Department of Computer Science & Mathematics

Title: Genetics in Horse Racing

By: Nour El Khoury

Aram Papazian

Genetics in Horse Racing

Introduction:

Horse racing is one of the oldest sports in the world, but its basic concept has not changed. The horse that finishes the race first is the winner. In the early years of the sport, the race only involved two horses racing each other in either a short distance speed race or long-distance endurance race. However, it later evolved into a massive spectacle with large fields of runners, advanced tools to monitor the horses, and large sums of money were being thrown to bet on the winning horse. This all attracted a massive crowd which led to it becoming a world-renowned sport and horse races were being organized all over the world now. The horse breeds that are often involved and best suited for these types of races are Thoroughbreds. In fact, Thoroughbreds have been the topic of selection for four centuries now in horse racing. This all might be due to its physical and physiological prowess which gives them a massive advantage over other horse breeds. The main physical and physiological phenotype that makes thoroughbreds superior is its high maximal oxygen intake which is due to its remarkable oxygen-carrying capacity and delivery aided by structural and functional adaptations involving the respiratory and cardiovascular system. To illustrate, some of these adaptations involve high levels of maximum hemoglobin and cardiac output, a large lung volume, as well as a large muscle mass (approximately 55%) to body weight ratio, high skeletal muscle mitochondrial density and oxidative enzyme activity and large intramuscular stores of energy substrates (mainly glycogen) (Gu, et al., 2009).

Unlike the physical and physiological adaptations that make a Thoroughbred elite athletically, the genes of this breed that also contribute to enhancing athletic performance are not

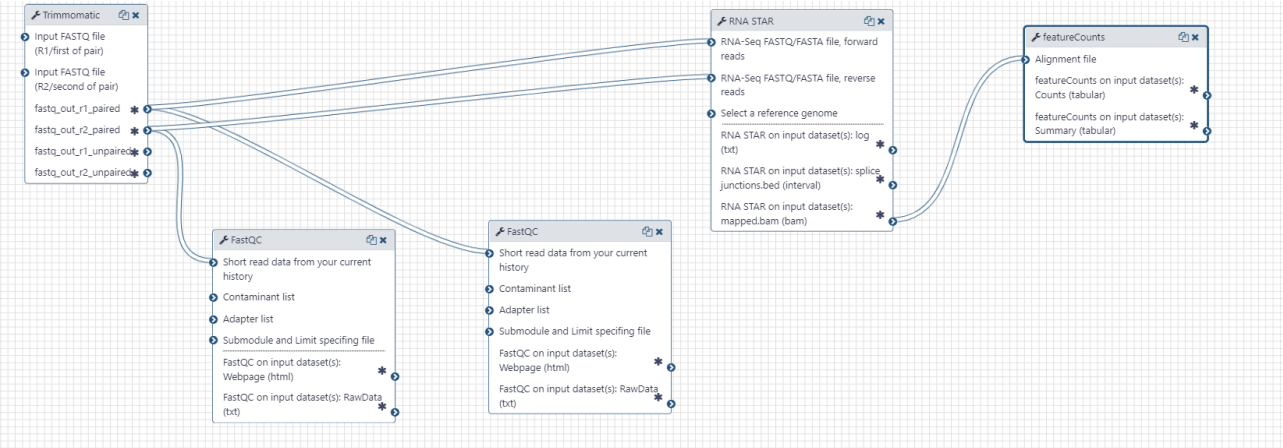
described and analyzed properly. Some companies offer genetic testing on horses, but with low accuracy results. Most of these companies offer the speed gene test, which is basically to see if the horse has the MSTN variant g.66493737C>T (rs397152648) which has been widely considered to be advantageous in the horse's racing ability (Hill, Mcgivney, Gu, Whiston, & Machugh, 2010). To illustrate, MSTN is known to be responsible for muscle growth in horses and this mutation in MSTN causes muscle hypertrophy which will allow the Thoroughbred to carry more muscle than normal Thoroughbreds. The horses that have a genotype homozygous C/C in MSTN are the most sought after horses when it comes to sprint racing (Hill E. , et al., 2010). This genotype has been proven to be the most successful during short distance races. The MSTN variant is one example of some of the genes that have been found to play a role in enhancing performance.

However, very few variants have been agreed upon to be directly related to horse racing. So we embarked upon studying this topic even further and find out which genes other than these may play a role in a horse's performance. To start with, we searched for articles that are directly related to our topic and collected all the genes that have been discussed in these articles. These genes found were shown to play a role in enhancing a horse's performance both physically and mentally. Then, a score was placed on each gene depending on the number of articles that showed the significance of that gene in a horse's performance. Also, enough information was collected on how these genes interact with each other and if there is a significant correlation between these genes. Furthermore, we also discovered how these genes are differentially expressed after specific training regimes. Separation of the genes that are differentially expressed after a single bout of high-intensity exercise from the ones that are differentially expressed after months of low-intensity training was done. After that, two datasets were taken that contained

samples of horses before and after exercise and analysis was done using the RNA seq protocol and EdgeRLimma Package to find out the different gene expressions that occurred post-exercise. Then, we formed a Venn-diagram to show the genes that have been expressed similarly between the two datasets and the genes that have been expressed differently between the two datasets. Also, we used the information that we collected to find out from these datasets which horses have the right genes with the favorable genotypes to be contenders to win the race. Finally, to find novel variants related to performance in horses, we performed an Enrichr analysis to check the KEGG pathways that these genes affect, took the gene that is directly involved in performance-related pathways and checked their log fold change and expression levels after exercise and then formed a Venn-diagram that compares all the variants that affect this specific gene in all the horses.

Methods:

Differential Expression Pipeline:



Galaxy Workflow for Trimming, Alignment, and Feature Counts

Data and Samples Used:

To study the differentially expressed genes in skeletal muscle, first we had to find two datasets that have horse samples before and after exercise. So by using ArrayExpress we were able to find the RNA seq data sets that we needed. The first data set contains skeletal muscle tissue samples of horses before exercise, 4 hours after exercise, and after 10 months of training. The second data contains skeletal muscle and blood tissue samples of horses before exercise and immediately after 30 min of intense exercise. Both of the links of the data will be available below. After that we decided to take 12 samples from each dataset (6 before exercise and 6 after exercise) to study the differentially expressed genes in both datasets and later compare their results

Preprocessing:

The paired-end reads were first uploaded to Galaxy. Then, Fastqc was applied to check the quality of these reads. All the reads needed trimming based on the Fastqc results. To perform paired-end trimming, the Trimmomatic tool on Galaxy was used with conditions to perform the initial Illumina adapter trimming and to trim the bases that had a per-base quality less than 20. We chose these conditions based on the Fastqc results that we got, and since these conditions were used for previous RNA seq studies. After that, Fastqc was performed again on the trimmed reads to check if any changing is still needed which was not the case for all the trimmed reads.

Alignment & Feature Counts:

First, the reference genome EquCab2.0 release 67 was uploaded to Galaxy and indexed directly (Done by Galaxy once uploaded). To align the reads to this reference genome the RNA

STAR tool on Galaxy was used. The mapped bam files were obtained directly from the RNA STAR tool since it has a built-in Samtools. After obtaining the mapped bam files it was time to perform Feature Counts which was also a tool on Galaxy. This tool is used to measure gene expression for each gene. However, before running it, the GTF (Gene transfer format) file from ensemble 67 was uploaded to Galaxy. This file contains information about gene structure and it is needed to run Feature Counts. Finally, Feature Counts was done to all the bam files, and results were downloaded locally.

EdgeR Limma Package:

To begin with, the feature counts results were read into Rstudio. After that, all the counts were joined together in a matrix that had column names that indicated the group they belong too (Before or After exercise), and row names with the gene stable ensemble IDs. After that, the data was scale normalized using the calcNormFactors function in the EdgeR package. The low expressed genes were filtered out. Then, the Voom transformation was applied to the normalized data using the Voom function in the EdgeR package. After that, the Empirical Bayes method was used to smoothen the standard errors and to calculate the P values. Finally, the genes that had an adjusted P value less than 0.05 were kept since they were considered significant, and the ones that had a fold change greater than or equal to 0.5 or less than or equal to -0.5 were also kept.

Variant Identification Pipeline:

Preprocessing:

The paired-end reads were first uploaded to Galaxy. Then, to check the quality of these reads Fastqc was used. All of the reads needed trimming based upon the Fastqc analysis that was

obtained. To perform paired-end trimming, the Trimmomatic tool on Galaxy was used with conditions to perform the initial Illumina adapter trimming and to trim the bases that had a per-base quality less than 20. We chose these conditions based upon the Fastqc results that we got and since these conditions were used for previous RNA seq studies.

Alignment:

First, the latest version of the reference genome EquCab3.0 was uploaded to Galaxy and indexed directly. To align the reads to this reference genome the TopHat tool on Galaxy was used. Then we directly obtained the mapped bam file from TopHat since it has a built-in Samtools.

Variant Calling:

The GATK package was used for variant identification. We begin by indexing the genome EquCab3.0 using samtools faidx and Creates a sequence dictionary for a reference sequence using the CreateSequenceDictionary tool. Following, the AddOrReplaceReadGroups tool enables the user to add a read group in the bam files (from tophat). Next, SortSam was used to sort the bam files by coordinates. It was also important to use MarkDuplicates to eliminate PCR duplicates and avoid over-representation. Afterward, HaplotypeCaller was applied to identify potential variation sites and assign possible genotype along with a likelihood score. Then, joint genotyping to generate a raw unfiltered vcf file was handled by GenotyGVCFs. After obtaining the vcf file containing all the potential variants, SelectVariants was used to identify and select SNPs only. Finally, VariantFiltration was used to filter the SNPs based on their

mapping quality coverage quality, mapping quality rank-sum, and other parameters detecting strand bias such as the Fischer score and read position rank-sum.

Enrichr Analysis:

After obtaining all the vcf files from the 12 samples (6 before exercise/6 after exercise), we uploaded those vcf files to the ensemble Variant Effect Predictor (VEP). We used the latest version of EquCab3.0 as our reference genome. After running the VEP, we took the gene stable ensemble IDs that we got and uploaded them to Biomart, to get the gene names. Then, we simply uploaded the gene names to Enrichr to better analyze which pathways to these variants affect the most.

SnpEff:

Also, to better understand the effect of these variants on the horses, we decided to analyze their effect using SnpEff tool. To do this, all the vcf files were first uploaded to Galaxy. After that, the SnpEff database EquCab2.86 was downloaded on Galaxy by using the SnpEff download database tool. Then, SnpEff eff tool was used with the downloaded database to annotate and predict the effect of each variant on a specific gene.

Results:

Differential Expression:

	logFC	AveExpr	t	P.Value	adj.P.Val
ENSECAG00000024623	3.1904602	6.140334	15.487885	8.367588e-11	1.052224e-06
ENSECAG00000016773	3.0806130	2.499810	11.491973	5.846691e-09	3.676107e-05
ENSECAG00000009799	2.4435146	4.943932	10.483341	2.068133e-08	6.292086e-05
ENSECAG00000012476	2.4418422	3.064237	10.466487	2.113932e-08	6.292086e-05
ENSECAG00000009402	1.3915101	6.770427	10.283435	2.686388e-08	6.292086e-05
ENSECAG00000021680	-1.7310924	3.267453	-10.199373	3.002188e-08	6.292086e-05
ENSECAG00000018816	-3.3881152	1.857410	-10.024291	3.792691e-08	6.813298e-05
ENSECAG00000013903	1.8456541	5.216520	9.844910	4.834370e-08	7.474411e-05
ENSECAG00000017011	2.9172670	8.191208	9.703190	5.869755e-08	7.474411e-05
ENSECAG00000004492	1.5124649	3.920156	9.625053	6.538479e-08	7.474411e-05

Figure 1: The top 10 differentially expressed (based on P-value) genes after exercise in Data 1.

	logFC	AveExpr	t	P.Value	adj.P.Val
ENSECAG00000000763	4.854148	4.385382	15.69863	4.859143e-11	6.651979e-07
ENSECAG00000024719	3.911435	4.730003	15.02220	9.336111e-11	6.651979e-07
ENSECAG00000021269	3.144957	5.640048	14.59482	1.429524e-10	6.790239e-07
ENSECAG00000014260	4.958241	7.361902	14.03262	2.546208e-10	6.970575e-07
ENSECAG00000022059	4.642770	5.533358	13.89964	2.927272e-10	6.970575e-07
ENSECAG00000021525	3.079742	7.173789	13.89714	2.934979e-10	6.970575e-07
ENSECAG00000022166	3.568485	6.181008	13.47989	4.580873e-10	8.371341e-07
ENSECAG00000017657	3.287006	4.464505	13.43217	4.823759e-10	8.371341e-07
ENSECAG00000009722	6.699675	4.594030	13.34779	5.287163e-10	8.371341e-07
ENSECAG00000003816	5.161633	6.704287	13.07362	7.147514e-10	1.018521e-06

Figure 2: the top 10 differentially expressed (based on P-value) genes after exercise in Data 2.

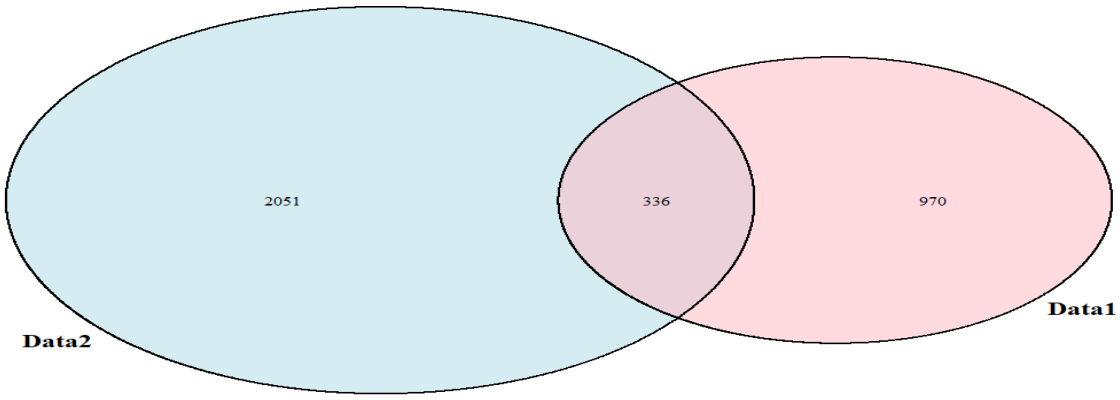
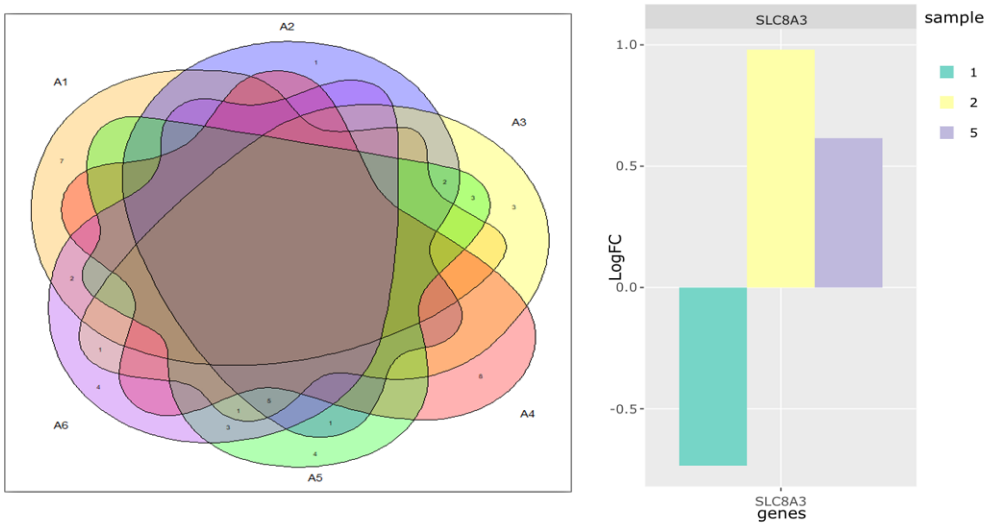


Figure 3: Venn-diagram comparing the differentially expressed genes between Data 1 and Data 2.

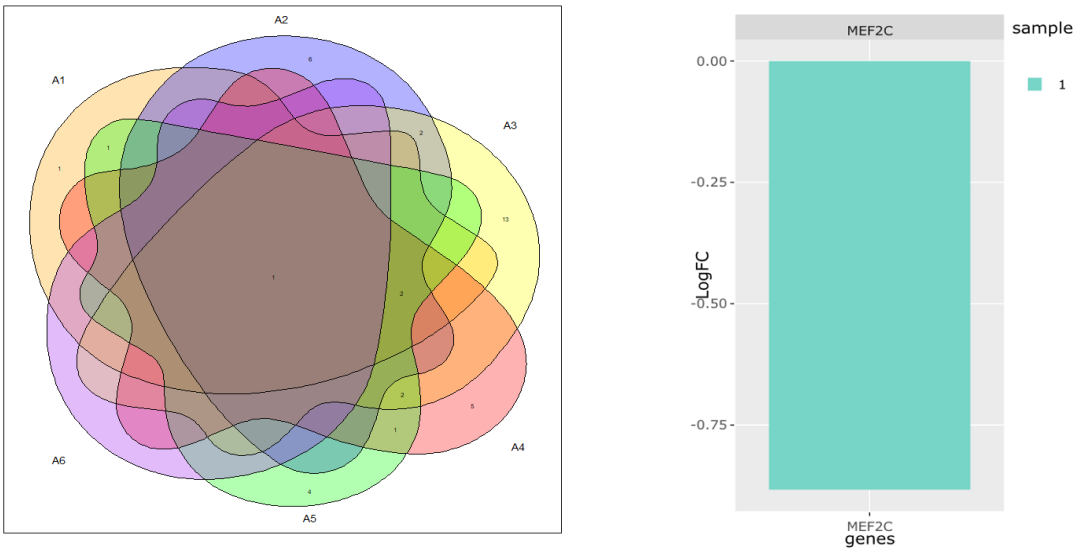
Novel Variants:

Figure 4: SLC8A3 Gene



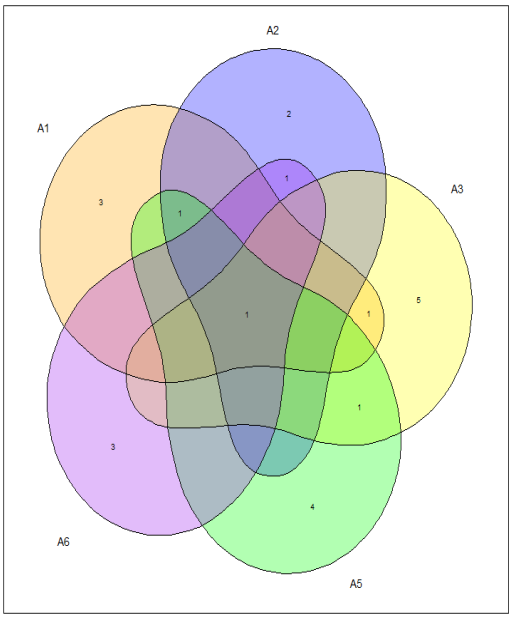
Horses 2 and 5 unique variant	Location
rs69361443	c.1785-6810C>T

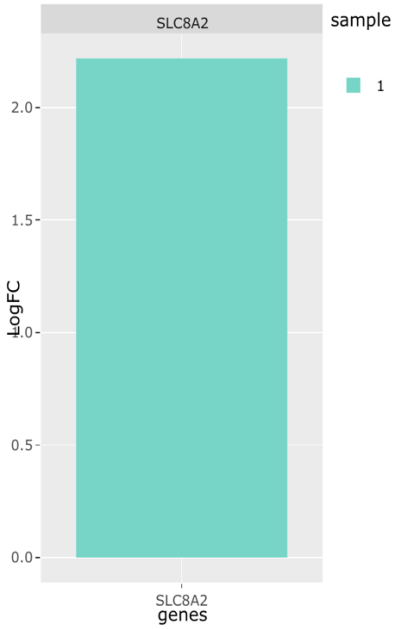
Figure 5: MEF2C Gene



Horse 1 unique variant	Location
“.”	n.78817493A>G

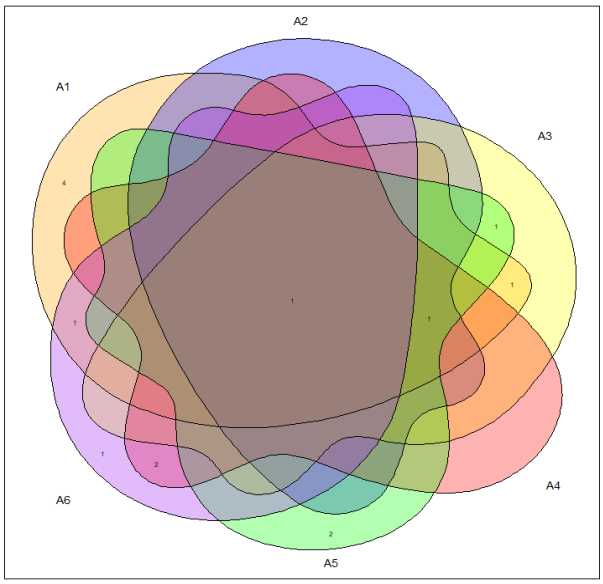
Figure 6: SLC8A2 Gene.

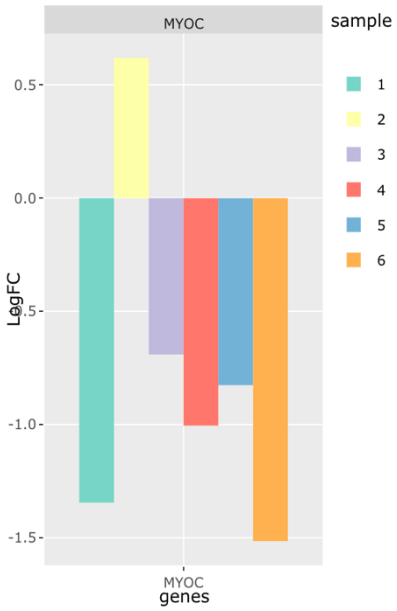




Horse1 unique variants	Location
rs1148975178	c.1618+90G>A
rs1147849609	c.1074+275C>T
rs1136481604	c.957-81A>C

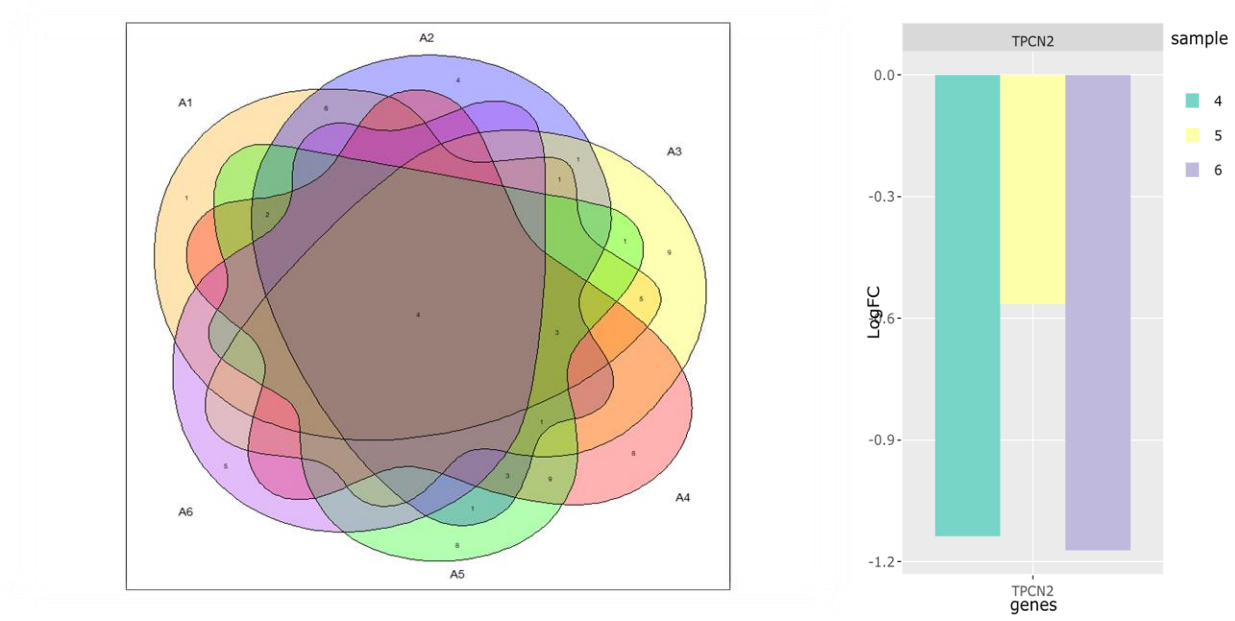
Figure 7: MYOC Gene.





Horses 1 and 6 unique variant	Location
rs782858080	n.7657993A>C

Figure 8: TPCN2 Gene



Horses 1,2, and 3 unique variant	Location
"."	n.28604047A>G

Discussion:

Differential Expression:

The Venn-diagram (Figure 3) is showing that there are 336 common differentially expressed genes between data 1 and data 2. However, Data 1 has 970 genes unique to it and Data 2 has 2051 unique genes. We can see that Data 2 has 1081 more differentially expressed genes than Data 1. This result was expected since in Data 2 the muscle tissue samples were taken directly after exercise, but in Data 1 the samples were taken 4 hours after exercise. To illustrate,

4 hours post-exercise is enough for the horse to rest and get back to homeostasis, which will lead to some of the genes getting back to normal expression levels as they were before exercise which leads to a decrease in the number of differentially expressed genes (Mcgivney, et al., 2010). Unlike the thoroughbreds in Data 1, the thoroughbreds in Data 2 did not have time to recover and their body still is considered to be under stress. So the expression of the immediate-early response genes will still be high. These genes are early regulators of cell growth and differentiation signals, which help the body to get back to homeostasis when under a wide variety of stress signals (McGivney, et al., 2009). An example of such a gene is FOS which was considered to be our fourth best differentially expressed gene in Data 2. FOS is just one of the many immediate-early response genes that we got and to further understand which pathways do these genes effect, we performed an Enrichr analysis. The results showed us the different KEGG pathways that these genes affect, which are: Apelin signaling pathway, spliceosome, HIF-1 signaling pathway, protein processing in the endoplasmic reticulum, and the insulin signaling pathway. These pathways perfectly match the variety of purposes of these genes which are energy metabolism, oxygen uptake, regulation of muscle contraction, regulation of transcription, and cell growth and development. Furthermore, these immediate-response genes have a response that is fast and transient, but some of these genes are also expressed in Data 1. This suggests that these genes may also contribute to long term adaptation which brings up the notion of the importance of training. If the horse is trained properly, these genes will not be differentially expressed when going through another bout of exercise since the horse's body will already be familiar and adapted with the stimulus that it is facing. Furthermore, these results also show how important exercise is and how it changes an organism not only physically, but also the

organism's genetic composition. Concerning the unique genes for Data 1, these genes have a delayed response to exercise and their purpose is also to restore homeostasis in the body.

Known Variants:

These are the variants that we have obtained from our thorough analysis of different articles related to horse racing and enhancing performance in horses. We took these variants and checked if any of the thoroughbreds in data set 2 had them. We only performed this on the vcf files of dataset 2 since we were provided with a relation file that indicated specifically which horse was which before and after exercise. We tested this on a total of six horses. The results we got were as follows: None of the six horses had the holy grail MSTN variant g.66493737C>T(rs397152648). However, horse 2 has the SLC16A1 variant g.55589063T>G(rs1149933209). This variant has been proven in several articles that it plays a role in enhancing performance (Musiał, Ropka-Molik, Piórkowska, Jaworska, & Stefaniuk-Szmukier, 2019). In fact, SLC6A1 is known to transport lactate acid and protons across the cell membrane which is important since lactate can act as a source of energy for working muscle especially during low levels of oxygen. Moreover, Horse 2 has also the T allele of this variant which is found to be very suitable for short-distance racing.

Besides, other than the SLC16A1 variant, the ACTN3 variant c.2334C > T(rs1141508235) was detected in both horse 1 and horse 2. Alpha-actinin-3 plays a main role in muscle contraction and cell metabolism not only in horses, but also in humans (Thomas, Hamilton, North, & Houweling, 2014). Also, it is considered as one of the main determinants of muscle strength in both humans and horses because of this it is one of the most

studied genes not only for horse racing but also for sports in general. Moreover, in our study both horses, one and two had the C allele of this variant which suits short distance racing. Furthermore, horses 1 and 2 that have the rs1141508235 show significantly higher expression than the other 4 horses. Also they are the only ones that have ACTN3 upregulated instead of downregulated after exercise. This finding was also supportive with previous studies which suggest such horses having this variant will have higher expressions of ACTN3 which will improve their ability in sprint races (Ropka-Molik K. , Stefaniuk-Szmukier, Musiał, Piórkowska, & Szmatoła, Sequence analysis and expression profiling of the equine ACTN3 gene during exercise in Arabian horses, 2019). So the higher the ACTN3 level the better the horse will be in sprints.

Another known SNP that was detected was the PDK4 variant g.38973231A>G (rs69586789). This was detected in horse 6 with the A allele of this variant which is advantageous to short distance racing. To elaborate, PDK4 is responsible for the oxidation of fatty acids which is highly efficient in the generation of ATP (Hill E. , Gu, McGivney, & MacHugh, 2010). This information shows us the importance of PDK4 towards performance since ATP is the main source of energy in horses.

Based on the SNPs detected, horse 2 is considered to be the best since it had both the SLC6A1 variant and the ACTN3 variant after that comes in horse 6 over horse 1 since the PDK4 variant was found in 8 articles, but the ACTN3 variant was found in 7 articles.

Novel Variants:

The variants that we are going to mention here below are going to be considered as novel since we did not find any references that consider these variants as a candidate in making a horse elite

in performance. We detected these variants by first looking at the results that were obtained from both the Enrichr analysis and the SnpEff and picking the ones that are related to pathways that play a role in athletic performance. After that, we looked at their log fold changes and expressions directly after high-intensity exercise in the six horses in dataset 2. Then we analyzed the SnpEff vcf file results and from that took all the variants from the six horses that affect a specific gene. After collecting the variants from the horses, we formed Venn-diagrams to compare the variants in the six horses for a specific gene. Also, if an unknown variant was detected(“.” Variant), we compared these based on their location and not the rs ID. So in total we analyzed 64 genes that we have found to be part of pathways that play a role in enhancing a horse’s athleticism after filtering, but out of the 64 genes we got five interesting findings that we believe may play a factor in improving a horse’s performance.

The first interesting finding that we discovered was with the SLC8A3 gene. As you can see in the bar graph (Figure 4), the expression levels and log fold change of SLC8A3 in horses 2 and 5 are the highest. Then by analyzing the Venn-diagram of the variants of this gene, it is clear that horses 2 and 5 have one similar variant between them that the others do not have, so this could be a reason behind the fold change and expression level of SLC8A3 in these horses standing out from the rest. The variant that is common between them uniquely is c.1785-6810C>T(rs69361443). We chose SLC8A3 as a candidate gene since it is known to play a role in the regulation of skeletal muscle contraction, synapse regulation, and also in learning and memory. So not only will it help the horse’s muscle contract faster, but it will form muscle memory to make the horse perform even better when put in the same situation next time. Furthermore, it may also make the horse easier to train since it improves training. So since horses 2 and 5 have higher expressions this may provide them with an advantage over the others when

racing. Unlike horses 2 and 5, horse 1 had SLC8A3 significantly downregulated after exercise which may be considered as a disadvantage to this horse.

The second variant that we have found that may play a role in performance is unknown in MEF2C (n.78817493A>G). When comparing the SNPs between the horses this variant was found to be unique in horse 1. Furthermore, as you can see in the bar graph (Figure 5) comparing the fold changes of MEF2C between all the six horses, horse 1 was the only one who had MEF2C significantly downregulated. This is considered to be a handicap to horse 1 since MEF2C is found to play a role in skeletal and cardiac muscle cell development and differentiation. So downregulation of this gene means that horse 1 will have a harder time growing skeletal muscle compared to the other horses. Also, this shows that horse 1 will not be able to pump blood in its body as fast as the others. This is a disadvantage since it means that it will not produce as much energy during exercise which will lower the level of mechanical work its body can handle.

To add to, another gene that we considered to play a role in athleticism was the SLC8A2. As you can see in the bar graph (Figure 6), horse 1 is the only horse that had this gene significantly upregulated, and by checking the results of the comparison between the variants of this gene we found that horse 1 had 3 unique variants (Table in Figure 6). By studying the GO annotations of SLC8A2, we found that this gene plays a role in learning and memory so this means that horse 1 will be considered more trainable than the other horses since it had higher expression of SLC8A2. Also, this will help in muscle memory which means that the horse's muscle will be well prepared next time when putting in the same position.

Furthermore, another gene that we discovered to have a positive impact on horse racing was the MYOC gene. As you can see from the bar graph (Figure 7), both horses 1 and 6 had

MYOC downregulated the most. By analyzing the SNPs, it was found that horses 1 and 6 had one variant in common unique to them(rs782858080) which may be the reason for MYOC being downregulated that much. This might give both horses a significant disadvantage over the others since MYOC was found to regulate skeletal muscle hypertrophy. This means that horses 1 and 6 will not have the ability to hold more muscle mass than the others since hypertrophy means the enlargement of the muscle organ. Another reason why this is disadvantageous is because MYOC also positively regulates stress fiber assembly which helps in faster reflexes when under stress. So both horses will react slower when engaged with a stimulus which may also deprive them of their chances of winning since race time differs by milliseconds so a slower reaction to the start signal may lead to a loss. Moreover, the bar graph also shows that horse 2 is the only one who had MYOC upregulated. This is a massive boost to its chances of winning the race.

Finally, the last gene that we noticed to provide an improvement to the horse's racing ability was the TPCN2 gene. The bar graph (Figure 8) shows us that horses 4,5, and 6 had TPCN2 downregulated. However, horses 1,2, and 3 had no significant fold change for this gene. When comparing the variants, we found that horses 1,2, and 3 had one SNP in common which was an unknown variant(n.28604047A>G). So we concluded that this SNP may be the reason why there was no fold change in these horses. Furthermore, TPCN2 is known to be part of smooth muscle contraction so horses 4, 5, and 6 who had it downregulated have a detriment in their chances of winning the race. To illustrate, smooth muscle contraction determines the speed of the muscle working when performing a mechanical movement, so downregulation of this gene will lead to slower contraction of the muscle in these horses which is a huge blow to their chances of winning.

Conclusion:

All in all, this study shows us how exercise changes not only an organism's phenotype but also the genetic composition. Also, how much of a factor can rest be for gene expressions as it was obvious when comparing the differentially expressed genes between both datasets. Moreover, concerning the novel variant that we found in SLC8A3, MEF2C, SLC8A2, MYOC, and TPCN2 further studies must be done in order to properly conclude that these variants play a role in enhancing performance. Finally, if you asked us which horse will win a race between the six thoroughbreds that we tested, based on the results that we got in the known and novel variants our money is on horse 2.

Data Links:

Data 1: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5447/?query=horse

Data 2: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-37870/?query=horse&page=2&pagesize=25

References

Bogossian, P. (2017, 10 12). Brief Review of Equine Genomics: Prospects toward Exercise and Sports Science. *MOJ Sports Medicine, 1*(5).

Bryan, K., McGivney, B., Farries, G., McGettigan, P., McGivney, C., Gough, K., . . . Hill, E. (2017, 8 9). Equine skeletal muscle adaptations to exercise and training: Evidence of differential regulation of autophagosomal and mitochondrial components. *BMC Genomics, 18*(1).

Costa, M., Davies, H., Anderson, G., & Slocombe, R. (2011, 7). Effects of two training protocols on Angiotensin I-converting enzyme (ACE) activity in horses. *Equine Veterinary Journal, 43*(4), 466-470.

Davie, A., Wen, L., Cust, A., Beavers, R., Fyfe, T., & Zhou, S. (2017, 7 15). The effects of moderate intensity training in a hypoxic environment on transcriptional responses in Thoroughbred horses. *Biology Open, 6*(7), 1035-1040.

De Mello Costa, M., Anderson, G., Davies, H., El-Hage, C., & Slocombe, R. (2010, 11). Circulating angiotensin converting enzyme in endurance horses: Effect of exercise on blood levels and its value in predicting performance. *Equine Veterinary Journal, 42*(SUPPL. 38), 152-154.

EATON, M., ROSE, R., EVANS, D., & HODGSON, D. (1995). Assessment of anaerobic capacity using maximal accumulated oxygen deficit in fit Thoroughbreds. *Equine Veterinary Journal, 27*(18 S), 29-32.

Effect-of-Training-on-PPARGC1A-and-FNDC5-Gene-Expression-in-Thoroughbred-Horses. (n.d.).

Eivers, S., Mcgivney, B., Fonseca, R., Machugh, D., Menson, K., Park, S., . . . Hill, E. (2010). Alterations in oxidative gene expression in equine skeletal muscle following exercise and training. *Physiol Genomics, 40*, 83-93.

Farries, G., McGettigan, P., Gough, K., McGivney, B., MacHugh, D., Katz, L., & Hill, E. (2018, 6 1). Genetic contributions to precocity traits in racing Thoroughbreds. *Animal Genetics, 49*(3), 193-204.

Filonzi, L., Franchini, N., Vaghi, M., Chiesa, S., & Marzano, F. (2015, 9 7). The potential role of myostatin and neurotransmission genes in elite sport performances. *Journal of Biosciences, 40*(3), 531-537.

Gomes Ferreira PADILHA, F., Balbi EL-JAICK, K., de CASTRO, L., Dos Santos MOREIRA, A., & Maria Reis FERREIRA, A. (n.d.). *Effect of selection for eventing on the MSTN gene in Brazilian sport horses.*

Gu, J., Orr, N., Park, S., Katz, L., Sulimova, G., MacHugh, D., & Hill, E. (2009, 6 2). A genome scan for positive selection in thoroughbred horses. *PLoS ONE, 4*(6).

Hamilton, N., Tammen, I., & Raadsma, H. (2013, 2 8). Multi-Species Comparative Analysis of the Equine ACE Gene Identifies a Highly Conserved Potential Transcription Factor Binding Site in Intron 16. *PLoS ONE, 8*(2).

Harrison, S., & Turrion-Gomez, J. (2006, 4). Mitochondrial DNA: An important female contribution to thoroughbred racehorse performance. *Mitochondrion, 6*(2), 53-66.

Hill, E., Gu, J., Eivers, S., Fonseca, R., McGivney, B., Govindarajan, P., . . . MacHugh, D. (2010, 1 20). A sequence polymorphism in MSTN predicts sprinting ability and racing stamina in thoroughbred horses. *PLoS ONE, 5*(1).

Hill, E., Gu, J., McGivney, B., & MacHugh, D. (2010). Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance. *Animal Genetics*, *41*, pp. 56-63.

Hill, E., Mcgivney, B., Gu, J., Whiston, R., & Machugh, D. (2010). *A genome-wide SNP-association study confirms a sequence variant (g.66493737C>T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses.*

McGivney, B., Eivers, S., MacHugh, D., MacLeod, J., O'Gorman, G., Park, S., . . . Hill, E. (2009, 12 30). Transcriptional adaptations following exercise in Thoroughbred horse skeletal muscle highlights molecular mechanisms that lead to muscle hypertrophy. *BMC Genomics, 10*.

McGivney, B., Hernandez, B., Katz, L., MacHugh, D., McGovern, S., Parnell, A., . . . Hill, E. (2019, 8 1). A genomic prediction model for racecourse starts in the Thoroughbred horse. *Animal Genetics, 50*(4), 347-357.

Mcgivney, B., Mcgettigan, P., Browne, J., Co Evans, A., Fonseca, R., Loftus, B., . . . Hill, E. (2010). *Characterization of the equine skeletal muscle transcriptome identifies novel functional responses to exercise training.*

Miyata, H., Itoh, R., Sato, F., Takebe, N., Hada, T., & Tozaki, T. (2018, 9 1). Effect of Myostatin SNP on muscle fiber properties in male Thoroughbred horses during training period. *Journal of Physiological Sciences, 68*(5), 639-646.

Musiał, A., Ropka-Molik, K., Piórkowska, K., Jaworska, J., & Stefaniuk-Szmukier, M. (2019, 12 1). ACTN3 genotype distribution across horses representing different utility types and breeds. *Molecular Biology Reports, 46*(6), 5795-5803.

Petersen, J., Mickelson, J., Rendahl, A., Valberg, S., Andersson, L., Axelsson, J., . . . McCue, M. (2013, 1). Genome-Wide Analysis Reveals Selection for Important Traits in Domestic Horse Breeds. *PLoS Genetics, 9*(1).

Rooney, M., Hill, E., Kelly, V., & Porter, R. (2018, 10 1). The "speed gene" effect of myostatin arises in Thoroughbred horses due to a promoter proximal SINE insertion. *PLoS ONE, 13*(10).

Ropka-Molik, K., Stefaniuk-Szmukier, M., Musiał, A., & Velie, B. (2019). The Genetics of Racing Performance in Arabian Horses. *International Journal of Genomics, 2019*. Hindawi Limited.

Ropka-Molik, K., Stefaniuk-Szmukier, M., Musiał, A., Piórkowska, K., & Szmatoła, T. (2019, 2 15). Sequence analysis and expression profiling of the equine ACTN3 gene during exercise in Arabian horses. *Gene, 685*, 149-155.

Ropka-Molik, K., Stefaniuk-Szmukier, M., Szmatoła, T., Piórkowska, K., & Bugno-Poniewierska, M. (2019, 9 11). The use of the SLC16A1 gene as a potential marker to predict race performance in Arabian horses. *BMC Genetics, 20*(1).

Ropka-Molik, K., Stefaniuk-Szmukier, M., Zukowski, K., Piórkowska, K., Gurgul, A., & Bugno-Poniewierska, M. (2017, 4 5). Transcriptome profiling of Arabian horse blood during training regimens. *BMC Genetics, 18*(1).

Selvaraj, P. (n.d.). Predicting The Outcome Of The Horse Race Using Data Mining Technique MSc Research Project Data Analytics.

Stefaniuk-Szmukier, M., Szmatoła, T., Łatka, J., Długosz, B., & Ropka-Molik, K. (2019, 8 1). The blood and muscle expression pattern of the equine TCAP gene during the race track training of arabian horses. *Animals, 9*(8).

Thomas, K., Hamilton, N., North, K., & Houweling, P. (2014, 3 15). Sequence analysis of the equine ACTN3 gene in Australian horse breeds. *Gene, 538*(1), 88-93.

Tozaki, T., Hill, E., Hirota, K., Kakoi, H., Gawahara, H., Miyake, T., . . . Kurosawa, M. (2012, 2). A cohort study of racing performance in Japanese Thoroughbred racehorses using genome information on ECA18. *Animal Genetics, 43*(1), 42-52.