicity			V	
		Toll-like receptor signaling pathway			V	
		T cell receptor signaling pathway			V	

For each set of up-regulated and down-regulated, DEG in skeletal muscle and blood, a KEGG pathway enrichment analysis was performed. Starting from the right, the table shows: tissue type, status of regulation, KEGG pathway terms, higher-level KEGG pathway terms, and the highest level of KEGG pathway terms.
doi:10.1371/journal.pone.0091418.t001

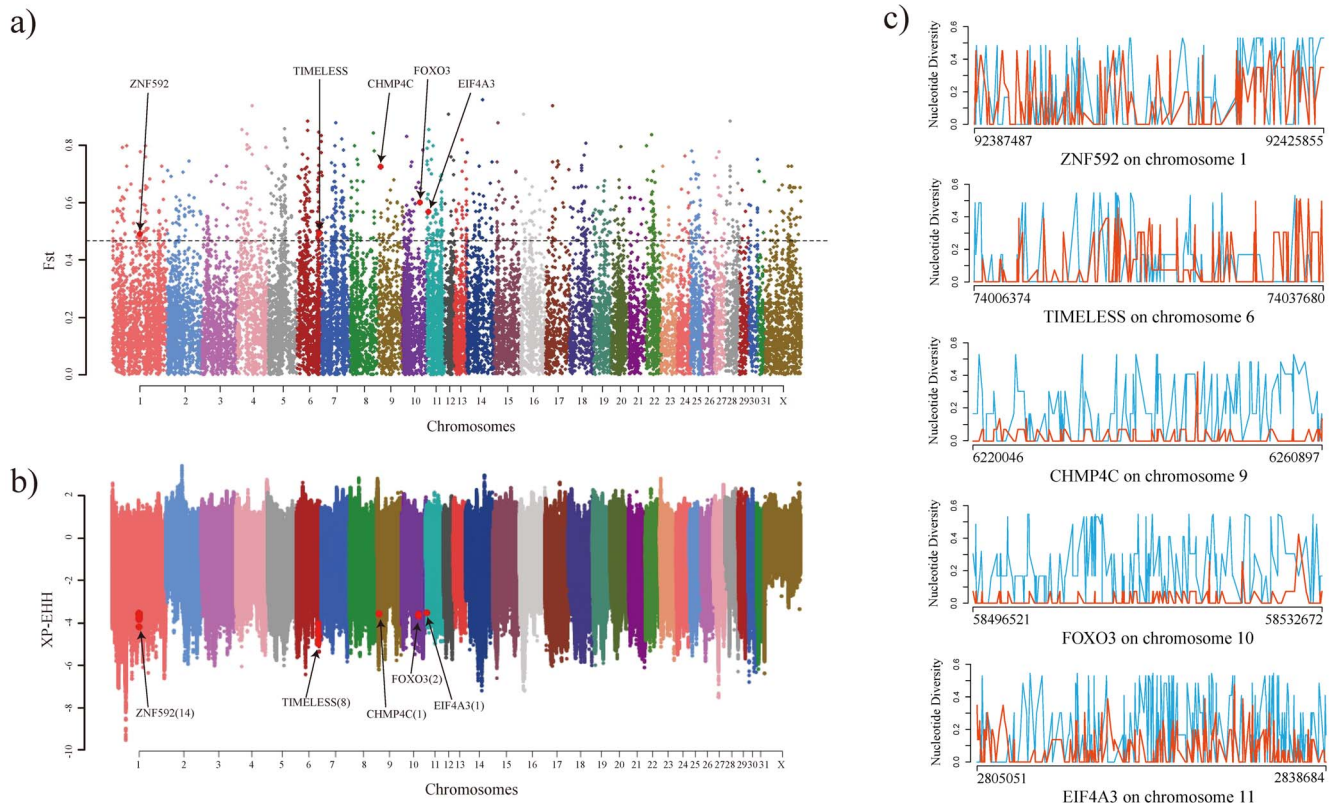


Figure 3. Signatures of correlation between DEGs from Thoroughbred RNA-seq and selected genes associated with nucleotide diversity, F_{ST} and XP-EHH from Thoroughbred and Jeju pony DNA sequence. a) Manhattan plot of F_{ST} (Dotted line = cut-off value of the top 5% with empirical p-values of <0.05). Red point = common genes between the DEGs, Thoroughbred selected genes associated with F_{ST} and XP-EHH. b) Manhattan plot of XP-EHH value (Red point = Common genes between the DEGs and Thoroughbred selected genes associated with F_{ST} and XP-EHH). c) Nucleotide diversity line plot of five common genes (sky blue color line = Jeju pony, orange color line = Thoroughbred). doi:10.1371/journal.pone.0091418.g003

also related with the regulation of immune responses, and caused the inflammatory responses from muscle damage [55,56]. In the KEGG pathways result (Table 1), the JAK-STAT signaling pathway and MAPK signaling pathway were also up-regulated in the skeletal muscle. The JAK-STAT signaling is a key pathway in myoblast proliferation [57] and plays a major role in inflammatory and immune responses [58]. The MAPK signaling pathway is implicated in inflammation and carbohydrate metabolism [59]. Death, Related with apoptosis of skeletal muscle caused by over-exercise also had a significant P-value (P-value = 1.75E-06). This was supported by the KEGG pathways results, which showed that the p53 signaling pathway and regulation of actin cytoskeleton were up-regulated. P53 protein has an important role in apoptosis of skeletal muscle and actin cytoskeleton, and is also a key pathway in regulation of apoptosis pathways [60,61]. However, we did not find immune responses and apoptosis related with BP-GO and KEGG pathway in blood.

Integration of conceptually new DEGs: DUDEGs and selected gene associated with nucleotide diversity, F_{ST} and Cross Population Extended Haplotype Homozygosity (XP-EHH)

In order to investigate the evolutionary history of domestication in relation to different experimental conditions, we approached the identification of DEGs with a new concept. The conceptually new DEGs were attained by screening for genes in common between DUDEGs from DBA in RNA-seq and selected genes

identified by evolutionary statistics, such as nucleotide diversity, F_{ST} and XP-EHH from RBA in WGS.

This comparison highlighted three genes (*EIF4A3*, *ZNF592* and *TIMELESS*) in blood and two genes (*CHMP4C*, and *FOXO3*) in skeletal muscle as being in common between DUDEGs, F_{ST} and XP-EHH. These five genes are not only DUDEGs in six Thoroughbreds, before and after exercise, but also selected genes (F_{ST} (empirical p-value <0.01) and XP-EHH (value <0 and p-value <0.01)). A pairwise test, XP-EHH, of the Thoroughbred and Jeju domestic pony populations was used to identify selective sweep regions between the two populations. As we are interested in locating selective sweep regions representing adaptation in Thoroughbreds, a cutoff of XP-EHH value <0 was used. If the XP-EHH value >0 is used, then the identified selective sweep region would correspond to the adaptations in Jeju domestic pony.

The five genes were conceptually new DEGs, and were related to the evolution of exercise response during the domestication process of Thoroughbred. Among them, three genes, *CHMP4C*, *EIF4A3* and *FOXO3*, showed relatively low levels of nucleotide diversity compared to that of the Jeju pony (Table 2 and Figure 3). This suggests that these three genes have been more strongly selected for in Thoroughbred than in Jeju pony. *EIF4A3* was mostly expressed in megakaryocytes, platelets and red blood cell. *EIF4A3*, an mRNA-localization protein in mammals, controls the synaptic strength, neuronal protein expression, and in megakaryocytes and platelets act as mRNA sorting machinery [62–64]. In a previous study, it was shown that over-exercise activates and

Table 2. Co-matching genes between the DEGs, selected genes associated with F_{st} (F_{st} cut-off value top 5% with empirical p -value <0.05) and Thoroughbred selected genes associated with XP-EHH (XP-EHH cut-off value empirical p -value <0.01 and XP-EHH value < -3.51551 significant SNPs).

Sample Tissue	Ens ID	CHR	Start	End	DEG logFC	DEG P-value	DEG FDR	Gene symbol	Reynolds	Fst	SNP region	XP-EHH	XP-EHH P-value			
Blood	ENSECAG00000015925	11	2814986	2828951	-17.23	1.81E-34	2.94E-31	EIF4A3 (eukaryotic translation initiation factor 4A3)	0.83481	0.56604	2821795	-3.52526	9.73E-03			
														ENSECAG00000016283	1	92397477
	92397597	-3.78677	4.48E-03													
	92398170	-3.74169	5.15E-03													
	92398472	-3.78551	4.50E-03													
	92398776	-3.83657	3.84E-03													
	92398791	-4.18477	1.23E-03													
	92398828	-3.56243	8.74E-03													
	92398986	-3.5556	8.92E-03													
	92399035	-3.76415	4.81E-03													
92402102	-3.60557	7.72E-03														
92402157	-3.58028	8.31E-03														
92402965	-3.58305	8.24E-03														
92403116	-3.61357	7.54E-03														
92403392	-3.55258	9.00E-03														
Muscle	ENSECAG0000002892	6	74016170	74027809	9.80616	1.34E-12	6.75E-11	TIMELESS (timeless circadian clock)	0.67349	0.49007	74019099	-4.02647	2.09E-03			
														74019114	-4.05026	1.93E-03
														74019135	-4.28442	8.68E-04
														74019141	-4.56193	3.15E-04
														74019149	-4.42328	5.27E-04
														74019152	-4.43649	5.02E-04
														74019153	-4.81799	1.16E-04
														74019241	-4.96438	6.40E-05
														6257522	-3.56699	8.63E-03
														58521033	-3.5866	8.15E-03
58521300	-3.67329	6.32E-03														

doi:10.1371/journal.pone.0091418.t002

increases platelets [65]. Although, it has an important role in blood post-exercise in Thoroughbreds, *EIF4A3* expression was up-regulated in our results. *CHMP4C* is a p53-regulated gene and plays an important role in exosome production [66]. The importance of p53 in apoptosis of skeletal muscle was implicated in a previous study, in which p53-null animals showed greater fatigability and less locomotory endurance than wild-type animals [60]. This suggests that p53 is closely related with exercise-induced stress in skeletal muscle. *CHMP4C* expression was down-regulated in our results implying the activation of p53 regulation in Thoroughbred skeletal muscle. *FOXO3*, also known as Forkhead box O3, has a role in triggering apoptosis by down-regulating the *FOXO3* gene. In addition, *FOXO3* causes a loss of muscle mass, and is closely related to PGC1 α , ATG4b, ATG12, Beclin1, Gabarapl1, and LC3b. PGC1 α , the transcription of atrophy-specific genes, inhibits the activity of the transcription factor *FOXO3*, with protects skeletal muscle from atrophy [67]. In human muscle after ultra-endurance exercise, the expression of several autophagy genes, ATG4b, ATG12, Beclin1, Gabarapl1 and LC3b, were increased [68,69]. For this reason, *FOXO3* was also closely related to exercise in Thoroughbred skeletal muscle.

Based on these results, *EIF4A3*, *CHMP4C* and *FOXO3* are conceptually new DEGs involved in exercise response that have been selected for during the domestication history of the Thoroughbred that cannot be acquired by RBA.

Supporting Information

File S1 Contains the following. Table S1, List of unique DEGs in skeletal muscle and blood in six Thoroughbred horses before and after exercise RNA-seq data by de novo assembly (FDR<0.01). Table S2, The number and rate of SNPs from different next-generation sequencing method (DNA and RNA sequencing) and different reference genome assembly in each Thoroughbred horse sample (F1, F2 and F3 = male, S3 = female). Table S3, GO terms of cellular components and molecular function of two tissues specific DEGs between before the exercise and after exercise in horses. Table S4, Common genes between DEGs and selected genes associated with F_{ST} (F_{ST} cut-off value top 5% with empirical p-value<0.05). Table S5, Common genes between DEGs and selected genes associated with XP-EHH: XP-EHH cut-off value empirical p-value<0.01 and XP-EHH value <-3.51551 significant SNPs in Thoroughbred were selected and >1.73481 significant SNPs in Jeju domestic pony were selected. Table S6, List of basic stats such as the number of transcripts, components, and contig N50 value in RNA-seq whole reads and unmapped reads by trinity de novo assembly. Table S7, Number of annotated transcripts from RNA-seq unmapped reads by trinity de novo assembly. The number in the parentheses is the number of transcripts that were not included in the results of the reference-based analysis. Table S8, Basic information of 4 horses re-sequencing data. Table S9, RT-PCR primer information such as the gene symbol, direction and sequence. Figure S1, Summary of comparative analysis between de novo assemble and reference genome assemble from blood in six Thoroughbred horses before and after exercise RNA-seq data (Total 12 samples). a) The number of common transcripts of 12 samples between de novo assemble and reference genome assemble b) MDS plot of six Thoroughbred horses before and after exercise using de novo assemble. c) The number of DEGs between de novo assemble and reference genome assemble. d) Heat-map visualization of common DEGs between de novo assemble and reference genome assemble:

rows represent DEGs from blood and columns represent assemble method from 6 horse samples (*First 'B' is for Blood and 'M' is for muscle. 'F1', 'F2', 'F3' and 'S3' are horse samples. Last 'B' is for 'before exercise' and 'P' is for 'after exercise'). Figure S2, MDS plot of six Thoroughbred horses before and after exercise using reference genome assemble in RNA-seq. a) MDS plot of blood tissue in six Thoroughbred horse before and after exercise. b) MDS plot of skeletal muscle tissue in six Thoroughbred horse before and after exercise. (*First 'B' is for Blood and 'M' is for muscle. 'F1', 'F2', 'F3' and 'S3' are horse samples. Last 'B' is for 'before exercise' and 'P' is for 'after exercise'). Figure S3, Hierarchical clustering of biological process GO terms associated with up-regulated DEGs in blood. The gene list of each GO term clustered using DAVID was compared to calculate the distance between the GO terms. For a distance value >0.5, GO terms were re-clustered, and GO term groups are shown as light-blue graduated blocks. The number of genes associated with the re-clustered GO term group is shown on the left side of the block. Figure S4, Hierarchical clustering of biological process GO terms associated with down-regulated DEGs in blood. The gene list of each GO term clustered using DAVID was compared to calculate the distance between the GO terms. For a distance value >0.5, GO terms were re-clustered, and GO term groups are shown as light-blue graduated blocks. The number of genes associated with the re-clustered GO term group is shown on the left side of the block. Figure S5, Hierarchical clustering of biological process GO terms associated with up-regulated DEGs in muscle. The gene list of each GO term clustered using DAVID was compared to calculate the distance between the GO terms. For a distance value >0.5, GO terms were re-clustered, and GO term groups are shown as light-blue graduated blocks. The number of genes associated with the re-clustered GO term group is shown on the left side of the block. Figure S6, Hierarchical clustering of biological process GO terms associated with down-regulated DEGs in muscle. The gene list of each GO term clustered using DAVID was compared to calculate the distance between the GO terms. For a distance value >0.5, GO terms were re-clustered, and GO term groups are shown as light-blue graduated blocks. The number of genes associated with the re-clustered GO term group is shown on the left side of the block. Figure S7, Histogram of conventional F_{ST} frequency between Thoroughbred and jeju pony. x-axis is conventional F_{ST} value, y-axis is gene frequency. Figure S8, qRT-PCR validation of de novo unique differentially expressed genes (DUDEGs) identified from the RNA-seq data set of Thoroughbred horses before and after exercise: a) RT-PCR of six DUDEG in horses before exercise and after exercise. b) qRT-PCR results depicted as Ct value was calculated using $2^{-\Delta\Delta C_t}$ method. *: p-value<0.05. **: p-value<0.01. #: The expression patters of genes supported the result of our analysis. (ZIP)

Acknowledgments

We thank Hyeongmin Kim, Taeheon Lee and woori Kwak for their comments on drafts of the manuscript.

Author Contributions

Conceived and designed the experiments: WP HK BWC. Performed the experiments: JC JWP HWC BWK MHP TSS SKC JKP BWC JYH CKL. Analyzed the data: WP JK HJK. Contributed reagents/materials/analysis tools: HWC BWK MHP TSS SKC JKP BWC HKL. Wrote the paper: WP HK SC BWC.

References

- Weatherby J (1791) An Introduction to a General Stud Book. Weatherby and Sons, London.
- Poole D (2004) Current concepts of oxygen transport during exercise. *Equine and Comparative Exercise Physiology* 1: 5–22.
- Young L, Marlin D, Deaton C, Brown-Feltner H, Roberts C, et al. (2002) Heart size estimated by echocardiography correlates with maximal oxygen uptake. *Equine veterinary journal* 34: 467–471.
- Kayser S, Hoppeler H, Lindstedt S, Claassen H, Jones J, et al. (1989) Total muscle mitochondrial volume in relation to aerobic capacity of horses and steers. *Pflügers Archiv* 413: 343–347.
- Gunn H (1987) Muscle, bone and fat proportions and muscle distribution of Thoroughbreds and other horses.
- McGivney BA, McGettigan PA, Browne JA, Evans ACO, Fonseca RG, et al. (2010) Characterization of the equine skeletal muscle transcriptome identifies novel functional responses to exercise training. *BMC genomics* 11: 398.
- Andersson L (2012) How selective sweeps in domestic animals provide new insight into biological mechanisms. *Journal of internal medicine* 271: 1–14.
- Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, et al. (2005) The effects of artificial selection on the maize genome. *Science* 308: 1310–1314.
- Storz JF (2005) INVITED REVIEW: Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* 14: 671–688.
- Chen G, Li R, Shi L, Qi J, Hu P, et al. (2011) Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. *BMC genomics* 12: 590.
- STEVEN L, SALZBERG J (2005) Beware of mis-assembled genomes. *Bioinformatics* 21: 320–324.
- Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics* 27: 1068–1075.
- Birzele F, Schaub J, Rust W, Clemens C, Baum P, et al. (2010) Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic acids research* 38: 3999–4010.
- Burset M, Seledtsov I, Solovjev V (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research* 28: 4364–4375.
- Robertson G, Schein J, Chiu R, Corbett R, Field M, et al. (2010) De novo assembly and analysis of RNA-seq data. *Nature methods* 7: 909–912.
- Chen G, Yin KP, Wang C, Shi TL (2011) De novo transcriptome assembly of RNA-Seq reads with different strategies. *Science China Life Sciences* 54: 1129–1133.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29: 644–652.
- Birrol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, et al. (2009) De novo transcriptome assembly with ABySS. *Bioinformatics* 25: 2872–2877.
- Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28: 1086–1092.
- Martin J, Bruno VM, Fang Z, Meng X, Blow M, et al. (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC genomics* 11: 663.
- Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Research* 20: 1432–1440.
- Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966–1967.
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18: 821–829.
- Park KD, Park J, Ko J, Kim BC, Kim HS, et al. (2012) Whole transcriptome analyses of six thoroughbred horses before and after exercise using RNA-Seq. *BMC genomics* 13: 473.
- Henschel R, Nista PM, Lieber M, Haas BJ, Wu L-S, et al. (2012) Trinity RNA-Seq assembler performance optimization. *ACM*. pp. 45.
- Tao T (2006) Program Parameters for blastall.
- Kim H, Lee T, Park W, Lee JW, Kim J, et al. (2013) Peeling Back the Evolutionary Layers of Molecular Mechanisms Responsive to Exercise-Stress in the Skeletal Muscle of the Racing Horse. *DNA Research*.
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.
- Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20: 1297–1303.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 43: 491–498.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.
- Da Wei Huang BTS, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4: 44–57.
- Hosack DA, Dennis Jr G, Sherman BT, Lane HC, Lempicki RA (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol* 4: R70.
- Alterovitz G, Ramoni MF (2010) Knowledge based bioinformatics: Wiley Online Library.
- Livak KJ, Schmittgen TD (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method. *methods* 25: 402–408.
- Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26: 589–595.
- Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics* 85: 847–861.
- Wright S (1949) The genetical structure of populations. *Annals of eugenics* 15: 323–354.
- Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105: 767–779.
- Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* 10: 564–567.
- Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome research* 16: 702–712.
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913–918.
- Sabeti P, Schaffner S, Fry B, Lohmueller J, Varrilly P, et al. (2006) Positive natural selection in the human lineage. *science* 312: 1614–1620.
- Peng Z, Cheng Y, Tan BC-M, Kang L, Tian Z, et al. (2012) Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology* 30: 253–260.
- Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* 8: 469–477.
- Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*.
- Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *The plant journal* 51: 910–918.
- Bray NJ, Buckland PR, Owen MJ, O'Donovan MC (2003) Cis-acting variation in the expression of a high proportion of genes in human brain. *Human genetics* 113: 149–153.
- Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, et al. (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *science* 329: 643–648.
- Xia Z, Xu H, Zhai J, Li D, Luo H, et al. (2011) RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*. *Plant molecular biology* 77: 299–308.
- Kingston S, Hoffman-Goetz L (1996) Effect of environmental enrichment and housing density on immune system reactivity to acute exercise stress. *Physiology & behavior* 60: 145–150.
- Stefansson S, Yepes M, Gorlatova N, Day DE, Moore EG, et al. (2004) Mutants of plasminogen activator inhibitor-1 designed to inhibit neutrophil elastase and cathepsin G are more effective in vivo than their endogenous inhibitors. *Journal of Biological Chemistry* 279: 29981–29987.
- Niess A, Dickhuth H, Northoff H, Fehrenbach E (1999) Free radicals and oxidative stress in exercise-immunological aspects. *Exercise immunology review* 5: 22.
- Douset E, Avela J, Ishikawa M, Kallio J, Kuitunen S, et al. (2007) Bimodal recovery pattern in human skeletal muscle induced by exhaustive stretch-shortening cycle exercise. *Medicine and science in sports and exercise* 39: 453.
- Sun L, Ma K, Wang H, Xiao F, Gao Y, et al. (2007) JAK1–STAT1–STAT3, a key pathway promoting proliferation and preventing premature differentiation of myoblasts. *The Journal of cell biology* 179: 129–138.
- O'Shea JJ, Pesu M, Borie DC, Changelian PS (2004) A new modality for immunosuppression: targeting the JAK/STAT pathway. *Nature Reviews Drug Discovery* 3: 555–564.
- Chau Long Y, Widegren U, Zierath JR (2004) Exercise-induced mitogen-activated protein kinase signalling in skeletal muscle. *Proceedings of the Nutrition Society* 63: 227–232.
- Saleem A, Adhietty PJ, Hood DA (2009) Role of p53 in mitochondrial biogenesis and apoptosis in skeletal muscle. *Physiological genomics* 37: 58–66.
- Gourlay CW, Ayscough KR (2005) The actin cytoskeleton: a key regulator of apoptosis and ageing? *Nature reviews Molecular cell biology* 6: 583–589.
- D'Alessandro A, Righetti PG, Zolla L (2009) The red blood cell proteome and interactome: an update. *Journal of proteome research* 9: 144–163.

63. Cecchetti L, Tolley ND, Michetti N, Bury L, Weyrich AS, et al. (2011) Megakaryocytes differentially sort mRNAs for matrix metalloproteinases and their inhibitors into platelets: a mechanism for regulating synthetic events. *Blood* 118: 1903–1911.
64. Giorgi C, Yeo GW, Stone ME, Katz DB, Burge C, et al. (2007) The EJC factor eIF4AIII modulates synaptic strength and neuronal protein expression. *Cell* 130: 179–191.
65. Kestin AS, Ellis PA, Barnard MR, Errichetti A, Rosner BA, et al. (1993) Effect of strenuous exercise on platelet activation state and reactivity. *Circulation* 88: 1502–1511.
66. Yu X, Riley T, Levine AJ (2009) The regulation of the endosomal compartment by p53 the tumor suppressor gene. *FEBS Journal* 276: 2201–2212.
67. Sandri M, Lin J, Handschin C, Yang W, Arany ZP, et al. (2006) PGC-1 α protects skeletal muscle from atrophy by suppressing FoxO3 action and atrophy-specific gene transcription. *Proceedings of the National Academy of Sciences* 103: 16260–16265.
68. Zhao J, Brault JJ, Schild A, Cao P, Sandri M, et al. (2007) FoxO3 coordinately activates protein degradation by the autophagic/lysosomal and proteasomal pathways in atrophying muscle cells. *Cell metabolism* 6: 472–483.
69. Jamart C, Benoit N, Raymackers J-M, Kim HJ, Kim CK, et al. (2012) Autophagy-related and autophagy-regulatory genes are induced in human muscle after ultraendurance exercise. *European journal of applied physiology* 112: 3173–3177.