Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative genomic and transcriptomic analyses identify functionally relevant causal genes for exercise in a large animal model

Mian Feng
University College Dublin
Thomas J. Hall
University College Dublin
John F. O'Grady
University College Dublin
Kseniia Maksimova
University College Dublin
Michael Salter-Townshend
University College Dublin
John A. Browne
University College Dublin
David E. MacHugh
University College Dublin
Lisa M. Katz
University College Dublin
Beatrice A. McGivney
Zinto Labs Ltd
Emmeline W. Hill
emmeline.hill@ucd.ie

University College Dublin

Research Article

Keywords: Thoroughbred, Exercise and training, reQTL, TWAS, Colocalisation, SMR

Posted Date: July 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9712807/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Thoroughbred horses have been strongly selected for exercise adaptations for competitive racing. Previous genome-wide association studies (GWAS) and expression quantitative trait locus (eQTL) analyses have examined the genomic architecture underlying variation in athletic performance and the transcriptional response to exercise in skeletal muscle in Thoroughbreds. However, no study has simultaneously leveraged a transcriptome-wide association study (TWAS), colocalisation analyses, and Summary-data-based Mendelian Randomisation (SMR) framework to identify causal mechanisms underpinning phenotypic variation for athletic performance traits.

Results

Here, we integrate RNA-seq data generated from Thoroughbred skeletal muscle sampled across three exercise states (untrained resting, untrained post-exercise, trained resting), and GWAS results for racing distance and speed traits, and use TWAS, colocalisation and SMR methodologies to characterise functionally relevant causal genes and variants for exercise traits. Exercise and training response eQTLs (reQTLs) were enriched in metabolic and stress response processes and showed distinct genomic localisation. Genes associated with racing distance (*MSTN* and *ORMDL1*) and speed (*ATAT1, DHDH, GPT2*, and *GYS1*) were identified.

Conclusion

This study provides important new omics insights into the distinction between inherited and acquired molecular mechanisms that underpin the adaptive exercise and training responses of skeletal muscle and demonstrates the Thoroughbred horse's importance as an animal model for exercise.

Introduction

Thoroughbred horses are a unique and valuable domestic animal resource for exercise genomics research, as they have been strongly selected for athleticism traits [1–3]. Originating in 17th century England, Thoroughbred horses are now globally widespread due to demand for horses with exceptional racing performance [4]. Following 300 years of intense selection, Thoroughbred horses possess a range of anatomical, metabolic, and physiological adaptations that enhance speed and stamina traits for athletic performance [5, 6].

Several studies have demonstrated the significant contribution of genomic variation to horse racing performance using genome-wide association study (GWAS) approaches [7–9]. From these and other studies, an intronic variant tagging an ERE1 short interspersed nuclear element (SINE) insertion

polymorphism at the myostatin gene (*MSTN*) locus, has been established as a key predictor of best race distance and speed in horses [7, 8, 10]. Moreover, several single-nucleotide polymorphisms (SNPs) within or in linkage disequilibrium (LD) with *MSTN* are strongly associated with best race distance [11, 12]. Functional studies have demonstrated that the ERE1 SINE insertion disrupts the transcriptional start site (TSS) and is likely to be the functional variant that produces the "speed gene" effect [10]. Most SNPs for complex traits identified through GWAS are located in noncoding regions and likely influence phenotypic variation by altering gene expression via sequence changes in genomic regulatory elements (GREs) [13]. Consequently, multi-omic integrative analyses—such as expression quantitative trait locus (eQTL) mapping, which combines matched transcriptomic and genomic data—represent an important approach in the identification of functionally relevant GREs [14–16]. Expression QTLs are defined based on the distance from the TSS; variants proximal to the associated gene (generally ≤ 1 Mb from the TSS) are characterised as *cis*-eQTLs, whereas variants located distally (generally > 1 Mb from the TSS) to the associated gene or on a different chromosome are considered *trans*-eQTLs [17, 18]. Previous work in the Thoroughbred has mapped *cis*- and *trans*-eQTLs in skeletal muscle, both at rest and post-exercise, and found that genes with significant eQTLs were predominantly related to metabolic function [6]. However, whether the effect size of such eQTLs varies across different exercise states (i.e., rest, exercise, training) has not yet been examined. Characterising eQTLs with substantial differences in effect size estimates on the expression levels of genes across biological states using matched samples, known as response eQTL (reQTL) mapping, represents an increasingly popular approach for both *in vivo* and *in vitro* studies examining variation in the interindividual response to diet and pathogen challenge [19–21]. However, this methodology has not been used to identify eQTLs with distinct effects across different exercise states in any mammalian species.

Recent advances in multi-omics approaches, including transcriptome-wide association study (TWAS) [16], colocalisation [22, 23], and Mendelian randomisation (MR) [14, 24] approaches, enable the integration of gene expression, genomic variation, and GWAS data for a trait of interest. These methodologies have not previously been applied to identify GREs and variants associated with athletic performance in the Thoroughbred. Applying such integrative omics approaches will enable more precise evaluation of how specific genetic variants influence exercise performance through modulation of gene expression at key genes [25].

The primary objective of this study, therefore, was to integrate genomic, transcriptomic, and GWAS data to characterise functional genomic variation for Thoroughbred horse racing performance traits. First, significant *cis*-eQTLs were identified in the skeletal muscle of Thoroughbred horses across three different exercise / training states. Response eQTLs were then mapped by comparing the effect sizes of inferred eQTLs across the three groups. To test the effects of trait-associated variants on gene expression, eQTLs were integrated with multiple GWAS results using TWAS, colocalisation, and MR analysis methodologies. The well-described race distance phenotype and knowledge of a functional variant at the *MSTN* locus was used to demonstrate the validity of the approach, which was then applied to a measured speed trait.

Materials and methods

Animal cohorts

The study subjects were a group of Thoroughbred horses trained for Flat racing at a single training establishment in Ireland. All horses were supervised by a single trainer and maintained under uniform management and feeding protocols. A total of 141 Thoroughbred horses (81 females, 60 males) were included in the study, with ages ranging from 1.4 years to 3.7 years.

The Thoroughbred horses were assigned to one or more of three distinct cohorts, corresponding to different exercise / training states: untrained at rest (UR, *n* = 94), untrained four hours post-exercise (UE, *n* = 71), and at rest following six months of training (TR, *n* = 65). A total of 50 horses were represented in both the UR and UE groups, 32 horses were common between the UR and TR groups, and 25 horses overlapped between the UE and TR groups, with 18 horses sampled at all three time points. Horses were categorised into untrained and trained cohorts according to their prior exercise workload, as reported previously [26]. The untrained horses had participated in fewer than five work days (WD; high-intensity exercise) prior to sampling, whereas the trained horses had an average of 17.1 WDs prior to sampling (standard deviation (SD) = 8.2).

Exercise and training protocols

The standard exercise protocol was as follows: horses underwent submaximal exercise conditioning (training) at canter six days per week, with WD exercises gradually introduced as part of their training regimen before transitioning to competitive racing. Based on each horse's soundness, fitness, and aptitude, a single trainer determined the specific training required. Each WD session began with 30–60 min of walking on an automated horse walker, followed by 5–10 min of hand-led walking. Horses then underwent a warm-up consisting of 300 metres of walking, 700 metres of trotting and slow cantering under saddle. Following the warm-up, horses performed high-intensity gallops over distances of 500–1,000 metres, with the final 800 metres set on a 2.7% incline on a 1,500-metre all-weather woodchip gallop track. In the same training facility, the WD data for a larger cohort of horses (*n* = 294) were recorded utilising concurrent global positioning system (GPS) and heart rate monitoring [6].

Biopsy, RNA extraction and quality control

Approximately 300 mg of skeletal muscle tissue was collected percutaneously from the ventral section of the middle gluteal muscle, following the methodology established by Valette *et al.* [27]. For both the UR group and the TR group, biopsies were taken between 7:30 AM and 11:30 AM prior to daily exercise. In the UE group, samples were collected four hours after the completion of WD exercise, a time point identified in prior research as optimal for detecting the most significant shifts in gene expression triggered by acute exercise [28, 29]. After collection, samples were initially preserved in RNAlater

solution (Thermo Fisher, Massachusetts, USA) for 24h at 4°C to stabilise RNA and subsequently transferred to − 20°C for long-term storage prior to RNA extraction.

Total RNA was extracted following a combined protocol that used TRIzol reagent (Thermo Fisher Scientific), DNase treatment (Qiagen, Hilden, Germany), and the RNeasy Mini Kit (Qiagen). The RNA concentration was determined using a NanoDrop ND1000 spectrophotometer (v3.5.27, Thermo Fisher Scientific). To evaluate RNA quality, the RNA integrity number (RIN) was measured using an Agilent Bioanalyzer in conjunction with the RNA 6000 Nano Kit (Agilent Technologies, Cork, Ireland).

RNA sequencing was conducted at the Research Technology Support Facility Michigan State University. Strand-specific, indexed Illumina sequencing libraries were generated using the TruSeq Stranded mRNA Library Preparation Kit LT (Illumina, San Diego, CA, USA). 18–20 indexed libraries were pooled per batch, and sequencing work was performed on an Illumina HiSeq 2500 platform utilising Rapid Run flow cells and reagents (Illumina). Each batch procedure included ten indexed libraries pooled. A dual-lane loading strategy was employed, allowing each pooled library to be distributed across both lanes of a single flow cell. Paired-end sequencing (2 × 100 bp reads) was performed.

RNA-seq data workflow

Quality control of the sequence data was performed using Trim Galore [v0.6.10] [30], which incorporated the FastQC (v0.12.1) [31] and Cutadapt (v1.18) [32] tools. FastQC was employed to assess the quality of the sequence data, while Cutadapt was used to trim adapter sequences. The STAR aligner (v2.7.10b) [33] was used for mapping the reads to the Equine reference genome, EquCab3 [34]. Transcript quantification for each sample was carried out using featureCounts (v2.2.1) [35].

Genotype data processing and imputation

Genomic DNA was extracted from blood samples collected from *n* = 3,231 Thoroughbred horses. Genotyping was performed using three platforms: the Illumina EquineSNP50 BeadChip (SNP50K, *n* = 320), the Illumina EquineSNP70 BeadChip (SNP70K, *n* = 2,167), and the Affymetrix Axiom™ Equine 670 SNP genotyping array (SNP670K, *n* = 743). Genome coordinates from all platforms were remapped from the EquCab2 to the EquCab3 reference genome using liftOver (v1.32.0) [36]. Quality control procedures were applied to the remapped genotype data using PLINK (v1.9) [37], ensuring that only SNPs with a missing genotype rate < 0.1 were retained. Genotypes from the SNP50K and SNP70K arrays were first imputed up to 488K SNPs based on the SNP670K genotype data with BEAGLE (v5.4) [1, 38, 39]. Subsequently, these imputed genotypes were phased and imputed to genome-wide resolution using a global reference panel comprising 11,442,356 SNPs from *n* = 125 horses spanning a broad geographic range (Supplementary Note 1). The imputation was also performed using BEAGLE with default parameters. Only variants with minor allele frequency (MAF) < 0.05 and dosage *R*-squared (DR^2) ≥ 0.8

were retained. A final set of 7,509,871 autosomal variants was retained for eQTL mapping and GWAS analysis.

Genome-wide association study

Using the complete set of imputed SNPs, we re-performed—at much higher resolution—GWAS analyses for race distance [40] and speed traits [41] in Thoroughbred horses. Prior to these analyses, the imputed variants were filtered within each GWAS cohort to exclude markers with a genotype missingness rate > 0.1 or a minor allele frequency (MAF) < 0.05.

For the GWAS for race distance traits, four phenotypes were examined: average race distance (AVD, n = 2,922), best race distance among all horses (BRD-ALL, n = 2,313), best race distance among elite horses (BRD-ELI, n = 820), and best race distance among non-elite horses (BRD-NON, n = 1,493). As previously described, horses were categorised as “elite” if they had won or were placed in at least one Stakes/Group/Listed race[40]. After quality control, 7,487,513, 7,479,207, 7,433,888, and 7,463,806 SNPs were retained for the AVD, BRD-ALL, BRD-ELI, and BRD-NON analyses, respectively. GWAS analyses were conducted using a mixed linear model implemented in GCTA (v1.94.1) [42]. Month of birth, sex, region, and the first five principal components (PCs) were included as covariates to control for potential confounding. Statistical significance was determined using both the Bonferroni (BF) and Benjamini-Hochberg (BH) correction methods, with adjusted P-values ($P_{adj.}$) < 0.05 considered significant. Because of the high LD exhibited in the Thoroughbred horse genome, LD clumping was used to obtain the significant independent GWAS signal sets using the PLINK parameter *–clump-r*^{2} 0.8 and *–clump-kb* 500.

For the GWAS for speed traits, the objective was to identify genetic variants associated with speed measured at two years of age (Speed2Y) in Thoroughbred horses (n = 122) [41]. Following quality control procedures (as above), a total of 7,026,748 SNPs remained for the Speed2Y GWAS analysis. A mixed linear model was implemented in GCTA, with sex and the first five PCs included as covariates. For this analysis, a relaxed significance threshold ($P < 5 \times 10^{-5}$) was applied to detect suggestive association signals for speed traits.

Conditional GWAS

To identify secondary association signals, a conditional GWAS analysis was performed using both summary- and individual-based data. The summary-based GWAS conditional analysis was performed using the *COJO* algorithm from GCTA [43], using the top SNPs (based on statistical significance) for each trait as the conditional variant. The individual-based GWAS conditional analysis used the top SNP as an additional covariate to re-run the GWAS.

Expression QTL (eQTL) mapping

Raw read counts were initially normalised using the Trimmed Mean of M-values (TMM) method with the edgeR (v3.36.0) package [44]. Following normalisation, the counts were subjected to inverse-normal transformation across samples. Genes with a transcript per million (TPM) normalised count > 0.1 in at least 20% of the samples were selected for eQTL mapping. To account for the unknown confounders in the gene expression data, the PCAForQTL R package (v0.1.0) [45] was used. The optimal number of PCs was determined using the *runElbow* or *runBE* function in PCAForQTL. Highly correlated known covariates, well captured by the inferred PCs (unadjusted $R^2 \geq 0.9$), were removed to avoid redundancy using the *filterKnownCovariates* function as part of the PCAForQTL R package. The remaining covariates, both known and inferred, were then combined. Known covariates included age, sex, sequencing batch, and the first two genotype PCs.

For *cis*-eQTL mapping, the *cis* window of each gene was set to ± 1 Mb from the transcriptional start site (TSS) [46, 47]. To identify significant *cis*-eQTLs, an adaptive permutation method [48] was applied to generate empirical *P*-values for all *cis* variants associated with each gene with the parameter --mode *cis* of tensorQTL (v1.0.9) [49], especially for lead variants with the most significant permutation *P*-value. Following this, a global BH false discovery rate (FDR) correction method [50] was applied to adjust the permutation *P*-values of these lead variants across all genes. A gene with an FDR-$P_{adj.} < 0.05$ was regarded as a *cis*-eGene, while a gene for the lead variants with FDR-$P_{adj.} \geq 0.05$ was regarded as a non-*cis*-eGene.

To identify all significant *cis*-eVariants associated with *cis*-eGenes, we calculated the nominal association statistics for all *cis*-eVariant-gene pairs with the parameter --model *cis_nominal* in tensorQTL. From this analysis, the empirical *P*-value of a gene closest to an FDR of 0.05 was defined as the genome-wide empirical *P*-value threshold (*pt*). Subsequently, the gene-level threshold for each gene was determined using the beta distribution, calculated with the function *qbeta* (*pt*, *beta_shape1*, *beta_shape2*) in R. The parameters of *beta_shape1* and *beta_shape2* were derived from the result of --mode *cis*. *cis*-eVariants with nominal *P*-value less than the gene-level threshold for the corresponding gene were considered as significant *cis*-eVariants.

To identify conditionally independent *cis*-eQTLs for each gene, the stepwise regression analysis from tensorQTL was performed with the parameter --mode *cis_independent*. Firstly, the gene-level significance threshold from the beta distribution was considered as the threshold to identify the significant *cis*-eQTLs as described above. By adding new *cis*-eQTLs and removing those without significance until no additional independent signals were found, the stepwise regression approach scanned all the *cis*-eQTLs.

Exercise response and training response eQTL (reQTLs)

To compare the effect size differences of eQTLs between different exercise states, and to identify the exercise reQTLs and training reQTLs, the multivariate adaptive shrinkage (mash) method was employed using the mashr package (v0.2.79) [51] in R. To ensure accuracy and avoid noise from LD, the reQTL analysis was restricted to paired samples. Two comparison groups were included: UE *vs*. UR to identify

response to exercise reQTLs, and TR *vs*. UR to identify response to training reQTLs. The eQTLs were reanalysed for the paired samples in the UE *vs*. UR (*n* = 50) and TR *vs*. UR (*n* = 32) comparison groups.

The summary statistics for all eVariant-gene pairs were obtained using tensorQTL with the parameter --mode *cis-nominal*. The summary statistics data were then used to generate a “strong” test and a “random” test. The “strong" test consisted of selecting the strongest gene-SNP pair for each expressed gene (i.e., the most significant eVariant for a gene) with the largest univariate test statistic in each comparison group and was used to learn data-driven covariance matrices. The “random” test consisted of a random subset of gene-SNP pairs, which was used to learn correlation structure and fit the mash mixture model.

After obtaining the estimated effect sizes (posterior mean) and the corresponding significance levels (local false sign rate, LFSR) from mashr, the top SNPs with LFSR < 0.05 were defined as significant eQTLs. To explore differences in the top eQTLs’ effect between exercise states, the effect sizes for the top SNPs, which were significant eQTLs in at least one state, were assessed. Exercise reQTLs were defined as the top SNPs that were significant eQTLs in at least one state (either UR or UE) and exhibited a fold change in effect size (posterior mean) between two states greater than 2.0. An exercise reGene referred to a gene containing at least one exercise reQTL.

Similarly, training reQTLs were defined as the top SNPs that were the significant eQTLs in at least one state (either UR or TR) and demonstrated a fold change in effect size between these conditions greater than 2.0. A training reGene was a gene with at least one training reQTL. In contrast, non-exercise reQTLs were defined as the top SNPs that were the significant eQTLs in at least one state (either UR or UE) but exhibited a fold change in effect size between states of less than 2.0. Non-exercise reGenes were genes without any exercise reQTLs. Non-training reQTLs were defined as the top SNPs that were the significant eQTLs in at least one state (either UR or TR) but exhibited a fold change in effect size between states of less than 2.0. Non-training reGenes were considered as genes without training reQTLs.

Functional enrichment analysis

To identify the functional roles of exercise and training reGenes, Gene set enrichment analysis (GSEA) was employed to identify significantly enriched Gene Ontology (GO) [52, 53] terms using the gProfiler2 (v0.2.3) R package [54]. The background gene set comprised all expressed genes (*n* = 16,709) after expression value filtering from eQTL mapping. For transcription factor (TF) enrichment analysis, exercise and training reGenes with Equine Ensembl IDs were first mapped to their annotated human orthologs by the ortholog search function (*Gorth*) of gProfiler2, based on the Ensembl database [55]. These mapped genes were then selected to perform the TF enrichment analysis using the *GOST* function of gProfilter2. The human orthologs (*n* = 13,950) of all expressed genes were selected as the background gene set. Additionally, to compare the proportion of TF enriched genes between exercise/training reGenes and non-exercise/non-training reGenes, they were cross-referenced with a list of human TFs [56], which includes 1,639 TFs. The difference in TF gene proportions between exercise/training reGenes and non-exercise/non-training reGenes was assessed using a two-sample proportion test [57].

Transcriptome-wide association studies

To identify the genes with expression patterns associated with racing performance traits, a transcriptome-wide association study (TWAS) was performed using the FUSION software [16]. The genomic relationship matrix was constructed using SNPs located within 1 Mb upstream and downstream of each gene's TSS. The covariates employed in this analysis were consistent with those used in the *cis*-eQTL mapping analysis. Initially, only genes identified as eGenes in the eQTL mapping analysis with a global BH FDR threshold of < 0.05 were included. For each of these genes, the heritability of *cis*-SNPs was estimated using the restricted maximum likelihood (REML) algorithm implemented in GCTA [42]. Genes with *cis*-heritability estimates significantly different from zero, determined by a likelihood ratio test (LRT) and nominal *P*-value threshold < 0.1, were considered as significantly heritable and retained for subsequent association tests. To construct expression prediction models, three statistical methods were employed: Top eQTL (top) based on standard marginal eQTL *Z*-scores, least absolute shrinkage and selection operator (LASSO) regression, and elastic-net (ENET) regression. The most significant model, determined through cross-validation, was utilised in the gene–trait association test. Using the BF correction method to adjust the TWAS *P*-value, the adjusted *P*-value threshold was < 0.05 for significant TWAS signals.

To investigate the independence of TWAS features and reduce the influence of LD, a conditional test was performed using the *FUSION.post_process.R* script in the FUSION package. Contiguous loci were set to 1 Mb, and the horse genotypes from all GWAS cohorts were used as an LD reference.

Colocalisation and Mendelian randomisation (MR) of eQTL and GWAS

To identify shared genetic variants between gene expression phenotypes and GWAS traits, we performed colocalisation analysis using the *cis*-eQTL results from TR groups and the GWAS results for the racing distance traits. All variants within the 100 kb flanking regions around the conditional independent cis-eQTLs were included. Colocalisation analysis was conducted using the Coloc R package (v5.2.3) [58], which includes the Sum of Single Effects (SuSiE) model [59] for genetic signals fine-mapping. The GWAS variants and eQTL pairs were considered as colocalised with the threshold of posterior probability for hypothesis 4 (i.e., a shared casual variant, PPH4) > 0.7 [60].

To infer the causality between gene expression phenotypes and GWAS traits, we performed MR analysis using the Summary-data-based Mendelian randomisation (SMR) software (v1.3.1) [14]. The top *cis*-eQTL in the TR group with nominal *P*-values < 5×10^{-5} were selected as the instrumental variables. The racing distance GWAS traits were set as the outcome. Firstly, SMR results were considered statistically significant if FDR-$P_{adj.}$ < 0.05. Furthermore, the significant gene-traits pairs were used to perform the heterogeneity in dependent instruments (HEIDI) analysis. We accepted the causality between gene expression and GWAS traits, when the HEIDI *P*-value (P_{HEIDI}) ≥ 0.05.

Results

Skeletal muscle transcriptome profiles in three different states

RNA sequencing (RNA-seq) generated a mean (± standard deviation) of 16,880,445 ± 3,126,337 clean reads per individual sample library (*n* = 230) following quality control. These reads were aligned to the EquCab3.0 reference genome, with a mean of 15,175,585 ± 2,776,929 (89.95% ± 2.87%) of the reads uniquely mapping to the genome (Table S1). Using an expression threshold of TPM ≥ 0.1 in ≥ 20% of samples, a total of 16,709 genes were considered expressed in skeletal muscle in the UR group, 16,442 genes in the UE group, and 16,444 genes in the TR group, respectively.

cis -eQTLs cluster near the transcription start site of their associated eGenes

To explore the genetic architecture underlying gene expression variation across the three distinct exercise/training states, we performed an eQTL mapping analysis to identify associations between genomic variants and gene expression levels within each group. In total, 3,044, 2,285, and 1,716 *cis*-eGenes were identified in the UR, UE, and TR groups, respectively (Fig. 1a, Table S2–S4). For each *cis*-eGene in each group, variants with nominal *P*-values below the gene-level threshold were considered significant *cis*-eVariants. In the UR, UE, and TR groups 3,479,712, 2,255,808, and 1,597,006 significant *cis*-eVariant-gene pairs were detected, respectively. Furthermore, 304, 153, and 69 eGenes were identified with at least two independent *cis*-eQTLs in the UR, UE, and TR groups, respectively (Fig. 1b). The top *cis*-eQTLs identified through permutation testing tended to cluster near the TSS of their associated eGenes (Fig. 1c), while the conditionally independent *cis*-eQTL signals were located further from the TSS (average median distance 391.7 kb from the TSS, compared to 253.6 kb for top cis-eQTLs; Wilcoxon rank-sum test, $P < 0.05$) (Fig. 1d). Additionally, a moderately negative but significant Spearman correlation (ρ) was observed between the effect sizes of *cis*-eQTLs and their respective distances from the TSS across all three states (Fig. 1e–1g).

Diverse cis-eQTL responses to exercise and training

To investigate the differential genetic responses to exercise and training, and to identify distinct sets of eQTLs activated in each condition, the effect sizes of the top SNPs for each gene across paired samples were compared. Using the UR condition as the baseline, two comparison groups were examined: UE *vs*. UR (exercise response, *n* = 50) and TR *vs*. UR (training response, *n* = 32). In UE *vs*. UR, we identified 583 and 1,110 genes with at least one significant *cis*-eQTL (FDR *P*-value < 0.05) in UR and UE, respectively. In TR *vs*. UR, 6 and 65 genes were found to have at least one significant *cis*-eQTL (FDR *P*-value < 0.05) in TR and UR, respectively. To improve the power of the comparisons, we then applied a flexible statistical method to identify reQTL between these two comparisons using mashr [51]. In the UE *vs*. UR, 3,713 and 4,706 significant eQTLs (LFSR < 0.05) were identified in the UR and UE conditions, respectively (Fig. 2a). Among these, 948 exercise reQTLs were observed (Table S5), indicating that genomic variants change

their effect on gene expression levels due to exercise. These exercise reQTLs represented 8.1% of the significant eQTLs in the UR condition and 13.9% in the UE condition. For the TR *vs*. UR comparison, 4,004 and 3,147 significant eQTLs (LFSR < 0.05) were identified in the UR and TR conditions, respectively (Fig. 2a). In this comparison, 2,366 training reQTLs were detected (Table S6), representing 39.5% of the significant eQTLs in the UR condition and 25.8% in the TR condition, respectively. It is well-established that eQTLs exhibit a strong enrichment in regions near TSS [61, 62], and the results of the eQTL mapping support this (Fig. 1c,1d). However, when the exercise/training reQTLs were compared to non-exercise/non-training reQTLs separately, a distinct localisation pattern emerged, with exercise/training reQTLs displaying only modest enrichment near TSS (Fig. 2b, S1a).

A significant proportion of the exercise reQTLs and training reQTLs showed consistent directional effects across both exercise-responsive and training-responsive conditions. For example, the top eVariant (chr13:20748509 A > C) for *ALDOA* (aldolase, fructose-bisphosphate A) was associated with much lower *ALDOA* expression in the rest (UR) than in the exercise state (UE) (Fig. 2g). However, a subset of exercise/training reQTLs exhibited opposite directional effects between UE *vs*. UR or TR *vs*. UR. For instance, the top eVariant (chr10:16157921 G > A) for *ERCC1* (ERCC excision repair 1, endonuclease non-catalytic subunit), which encodes a protein involved in recognising, excising, and repairing damaged DNA, was associated with increased *ERCC1* expression in the exercise state but decreased expression in the rest state (Fig. 2f). Similarly, the top eVariant (chr3:64961808 C > G) for *GRSF1* (G-rich RNA sequence binding factor 1) was associated with increased *GRSF1* expression in the rest state but decreased *GRSF1* expression in the exercise state (Fig. 2h). The top eVariant (chr4:44918196 C > T) for *SLC25A3* (solute carrier family 25 member 3) was associated with increased *SLC25A3* expression in the untrained state, but decreased expression in the trained state (Fig. 2i).

To elucidate the functional features of the exercise/training reGenes, GO enrichment analysis was performed. Exercise reGenes were mainly implicated in *metabolic process* (*GO:0008152*; FDR-$P_{adj.}$ = 3.50×10^{-12}), *developmental process* (*GO:0032502*, FDR-$P_{adj.}$ = 1.85×10^{-5}), *cellular response to endogenous stimulus* (*GO:0071495*, FDR-$P_{adj.}$ = 9.99×10^{-4}), and *muscle organ development* (*GO:0007517*, FDR-$P_{adj.}$ = 2.82×10^{-2}) (Figure S1b, Table S7). Training reGenes were mainly implicated in *metabolic process* (*GO:0008152*; FDR-$P_{adj.}$ = 1.22×10^{-21}), *developmental process* (*GO:0032502*; FDR-$P_{adj.}$ = 7.32×10^{-8}), *cell differentiation* (*GO:0030154*; FDR-$P_{adj.}$ = 3.02×10^{-4}), *regulation of response to stimulus* (*GO:0048583*, FDR-$P_{adj.}$ = 8.04×10^{-3}), and *muscle cell differentiation* (*GO:0042692*, FDR-$P_{adj.}$ = 4.23×10^{-2}) (Figure S1c, Table S8). Furthermore, many TF terms were enriched for both exercise reGenes and training reGenes (Table S9, S10). To investigate differences in TF representation, the proportion of TFs among exercise/training reGenes compared to non-exercise/non-training reGenes was calculated. The analysis revealed significantly higher TF enrichment in exercise/training reGenes than non-exercise/non-training reGenes (Figure S1d).

Genome-wide association studies for distance and speed demonstrate the robustness of the methodology

For average race distance (AVD, *n* = 2922), the GWAS identified 4,071 SNPs on chromosome 18 that achieved genome-wide significance after Bonferroni correction with the threshold $P = 6.67 \times 10^{-9}$ (Figure S2a, S3a, Table S11). After LD-based clumping ($r^2 = 0.8$, window size = ± 500 kb), these significant SNPs represented 84 independent genomic loci, each tagged by its lead SNP. Among them, 38 SNPs with an FDR P_{adj}. value < 0.05 overlapped with the top 40 significant SNPs identified in the original analysis [40]. In the GWAS for best race distance among all horses (BRD-ALL, *n* = 2313), 4,871 SNPs on chromosome 18 passed the genome-wide significance threshold following BF correction with a threshold of $P = 6.68 \times 10^{-9}$ (Fig. 3a, S3b, Table S12), which represented 105 significant independent genomic loci. Among these, 40 SNPs with FDR significance were also included among the top 40 significant SNPs from the previous study. Similarly, for the best race distance among elite horses (BRD-ELI, *n* = 820), the GWAS identified 3,459 SNPs on chromosome 18 that reached genome-wide significance after BF correction with a threshold of $P = 6.72 \times 10^{-9}$ (Figure S2c, S3c, Table S13), which represented 61 significant independent genomic loci. Thirty of the top 40 significant SNPs from the previous study were also statistically significant after FDR adjustment. Additionally, in the analysis of the best winning race distance among non-elite horses (BRD-NON, *n* = 1493), 3,905 SNPs on chromosome 18 were genome-wide significant following BF correction with a threshold of $P = 6.69 \times 10^{-9}$ (Figure S2d, S3d, Table S14), which represented 70 significant independent genomic loci. Forty SNPs that were significant after FDR adjustment were also included among the top 40 significant SNPs from the previous study. For the BRD-ALL, BRD-AVD and BRD-NON GWAS, SNP chr18: 66501541 C > T was the lead SNP, while SNP chr18: 66857339 A > G was the lead SNP for BRD-ELI. In the GWAS for two-year-old speed (Speed2Y, *n* = 122), no SNP reached genome-wide significance, which is consistent with our previous results. However, 81 SNPs on chromosome 8 surpassed the suggestive threshold ($P < 5 \times 10^{-5}$), defining a 1.07 Mb region (Fig. 3e, S3e, Table S15). Therefore, our previous findings were well replicated in these GWAS results, providing good validation of the reliability and robustness of both the methodologies employed and the new results obtained.

The lead SNP accounts for a substantial proportion of the race distance GWAS signals

The race distance GWAS identified a predominant associated peak on chromosome 18 (top SNP, chr18: 66501541 C > T) for the BRD-ALL, BRD-AVD, and BRD-NON traits, and (top SNP, chr18: 66857339 A > G) for the BRD-ELI trait. After conditional GWAS analysis, the significant signals within this region disappeared. No signals passed the FDR corrected threshold of $P < 5 \times 10^{-8}$ for any of the traits (Figure S4a–Sd).

TWAS analysis identifies MSTN as a key contributor to the training response to exercise

The reference panels of genotype-transcriptome data for exercise and training were generated from 71 UE and 65 TR Thoroughbred horses, respectively. The reference panel generated from the UE group was used to predict the expression of *cis*-heritable genes for the Speed2Y GWAS, while the reference panel generated from the TR group were used to predict the expression of *cis*-heritable genes for the race distance series of GWAS. Using these reference panels, 2,285 and 1,716 *cis*-eGenes were obtained for the UE and TR groups, respectively. For the four race distance phenotypes, 16 significant gene-trait pairs, representing four unique genes (*MSTN*, *ORMDL1*, *ANKAR*, and *OSGEPL1*) were identified (Fig. 4a–4b, Table S16–S19). Notably, these four genes were located within a 0.31 Mb region on chromosome 18 spanning the 66,307,427–66,610,122 bp interval. *MSTN* was the most significant associated gene in the BRD-ALL, BRD-AVD, and BRD-ELI TWAS, while *ANKAR* was the most significant associated gene in the BRD-NON TWAS. For the Speed2Y trait, no significant TWAS signal remained after correction for multiple testing using either BF or FDR adjustments. The top 10 genes with TWAS *P*-values > 0.0027 (*DHDH*, *ENSECAG00000050431*, *ENSECAG00000050186*, *ENSECAG00000051521*, *CDC37L1*, *GPT2*, *RRAGD*, *HMGN3*, *ATAT1*, and *GYS1*) were selected as the associated genes for the Speed2Y traits (Fig. 4e, Table S20).

After the conditional analysis, *ANKAR* was excluded because the conditional *P*-value was greater than 0.05 in the TWAS results for BRD-ALL and BRD-ELI (Table S21) and because it showed a high correlation (r^2 > 0.9) in BRD-AVD (Table S21). Similarly, *ORMDL1* was removed because the conditional *P*-value exceeded 0.05 in the TWAS results for BRD-ALL, BRD-ELI, and BRD-NON. In addition, the *OSGEPL1* gene was excluded from the BRD-NON TWAS for the same reason (conditional *P*-value > 0.05) (Table S21).

Colocalisation and Mendelian randomisation reveal causality between candidate genes and racing performance traits

Colocalisation and SMR analysis used the *cis*-eQTL data from the TR group and GWAS summary statistics for the racing distance trait. The results revealed that the eQTL for two genes were colocalised with the GWAS signals for the four racing distance traits, corresponding to eight gene-trait pairs. Specifically, the eQTLs for *MSTN* and *ORMDL1* were colocalised with BRD-ALL, BRD-AVD, BRD-ELI and BRD-NON (Fig. 5a, 5b, Table S22). Furthermore, SMR analysis indicated potential causal relationships between the expression of three genes (*MSTN*, *ORMDL1*, and *ANKAR*) and the four racing distance traits. On the other hand, *ORMDL1* did not exhibit a significant association between its expression and BRD-ELI traits (SMR *P* > 0.05) (Table S23).

Discussion

Genetic variants can influence complex traits by modulating gene expression, with the potential downstream consequence of altering the abundance of proteins [63, 64]. Recently, integration approaches, including TWAS, colocalisation and MR, have emerged as effective methodologies for

associating gene expression levels with phenotypic variability for a wide range of traits in both humans and livestock [47, 65–67]. Despite increasing application of these integration methodologies in mammals, to the best of our knowledge, no study published heretofore has employed a TWAS, colocalisation, and SMR framework for quantitative traits in horses, or for exercise traits in any species. Numerous genetic variants have been shown to influence racing performance in the Thoroughbred [7, 11, 68, 69]. By integrating high-resolution transcriptomics data, exploring the genomic architecture underpinning variability in molecular phenotypes and complex racing characteristics, we can improve our understanding of how the regulatory genome contributes to inter-individual variability in response to exercise. Therefore, we report here the first comprehensive multi-omics analysis of muscle-associated exercise traits, integrating genomic, transcriptomic, and GWAS data from Thoroughbred horses.

Building on our previous evaluation of eQTLs in Thoroughbred horse skeletal muscle at rest (UR) and four hours after high-intensity exercise (UE) [6], this study introduced a new timepoint: at rest after approximately six months of training (TR; *n* = 65). Here, we identified 3,044 and 2,285 eGenes in the UR and UE groups, respectively, of which 1107 (UR) and 952 (UE) were also detected in our previous study [6]. The discrepancy in the eGenes identified between the two studies may stem from differences in analytical methodologies, reference genome version, and genotype density. A total of 1,716 eGenes were identified in the TR group, of which 1,378 (76.7%) and 1,204 (67.0%) were identified as eGenes in the UR and UE groups, respectively. We observed that the top *cis*-eQTLs were located closer to the TSS than conditional independent *cis*-eQTLs, a pattern similar to that observed in multiple tissues and cells in human studies [13, 62]. Furthermore, consistent with published reports [65], a moderately negative but significant correlation between the effect sizes of the *cis*-eQTLs and their respective distances from the TSS across all three states showed an upstream bias of eQTL signals.

To decipher the genetic architecture modulating the transcriptional landscape in response to exercise or training in skeletal muscle, respectively, we compared the eVariant effect size (i.e., the posterior means derived from the mashr analysis) of the top SNPs for each gene across paired samples in the UE *vs*. UR and TR *vs*. UR comparisons. We observed that exercise or training reQTLs possessed distinct localisation patterns, differing from those of non-exercise or non-training reQTLs. Specifically, exercise- and training-related QTLs showed a more uniform distribution of distances from the TSS. This is in line with previous reports that reQTLs are particularly enriched in promoter flanking and enhancer regions [21, 70], distal regulatory regions that play a critical role in promoting gene expression in response to environmental cues and stimuli [71, 72]. Functional enrichment analysis of the genes associated with these reQTLs (i.e., reGenes) further supports this point. The exercise and training reGenes were enriched in GO terms that included *metabolic process* and *response to stress*, respectively. The *metabolic process* GO term—the most significant GO term—encompassed exercise reGenes such as *ATP2B4*, *ATP5PO*, *ATP5F1A, SLC25A12*, *ERCC1*, *GRSF1* and training reGenes such as *SLC4A7*, *SLC2A1*, *MAP4K2*, and *PPARGC1A*, which have previously been implicated in electron transport and energy supply during exercise [73–76]. Many stress response-related GO terms, including *cellular response to stimulus* and *cellular response to stress*, were enriched for both exercise reGenes (such as *ATF4*, *SOD2*, and *ERCC1*) and training reGenes (*SOD2*, *TP53*, and *PRKAA2*), which have mechanistic roles in mitochondrial function

and muscle contraction capacity, respectively [77–80]. Moreover, exercise reGenes, including *ACTN2*, *MYL11*, *ANKRD1*, and *CASQ1*, were significantly enriched in the *muscle structure development* GO term, whereas training reGenes, including *SPEG*, *MYF5*, *ACTN3*, and *MYBPC2*, were enriched in the *muscle cell differentiation* GO term. Several of these genes play well-established roles in skeletal muscle function. For instance, *CASQ1* contributes to the strengthening of skeletal muscle contraction by modulating the dihydropyridine receptor channel [81]. *ACTN2* is essential for maintaining sarcomere stability [82], and *ACTN3*, a key candidate gene for athletic traits in humans, is implicated in generating force in fast-twitch myofibres during sprint performance [83, 84].

Interestingly, exercise and training reGenes were also enriched in multiple transcription factor (TF) terms. TFs modulate wide biological regulation, and play a critical role in cell cycle control, signal transduction, and response to environmental changes [70, 85, 86]. Compared with non-exercise/non-training reGenes, the significantly higher proportion of exercise- and training-associated reGenes classified as TFs underscores the important role of TF-mediated regulation in modulating gene expression in response to exercise and training. Together, these findings suggest that context-specific eQTLs (or reQTLs) exhibit more complex regulatory effects on skeletal muscle in response to exercise and training stress.

The eQTL effects were observed to vary in their direction of action across the different exercise states. For example, we characterised an exercise reQTL (chr10:16301694 G > A) for the reGene *ERCC1* whereby horses with additional copies of the A allele had lower expression of *ERCC1* in the rest cohort (UR), in contrast to the exercise cohort (UE), where additional copies of the alternative allele yielded higher expression of *ERCC1*. The *ERCC1* gene encodes a DNA repair enzyme that recognises, excises, and repairs damaged DNA[87]. In humans, *ERCC1* polymorphisms are associated with several diseases [88, 89], and aerobic exercise significantly enhances *ERCC1* expression in rats [79]. Given the relationship between the response to exercise identified here, we hypothesise that increased expression of *ERCC1* protects against DNA damage, thereby inhibiting the occurrence of muscle-related disease and/or damage and delaying ageing [90, 91]. This observation is particularly relevant in the context of the well-established positive impact of exercise on ageing [92]. Similarly, we characterised an exercise reQTL (chr3:64961808 C > G) for the reGene *GRSF1* whereby horses that possessed additional copies of the G allele had lower expression of the gene in the exercise state (UE), compared to the rest state (UR). The *GRSF1* gene encodes a mitochondrial RNA-binding protein and has been implicated in maintaining mitochondrial function [93]. Additionally, *GRSF1*$^{-/-}$ mice have been shown to have significantly reduced muscle endurance compared to their wild-type littermates [94]. We also identified a training reQTL (chr28: 25071416 G > A) for the reGene *SLC25A3*, for which horses carrying additional copies of the A allele had decreased expression of this gene in the trained state (TR) relative to the resting state (UR). The SLC25A3 protein is a mitochondrial phosphate carrier that plays a pivotal role in the aerobic synthesis of adenosine triphosphate (ATP) by transporting inorganic phosphate into the mitochondrial matrix [95]. The different effect direction of this variant for *SLC25A3* may be related to metabolic strategy in the response to the increasing energy demands of skeletal muscle tissue initiated by training.

The TWAS analysis identified four genes (*MSTN*, *OSGEPL1*, *ANKAR*, and *ORMDL1*) that were significantly associated with racing distance traits. However, subsequent conditional testing demonstrated that only *MSTN* was consistently independently associated with the four racing distance traits. These results demonstrate that *MSTN* gene expression exerts a primary influence on the racing performance of Thoroughbred horses, consistent with previous studies from our group [7, 96]. The eQTLs for *MSTN* and *ORMDL1*, respectively, were found to colocalise with racing distance traits. These genes are located within a 0.21 Mb region on chromosome 18, and the high LD in the Thoroughbred population is likely to contribute to the collinearity and high correlation among neighbouring genes. Accordingly, SMR provides a robust method for interpreting causality and distinguishing pleiotropic effects from associations driven by LD [14]. The SMR results demonstrated that variation at the *MSTN*, *ANKAR* and *ORMDL1* loci had causal effects on racing distance traits in Thoroughbred horses, whereas *ORMDL1* did not show statistically significant evidence of a causal relationship with best racing distance for elite horses. The *MSTN* gene, which encodes a secreted ligand of the transforming growth factor-beta superfamily of proteins, has been widely reported to influence muscle hypertrophy in mammals [97–99] and race distance aptitude in Thoroughbred horses [7, 40]. The identification of other genes in the region further supports our previous observation that the region surrounding *MSTN* is highly responsive to exercise and training in muscle [11]. Furthermore, the genes we describe here have also been shown to be associated with muscle biology in other species; for example, comparative studies in dairy and beef cattle have identified SNPs in the *OSGEPL1*, *MSTN*, *ORMDL1*, and *ANKAR* genes as significantly associated with direct calving difficulty in beef cattle, likely related to roles in muscle development leading to overgrowth [100]. The *OSPEL1* gene product plays a critical role in mitochondrial function [101, 102], and variation in *ORMDL1* has been implicated in carbohydrate and lipid metabolism in the *gluteus medius* and *longissimus dorsi* of pigs [103, 104]. In addition, *ORMDL1* copy number variants (CNVs) are significantly associated with several growth traits in Chinese sheep breeds [105].

The identification of *MSTN* in the TWAS for Thoroughbred race distance traits demonstrates the power of the approach to identify functionally relevant genes for other traits of interest. Therefore, we used GWAS data for two-year-old speed (Speed2Y) to identify novel genes functionally relevant to exercise. In the Speed2Y TWAS, although they did not meet the significance threshold, four of the top ten genes (*DHDH*, *GPT2*, *ATAT1*, and *GYS1*) were selected as candidate genes based on their known biological functions. The *GPT2* gene encodes a nuclear-encoded mitochondrial enzyme that reversibly catalyses the transamination between alanine and a-ketoglutarate [106], directly implicated in the tricarboxylic acid (TCA) cycle [107]. Relevant to our findings is the observation that GPT2 regulates muscle mass and fibre diameter via thyroid hormone signalling in response to cellular energy requirements [108]. Specifically in horses, *GPT2* is differentially expressed in equine skeletal muscle satellite cells following the addition of leucine [109], a common component of dietary supplements used to aid muscle growth and recovery in exercising horses, since it has been shown to boost protein synthesis and muscle repair [110]. *GYS1*, which encodes glycogen synthase 1, catalyses the conversion of glucose to glycose, thereby regulating blood glucose level and energy reserves [111]. *GYS1* knockout mice have been reported to exhibit poor performance in glucose turnover and muscle glucose uptake, leading to a significant decrease in

exercise and endurance capacity [112]. In horses, *GYS1* expression has been shown to change following intensive exercise or endurance testing [113, 114]. Mutations in *GYS1* have also been associated with exertional rhabdomyolysis in the Quarter Horse breed [115]. Dihydrodiol dehydrogenase (DHDH) has been reported to regulate the breast muscle growth rate and intramuscular fat in chicken [116], while genetic variants proximal to *DHDH* have been associated with multiple traits in pigs, including leg conformation, growth, and meat quality [117]. Alpha tubulin acetyltransferase 1 (ATAT1) has been implicated in several cellular stresses [118, 119], and *ATAT1* knockout mice have deficits in motor coordination[120].

Conclusion

In summary, using a large animal model for exercise, we have demonstrated that combining genomics and transcriptomics data in a TWAS is a powerful approach for detecting genes with functional effects on exercise phenotypes, which we validated by the identification of *MSTN* as the greatest contributor to the racing distance phenotype [7, 40]. The TWAS for speed revealed novel genes that likely play central roles in exercise performance in Thoroughbred horses. Further utilisation of this highly selected animal model for exercise will continue to provide new insights into the regulatory and functional responses of skeletal muscle to exercise and training.

Abbreviations

GWAS
genome-wide association studies
LD
linkage disequilibrium
eQTL
expression quantitative trait locus
TWAS
transcriptome-wide association study
SMR
Summary-data-based Mendelian Randomisation
reQTL
response eQTL
SNPs
single-nucleotide polymorphisms
GREs
genomic regulatory elements
TSS
transcriptional start site
MAF
minor allele frequency

TMM
Trimmed Mean of M-values
FDR
false discovery rate
BF
Bonferroni
BH
Benjamini-Hochberg
PCs
principal components
MASH
multivariate adaptive shrinkage
LFSR
local false sign rate
GSEA
Gene set enrichment analysis
GO
Gene Ontology
TF
transcription factor
REML
restricted maximum likelihood
LRT
likelihood ratio test
LASSO
least absolute shrinkage and selection operator
ENET
elastic-net
HEIDI
heterogeneity in dependent instruments

Declarations

Acknowledgements

We thank Mr JS Bolger for granting access to his horses, B O'Connor, P O'Donovan, and yard staff at Glebe House Stables, Coolcullen, Co. Kilkenny, Ireland for assistance with the collection of the research samples.

Authors' contributions

MF managed and performed the data analysis, interpreted and visualized the results, and wrote the manuscript. TJH assisted with data management, installation and updating of software packages. JFO assisted with the methodology and data analysis in the eQTL mapping and reQTL analysis. KM and MS-T assisted with the WGS sequencing and performed the upstream analysis for WGS data. JAB assisted with the original sample collection and storage, RNA, DNA extraction, and sequence work. DEM provided the study guideline, data analysis methods, and software application. LMK assisted with the study design, horse management, sample collection and sequence work. BAM was involved in the phenotype management, SNP filtering and imputation. EWH conceived, designed and coordinated the study, managed the project, supervised the analyses, interpreted the results and had oversight of the development of and revisions to the manuscript. All authors reviewed and revised the manuscript.

Funding

This work has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant numbers SFI/11/PI/1166, 18/TIDA/6019 and 19/FFP/6879.

Availability of data and materials

The WGS data for imputation in this study are available at European Variation Archive with the accession number PRJEB108191. The RNA-seq data have been deposited at EBI ArrayExpress under the accession number E-MTAB-16558 (https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-16558).

Ethics approval and consent to participate

University College Dublin Animal Research Ethics Committee approval (AREC-P-12-55-Hill), a Department of Health license for the use of all horses in this study (B100/3525), and informed owner consent were obtained.

Consent for publication

Not applicable.

Competing interests

B.A.M is an employee, and E.W.H is a founder, director, shareholder, and scientific advisor to Zinto Labs that offers commercial genetic tests for *MSTN* gene variants. E.W.H, L.M.K, and D.E.M are named inventors on patents licenced by Zinto Labs pertaining to molecular assays for *MSTN* genetic tests.

References

1. Hill EW, Stoffel MA, McGivney BA, MacHugh DE, Pemberton JM. Inbreeding depression and the probability of racing in the Thoroughbred horse. *Proc Biol Sci* 2022, 289(1977):20220487.
2. Marvin G. The Sport of Kings: Kinship, Class and Thoroughbred Breeding in Newmarket Rebecca Cassidy. Anthrozoös. 2003;16(1):88–91. https://doi.org/10.2752/089279303786992369.
3. McManus P, Albrecht G, Graham R. The Global Horseracing Industry: Social, Economic, Environmental and Ethical Perspectives (1st ed.). Routledge. https://doi.org/10.4324/9780203132432; 2012.
4. Bailey E, Petersen JL, Kalbfleisch TS. Genetics of Thoroughbred Racehorse Performance. Annu Rev Anim Biosci. 2022;10:131–50.
5. Petersen JL, Mickelson JR, Rendahl AK, Valberg SJ, Andersson LS, Axelsson J, Bailey E, Bannasch D, Binns MM, Borges AS, et al. Genome-wide analysis reveals selection for important traits in domestic horse breeds. PLoS Genet. 2013;9(1):e1003211.
6. Farries G, Bryan K, McGivney CL, McGettigan PA, Gough KF, Browne JA, MacHugh DE, Katz LM, Hill EW. Expression Quantitative Trait Loci in Equine Skeletal Muscle Reveals Heritable Variation in Metabolism and the Training Responsive Transcriptome. Front Genet. 2019;10:1215.
7. Hill EW, McGivney BA, Gu J, Whiston R, Machugh DE. A genome-wide SNP-association study confirms a sequence variant (g.66493737C > T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. BMC Genomics. 2010;11:552.
8. Hill EW, Fonseca RG, McGivney BA, Gu J, MacHugh DE, Katz LM. MSTN genotype (g.66493737C/T) association with speed indices in Thoroughbred racehorses. J Appl Physiol (1985). 2012;112(1):86–90.
9. Shin DH, Lee JW, Park JE, Choi IY, Oh HS, Kim HJ, Kim H. Multiple Genes Related to Muscle Identified through a Joint Analysis of a Two-stage Genome-wide Association Study for Racing Performance of 1,156 Thoroughbreds. Asian-Australas J Anim Sci. 2015;28(6):771–81.
10. Rooney MF, Hill EW, Kelly VP, Porter RK. The speed gene effect of myostatin arises in Thoroughbred horses due to a promoter proximal SINE insertion. PLoS ONE. 2018;13(10):e0205664.
11. Farries G, McGettigan PA, Gough KF, McGivney BA, MacHugh DE, Katz LM, Hill EW. Genetic contributions to precocity traits in racing Thoroughbreds. Anim Genet. 2018;49(3):193–204.
12. Rooney MF, Porter RK, Katz LM, Hill EW. Skeletal muscle mitochondrial bioenergetics and associations with myostatin genotypes in the Thoroughbred horse. PLoS ONE. 2017;12(11):e0186247.
13. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis, Working G, Enhancing Gg, Fund NIHC, Nci N, Nhgri N, Nimh N et al. Nih/Nida : Genetic effects on gene expression across human tissues. *Nature* 2017, 550(7675):204–213.

14. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48(5):481–7.
15. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–.
16. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52.
17. Joo JW, Sul JH, Han B, Ye C, Eskin E. Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. Genome Biol. 2014;15(4):r61.
18. Cheung VG, Spielman RS. The genetics of variation in gene expression. Nat Genet. 2002;32 Suppl:522–5.
19. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, Imboywa SH, Chipendo PI, Ran FA, Slowikowski K, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science. 2014;343(6175):1246980.
20. Nedelec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier JC, Freiman A, Sams AJ, Hebert S, et al. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell. 2016;167(3):657–69. e621.
21. Lin W, Wall JD, Li G, Newman D, Yang Y, Abney M, VandeBerg JL, Olivier M, Gilad Y, Cox LA. Genetic regulatory effects in response to a high-cholesterol, high-fat diet in baboons. Cell Genom. 2024;4(3):100509.
22. Giambartolomei C, Zhenli Liu J, Zhang W, Hauberg M, Shi H, Boocock J, Pickrell J, Jaffe AE, CommonMind C, Pasaniuc B, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. Bioinformatics. 2018;34(15):2538–45.
23. Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99(6):1245–60.
24. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, Lloyd-Jones LR, Marioni RE, Martin NG, Montgomery GW, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun. 2018;9(1):918.
25. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51(4):592–9.
26. Bryan K, McGivney BA, Farries G, McGettigan PA, McGivney CL, Gough KF, MacHugh DE, Katz LM, Hill EW. Equine skeletal muscle adaptations to exercise and training: evidence of differential regulation of autophagosomal and mitochondrial components. BMC Genomics. 2017;18(1):595.

27. Valette JP, Barrey E, Jouglin M, Courouce A, Auvinet B, Flaux B. Standardisation of muscular biopsy of gluteus medius in French trotters. Equine Vet J Suppl 1999(30):342–4.
28. McGivney BA, Eivers SS, MacHugh DE, MacLeod JN, O'Gorman GM, Park SD, Katz LM, Hill EW. Transcriptional adaptations following exercise in thoroughbred horse skeletal muscle highlights molecular mechanisms that lead to muscle hypertrophy. BMC Genomics. 2009;10:638.
29. Eivers SS, McGivney BA, Fonseca RG, MacHugh DE, Menson K, Park SD, Rivero JL, Taylor CT, Katz LM, Hill EW. Alterations in oxidative gene expression in equine skeletal muscle following exercise and training. Physiol Genomics. 2010;40(2):83–93.
30. Felix Krueger F, James P, Ewels E, Afyounian M, Weinstein. Benjamin Schuster-Boeckler, Gert Hulselmans, sclamons: GitHub repository. https://github.com/FelixKrueger/TrimGalore. 2023.
31. Andrews S. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.
32. Martin M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. EMBnet J. 2011;17:10–2.
33. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
34. Kalbfleisch TS, Rice ES, DePriest MS Jr., Walenz BP, Hestand MS, Vermeesch JR, BL OC, Fiddes IT, Vershinina AO, Saremi NF, et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun Biol. 2018;1:197.
35. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
36. Bioconductor Package Maintainer. liftOver: Changing genomic coordinate systems with rtracklayer::liftOver. R package version 1.28.0. https://www.bioconductor.org/help/workflows/liftOver/. 2024.
37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
38. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103(3):338–48.
39. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021;108(10):1880–90.
40. Hill EW, McGivney BA, Rooney MF, Katz LM, Parnell A, MacHugh DE. The contribution of myostatin (MSTN) and additional modifying genetic loci to race distance aptitude in Thoroughbred horses racing in different geographic regions. Equine Vet J. 2019;51(5):625–33.
41. Farries G, Gough KF, Parnell AC, McGivney BA, McGivney CL, McGettigan PA, MacHugh DE, Katz LM, Hill EW. Analysis of genetic variation contributing to measured speed in Thoroughbreds identifies genomic regions involved in the transcriptional response to exercise. Anim Genet. 2019;50(6):670–85.

42. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
43. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG, Meta-analysis C, Madden PA, Heath AC, Martin NG et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012, 44(4):369–375, S361-363.
44. Chen Y, Lun AT, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Res. 2016;5:1438.
45. Zhou HJ, Li L, Li Y, Li W, Li JJ. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. Genome Biol. 2022;23(1):210.
46. Vosa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Yazar S, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53(9):1300–10.
47. Teng J, Gao Y, Yin H, Bai Z, Liu S, Zeng H, Pig GC, Bai L, Cai Z, Zhao B, et al. A compendium of genetic regulatory effects across pig tissues. Nat Genet. 2024;56(1):112–23.
48. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 2016;32(10):1479–85.
49. Taylor-Weiner A, Aguet F, Haradhvala NJ, Gosai S, Anand S, Kim J, Ardlie K, Van Allen EM, Getz G. Scaling computational genomics to millions of individuals with GPUs. Genome Biol. 2019;20(1):228.
50. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal Royal Stat Society: Ser B (Methodological) January. 1995;57(1):289–300.
51. Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet. 2019;51(1):187–95.
52. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 2004;32(Database issue):D262–266.
53. Tuggle CK, Giuffra E, White SN, Clarke L, Zhou H, Ross PJ, Acloque H, Reecy JM, Archibald A, Bellone RR, et al. GO-FAANG meeting: a Gathering On Functional Annotation of Animal Genomes. Anim Genet. 2016;47(5):528–33.
54. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res* 2020, 9.
55. Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, Becker A, Bennett R, Beracochea M, Berry A, et al. Ensembl 2025. Nucleic Acids Res. 2025;53(D1):D948–57.
56. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The Human Transcription Factors. Cell. 2018;172(4):650–65.

57. Julious SA: Two-sided confidence intervals for the single proportion: comparison of seven methods by Robert G., Newcombe. Statistics in Medicine 1998; 17:857–872. *Stat Med* 2005, 24(21):3383–3384.
58. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383.
59. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Ser B Stat Methodol. 2020;82(5):1273–300.
60. Cai W, Zhang Y, Chang T, Wang Z, Zhu B, Chen Y, Gao X, Xu L, Zhang L, Gao H, et al. The eQTL colocalization and transcriptome-wide association study identify potentially causal genes responsible for economic traits in Simmental beef cattle. J Anim Sci Biotechnol. 2023;14(1):78.
61. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008;4(10):e1000214.
62. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009;325(5945):1246–50.
63. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16(4):197–212.
64. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45(10):1238–43.
65. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60.
66. Zhang Y, Wang M, Li Z, Yang X, Li K, Xie A, Dong F, Wang S, Yan J, Liu J. An overview of detecting gene-trait associations by integrating GWAS summary statistics and eQTLs. Sci China Life Sci. 2024;67(6):1133–54.
67. O'Grady JF, McHugo GP, Ward JA, Hall TJ, Faherty O'Donnell SL, Correia CN, Browne JA, McDonald M, Gormley E, Riggio V, et al. Integrative genomics sheds light on the immunogenetics of tuberculosis in cattle. Commun Biol. 2025;8(1):479.
68. Han H, McGivney BA, Allen L, Bai D, Corduff LR, Davaakhuu G, Davaasambuu J, Dorjgotov D, Hall TJ, Hemmings AJ, et al. Common protein-coding variants influence the racing phenotype in galloping racehorse breeds. Commun Biol. 2022;5(1):1320.
69. Han H, McGivney BA, Farries G, Katz LM, MacHugh DE, Randhawa IAS, Hill EW. Selection in Australian Thoroughbred horses acts on a locus associated with early two-year old speed. PLoS ONE. 2020;15(2):e0227212.

70. Kim-Hellmuth S, Bechheim M, Putz B, Mohammadi P, Nedelec Y, Giangreco N, Becker J, Kaiser V, Fricker N, Beier E, et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. Nat Commun. 2017;8(1):266.
71. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczynski B, Riddell A, Furlong EE. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet. 2012;44(2):148–56.
72. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drablos F, Lennartsson A, Ronnerblad M, Hrydziuszko O, Vitezic M, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015;347(6225):1010–4.
73. Chang J, Wu W, Qian P, Lu Z, He X, Wang F, Zhang T. Multi-omics study on the effect of moderate-intensity exercise on protein lactylation in mouse muscle tissue. Front Cell Dev Biol. 2024;12:1472338.
74. Mao Y, Zhang J, Zhou Q, He X, Zheng Z, Wei Y, Zhou K, Lin Y, Yu H, Zhang H, et al. Hypoxia induces mitochondrial protein lactylation to limit oxidative phosphorylation. Cell Res. 2024;34(1):13–30.
75. Wang X, Miao D, Ye S, Xian H, Ge W. Abnormal Expression of COX5B Gene and Disorder of Mitochondrial Function in Cryptorchid Rats. J Cell Mol Med. 2024;28(22):e70234.
76. Coluccino G, Negro A, Filippi A, Bean C, Muraca VP, Gissi C, Canetti D, Mimmi MC, Zamprogno E, Ciscato F, et al. N-terminal cleavage of cyclophilin D boosts its ability to bind F-ATP synthase. Commun Biol. 2024;7(1):1486.
77. Sailani MR, Halling JF, Moller HD, Lee H, Plomgaard P, Pilegaard H, Snyder MP, Regenberg B. Lifelong physical activity is associated with promoter hypomethylation of genes involved in metabolism, myogenesis, contractile properties and oxidative stress resistance in aged human skeletal muscle. Sci Rep. 2019;9(1):3272.
78. Sanfrancesco VC, Hood DA. Acute contractile activity induces the activation of the mitochondrial integrated stress response and the transcription factor ATF4. J Appl Physiol (1985). 2025;138(3):857–71.
79. Ji N, Zhao W, Qian H, Yan X, Zong R, Zhang Y, Lao K. Aerobic exercise promotes the expression of ERCC1 to prolong lifespan: A new possible mechanism. Med Hypotheses. 2019;122:22–5.
80. Beyfuss K, Hood DA. A systematic review of p53 regulation of oxidative stress in skeletal muscle. Redox Rep. 2018;23(1):100–17.
81. Mosca B, Delbono O, Laura Messi M, Bergamelli L, Wang ZM, Vukcevic M, Lopez R, Treves S, Nishi M, Takeshima H, et al. Enhanced dihydropyridine receptor calcium channel activity restores muscle strength in JP45/CASQ1 double knockout mice. Nat Commun. 2013;4:1541.
82. Ranta-Aho J, Olive M, Vandroux M, Roticiani G, Dominguez C, Johari M, Torella A, Bohm J, Turon J, Nigro V, et al. Mutation update for the ACTN2 gene. Hum Mutat. 2022;43(12):1745–56.
83. Yang N, MacArthur DG, Gulbin JP, Hahn AG, Beggs AH, Easteal S, North K. ACTN3 genotype is associated with human elite athletic performance. Am J Hum Genet. 2003;73(3):627–31.
84. Pickering C, Kiely J. ACTN3: More than Just a Gene for Speed. Front Physiol. 2017;8:1080.

85. Reed EB, Sitikov A, Shin KWD, Hamanaka RB, Cetin-Atalay R, Mutlu GM, Mongin AA, Dulin NO. Ga12 and Ga13 proteins are required for TGF-beta-induced myofibroblast differentiation. Biochem J 2024.
86. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The Human Transcription Factors. Cell. 2018;175(2):598–9.
87. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JH. Understanding nucleotide excision repair and its roles in cancer and ageing. Nat Rev Mol Cell Biol. 2014;15(7):465–81.
88. Khalouei A, Masoumi-Ardakani Y, Jafarzaheh A, Kalantari Khandani B, Sedghy F, Khosravi Mashizi A, Yaghoobi MM, Zangouey M, Shahouzehi B. Association of ERCC1 Gene Polymorphisms (rs3212986 and rs11615) With the Risk of Lung Cancer in a Population From Southeast Iran. J Res Health Sci. 2024;24(4):e00631.
89. Yiu WS, Chu TSM, Meng Y, Kong FS. DNA Repair Genetics and the Risk of Radiation Pneumonitis in Patients With Lung Cancer: A Systematic Review and Meta-analysis. Clin Oncol (R Coll Radiol). 2024;36(7):e182–96.
90. He Y, Yang W, Huang L, Mever MA, Ramautar R, Harms A, Rijksen Y, Brandt RMC, Barnhoorn S, Smit K, et al. Metabolomic analysis of dietary-restriction-induced attenuation of sarcopenia in prematurely aging DNA repair-deficient mice. J Cachexia Sarcopenia Muscle. 2024;15(3):868–82.
91. Juttner AA, Ataei Ataabadi E, Golshiri K, de Vries R, Garrelds IM, Danser AHJ, Visser JA, Roks AJM. Adiponectin secretion by perivascular adipose tissue supports impaired vasodilation in a mouse model of accelerated vascular smooth muscle cell and adipose tissue aging. Vascul Pharmacol. 2024;154:107281.
92. Distefano G, Goodpaster BH. Effects of Exercise and Aging on Skeletal Muscle. Cold Spring Harb Perspect Med 2018, 8(3).
93. Yin W, Yang L, Kong D, Nie Y, Liang Y, Teng CB. Guanine-rich RNA binding protein GRSF1 inhibits myoblast differentiation through repressing mitochondrial ROS production. Exp Cell Res. 2019;381(1):139–49.
94. Driscoll RK, Krasniewski LK, Cockey SG, Yang JH, Piao Y, Lehrmann E, Zhang Y, Michel M, Noh JH, Cui CY, et al. GRSF1 deficiency in skeletal muscle reduces endurance in aged mice. Aging. 2021;13(11):14557–70.
95. Mayr JA, Merkel O, Kohlwein SD, Gebhardt BR, Bohles H, Fotschl U, Koch J, Jaksch M, Lochmuller H, Horvath R, et al. Mitochondrial phosphate-carrier deficiency: a novel disorder of oxidative phosphorylation. Am J Hum Genet. 2007;80(3):478–84.
96. Hill EW, Gu J, Eivers SS, Fonseca RG, McGivney BA, Govindarajan P, Orr N, Katz LM, MacHugh DE. A sequence polymorphism in MSTN predicts sprinting ability and racing stamina in thoroughbred horses. PLoS ONE. 2010;5(1):e8645.
97. McPherron AC, Lawler AM, Lee SJ. Regulation of skeletal muscle mass in mice by a new TGF-beta superfamily member. Nature. 1997;387(6628):83–90.
98. Gonzalez-Cadavid NF, Taylor WE, Yarasheski K, Sinha-Hikim I, Ma K, Ezzat S, Shen R, Lalani R, Asa S, Mamita M, et al. Organization of the human myostatin gene and expression in healthy men and HIV-

infected men with muscle wasting. Proc Natl Acad Sci U S A. 1998;95(25):14938–43.

99. Bass J, Oldham J, Sharma M, Kambadur R. Growth factors controlling muscle development. Domest Anim Endocrinol. 1999;17(2–3):191–7.
100. Purfield DC, Bradley DG, Evans RD, Kearney FJ, Berry DP. Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. Genet Sel Evol. 2015;47(1):47.
101. Zhou JB, Wang Y, Zeng QY, Meng SX, Wang ED, Zhou XL. Molecular basis for t6A modification in human mitochondria. Nucleic Acids Res. 2020;48(6):3181–94.
102. Zhang Y, Zhou JB, Yin Y, Wang ED, Zhou XL. Multifaceted roles of t6A biogenesis in efficiency and fidelity of mitochondrial gene expression. Nucleic Acids Res. 2024;52(6):3213–33.
103. Gonzalez-Prendes R, Quintanilla R, Marmol-Sanchez E, Pena RN, Ballester M, Cardoso TF, Manunza A, Casellas J, Canovas A, Diaz I, et al. Comparing the mRNA expression profile and the genetic determinism of intramuscular fat traits in the porcine gluteus medius and longissimus dorsi muscles. BMC Genomics. 2019;20(1):170.
104. Hjelmqvist L, Tuson M, Marfany G, Herrero E, Balcells S, Gonzalez-Duarte R. ORMDL proteins are a conserved new family of endoplasmic reticulum membrane proteins. Genome Biol. 2002;3(6):RESEARCH0027.
105. Wang X, Cao X, Wen Y, Ma Y, Elnour IE, Huang Y, Lan X, Chaogetu B, Hu L, Chen H. Associations of ORMDL1 gene copy number variations with growth traits in four Chinese sheep breeds. Arch Anim Breed. 2019;62(2):571–8.
106. Ruaud L, Keren B, Debs R, Mignot C, Mochel F. Demyelinating motor neuropathy associated with a homozygous GPT2 pathogenic variant. Muscle Nerve. 2021;63(5):E41–4.
107. McKenna MC. Glutamate dehydrogenase in brain mitochondria: do lipid modifications and transient metabolon formation influence enzyme activity? Neurochem Int. 2011;59(4):525–33.
108. Cicatiello AG, Sagliocchi S, Nappi A, Di Cicco E, Miro C, Murolo M, Stornaiuolo M, Dentice M. Thyroid hormone regulates glutamine metabolism and anaplerotic fluxes by inducing mitochondrial glutamate aminotransferase GPT2. Cell Rep. 2022;38(12):110562.
109. Xing J, Xie L, Qi X, Liu G, Akhtar MF, Li X, Bou G, Bai D, Zhao Y, Dugarjaviin M, et al. Integrated analysis of transcriptome and proteome for exploring mechanism of promoting proliferation of equine satellite cells associated with leucine. Comp Biochem Physiol Part D Genomics Proteom. 2023;48:101118.
110. Casperson SL, Sheffield-Moore M, Hewlings SJ, Paddon-Jones D. Leucine supplementation chronically improves muscle protein synthesis in older adults consuming the RDA for protein. Clin Nutr. 2012;31(4):512–9.
111. Rose AJ, Richter EA. Skeletal muscle glucose uptake during exercise: how is it regulated? Physiol (Bethesda). 2005;20:260–70.
112. Xirouchaki CE, Mangiafico SP, Bate K, Ruan Z, Huang AM, Tedjosiswoyo BW, Lamont B, Pong W, Favaloro J, Blair AR, et al. Impaired glucose metabolism and exercise capacity with muscle-specific glycogen synthase 1 (gys1) deletion in adult mice. Mol Metab. 2016;5(3):221–32.

113. Gim JA, Ayarpadikannan S, Eo J, Kwon YJ, Choi Y, Lee HK, Park KD, Yang YM, Cho BW, Kim HS. Transcriptional expression changes of glucose metabolism genes after exercise in thoroughbred horses. Gene. 2014;547(1):152–8.
114. Mycka G, Ropka-Molik K, Cywinska A, Szmatola T, Stefaniuk-Szmukier M. Molecular insights into the lipid-carbohydrates metabolism switch under the endurance effort in Arabian horses. Equine Vet J. 2024;56(3):586–97.
115. Valberg SJ, Williams ZJ, Finno CJ, Schultz A, Velez-Irizarry D, Henry ML, Gardner K, Petersen JL. Type 2 polysaccharide storage myopathy in Quarter Horses is a novel glycogen storage disease causing exertional rhabdomyolysis. Equine Vet J. 2023;55(4):618–31.
116. Wang Z, Tian W, Guo Y, Wang D, Zhang Y, Zhi Y, Li D, Li W, Li Z, Jiang R, et al. Dynamic alternations of three-dimensional chromatin architecture contribute to phenotypic characteristics of breast muscle in chicken. Commun Biol. 2024;7(1):910.
117. Wu Z, Gong H, Zhou Z, Jiang T, Lin Z, Li J, Xiao S, Yang B, Huang L. Mapping short tandem repeats for liver gene expression traits helps prioritize potential causal variants for complex traits in pigs. J Anim Sci Biotechnol. 2022;13(1):8.
118. Mackeh R, Lorin S, Ratier A, Mejdoubi-Charef N, Baillet A, Bruneel A, Hamai A, Codogno P, Pous C, Perdiz D. Reactive oxygen species, AMP-activated protein kinase, and the transcription cofactor p300 regulate alpha-tubulin acetyltransferase-1 (alphaTAT-1/MEC-17)-dependent microtubule hyperacetylation during cell stress. J Biol Chem. 2014;289(17):11816–28.
119. Ortiz HR, Cruz Flores P, Podgorski J, Ramonett A, Ahmed T, Hempel N, Charest PG, Ellis NA, Langlais PR, Montfort WR, et al. Extracellular signals induce dynamic ER remodeling through alphaTAT1-dependent microtubule acetylation. Neoplasia. 2024;53:101003.
120. Li L, Jayabal S, Ghorbani M, Legault LM, McGraw S, Watt AJ, Yang XJ. ATAT1 regulates forebrain development and stress-induced tubulin hyperacetylation. Cell Mol Life Sci. 2019;76(18):3621–40.

Figures

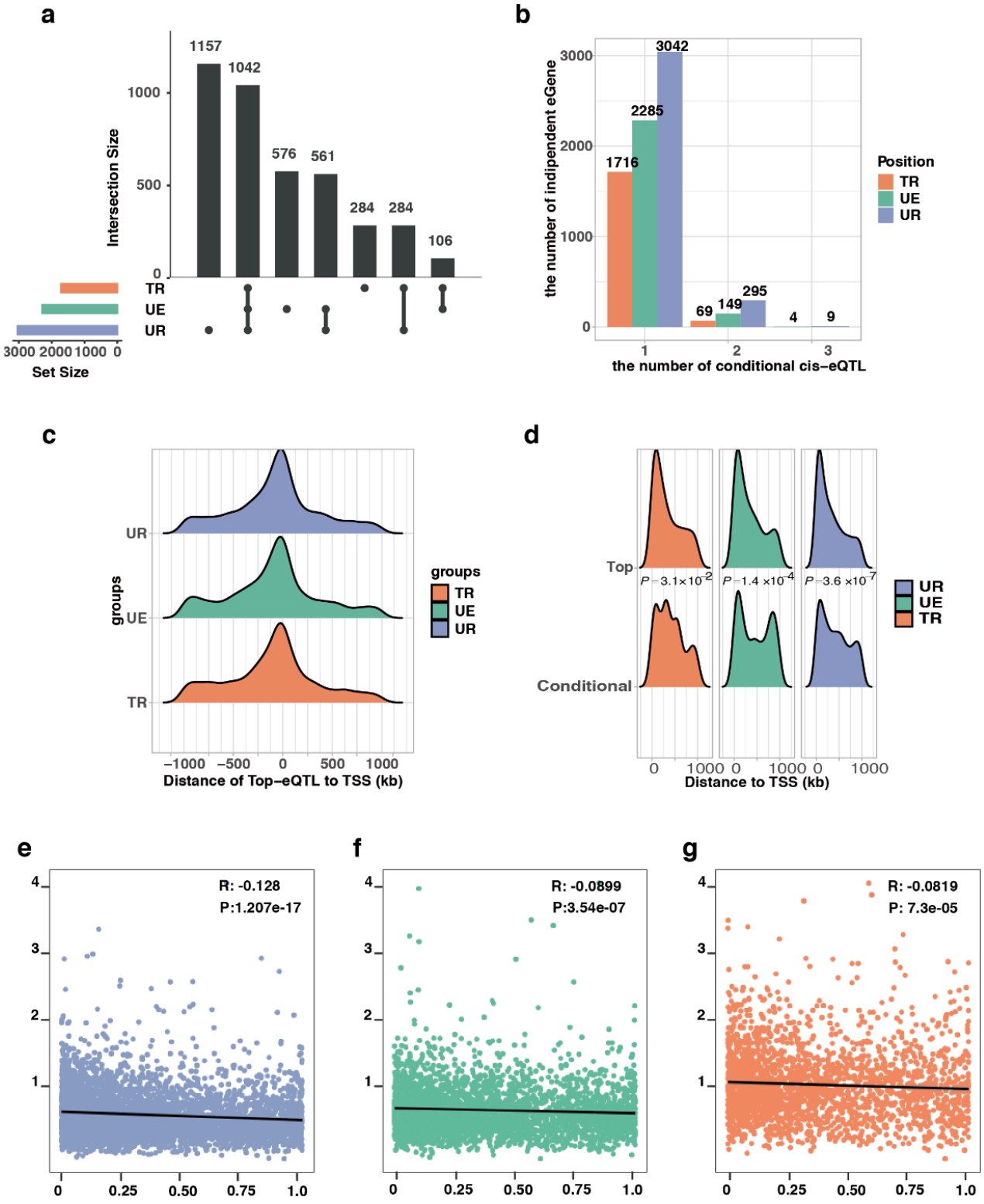

Figure 1

***Cis*-eQTL mapping results. (a)** The upsetR plot showed the interaction of *cis*-eGenes identified in the UR, UE, and TR cohorts, respectively. There were 3,044, 2,285 and 1,716 eGenes in UR, UE, and TR cohorts, respectively. **(b)** Barplot illustrated the number of eGenes with significant conditionally independent signal rank across the three states. **(c)** The ridgeline plots showed the distribution of the distance from the transcriptional start site (TSS) to top *cis*-eQTLs (right) in all three groups. More *cis*-eQTLs were

located in the regions close to TSS. **(d)** The ridgeline plots illustrate the distributions of absolute distance to TSS with the contrast between top *cis*-eQTLs and conditionally independent *cis*-eQTLs. Statistical significance of differences between the distributions was calculated by Wilcoxon Rank-Sum test. **(e, f, g)** The scatter plots showed the relationship between the distance of *cis*-eQTLs to TSS and the effect size of *cis*-eQTLs. The Spearman correlation coefficient and corresponding *P*-value were also marked in each plot. The absolute correlation coefficient of the UR cohort was larger than in the UE and TR cohorts. The three cohorts were represented by distinct colours: orange for the TR group, blue for the UE group, and purple for the UR group.

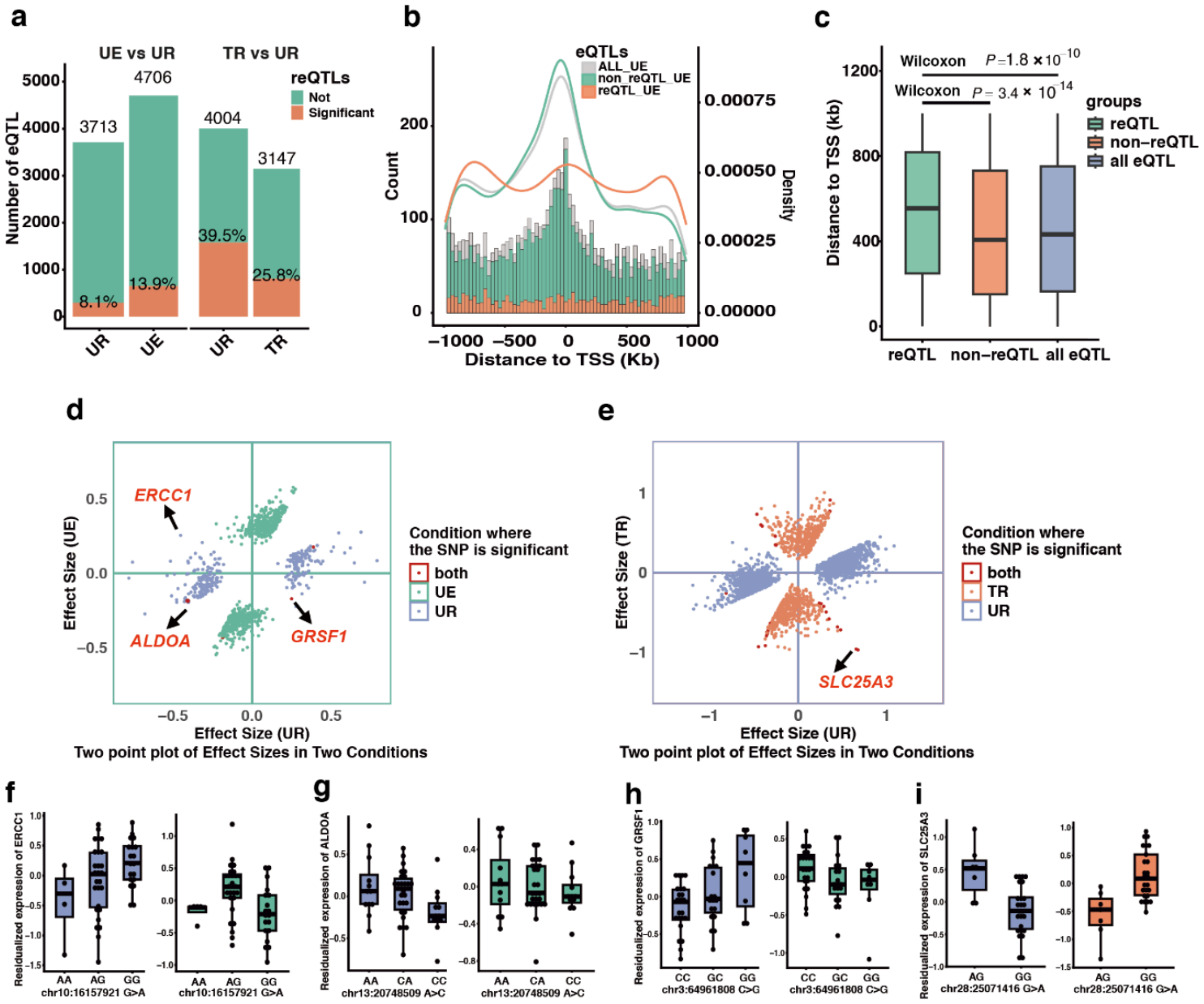

Figure 2

Response eQTLs (reQTLs) in UR vs. UE and UR vs. TR. (a) A bar plot illustrated the number of genes with significant cis-eQTLs identified by a local false sign rate (LFSR) of less than 0.05 across different

comparison groups and exercise conditions. The percentages on each bar represented the proportion of genes with significant reQTLs that exhibit fold changes greater than 2 or less than 1/2. **(b)** Distance of exercise reQTLs (orange), non-exercise reQTLs (blue), and all eQTLs (grey) to TSS of the target gene. The histogram showed the distribution of each kind of eQTL. The curves show the kernel density estimate for each kind of eQTL. **(c)** Comparison of distances to TSS between exercise reQTLs and non-exercise reQTLs, and between exercise reQTLs and all top *cis*-eQTLs. Differences in distance distributions were assessed using the Wilcoxon rank-sum test. **(d, e)** Two scatter plots presented the effect sizes of significant exercise reQTLs and training reQTLs within the two comparison groups: exercise comparison (UR vs. UE) and training comparison (UR vs. TR). The colour coding indicated the specific condition in which the reQTL was found to be significant. Red indicated eQTLs that were significant in both conditions. **(f-i)** The violin plots display the residual expression of genotypes for reGenes (*ERCC1*, *ALDOA*, *GRSF1* and *SLC25A3*).

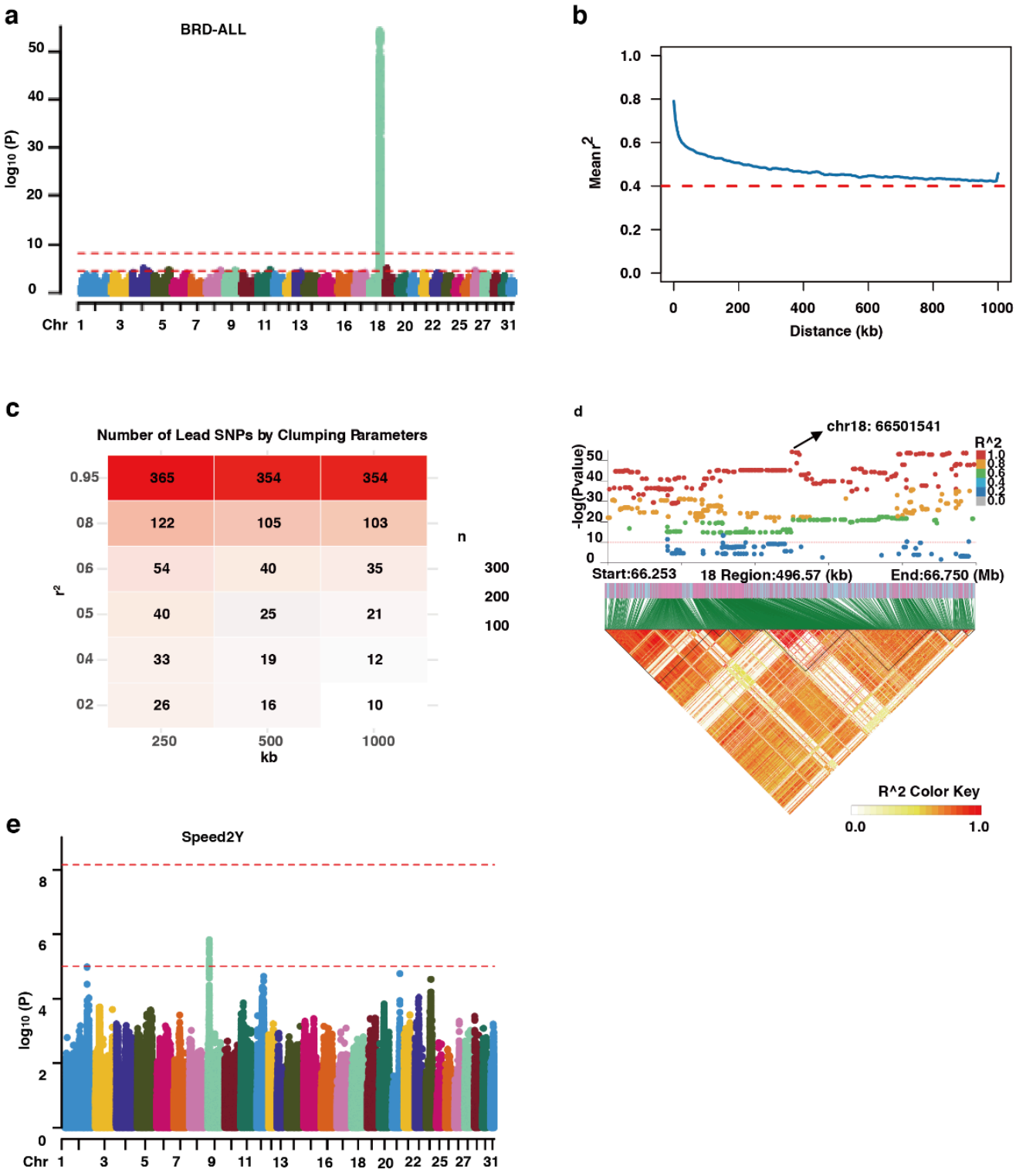

Figure 3

Manhattan plot depicting results from GWAS. (a) represented the GWAS outcomes for best racing distance for all horses. The two red horizontal line indicated the *P*-value threshold for genome wide significance using the Bonferroni (BF) and Benjamini–Hochberg (BH) for multiple testing. (b) represented the LD decay curve of chromosome 18 for all horses in this study. (c) represented the number of lead SNP for the significant BRD-ALL GWAS signals with different clumping parameters,

including different correlation r square and clumping window sizes. (d) Regional association and linkage disequilibrium (LD) plots for the top associated locus for BRD-ALL GWAS. Regional association plot showing the $-\log_{10}(P)$ values of SNPs from the GWAS around the lead SNP (chr18: 66501541). The colour of each SNP represented the LD (r^2) with the lead SNP. The bottom triangle represented the pairwise LD (r^2) pattern, where red indicated strong LD and yellow indicated weaker LD. (e) represented the GWAS results for speed of horses at two years of age (Speed2Y). The top red line marks genome-wide significance threshold using Bonferroni (BF) method, while the lower red line indicates the suggestive threshold of $P < 5 \times 10\text{-}5$. Chromosomes have been distinguished from different colours.

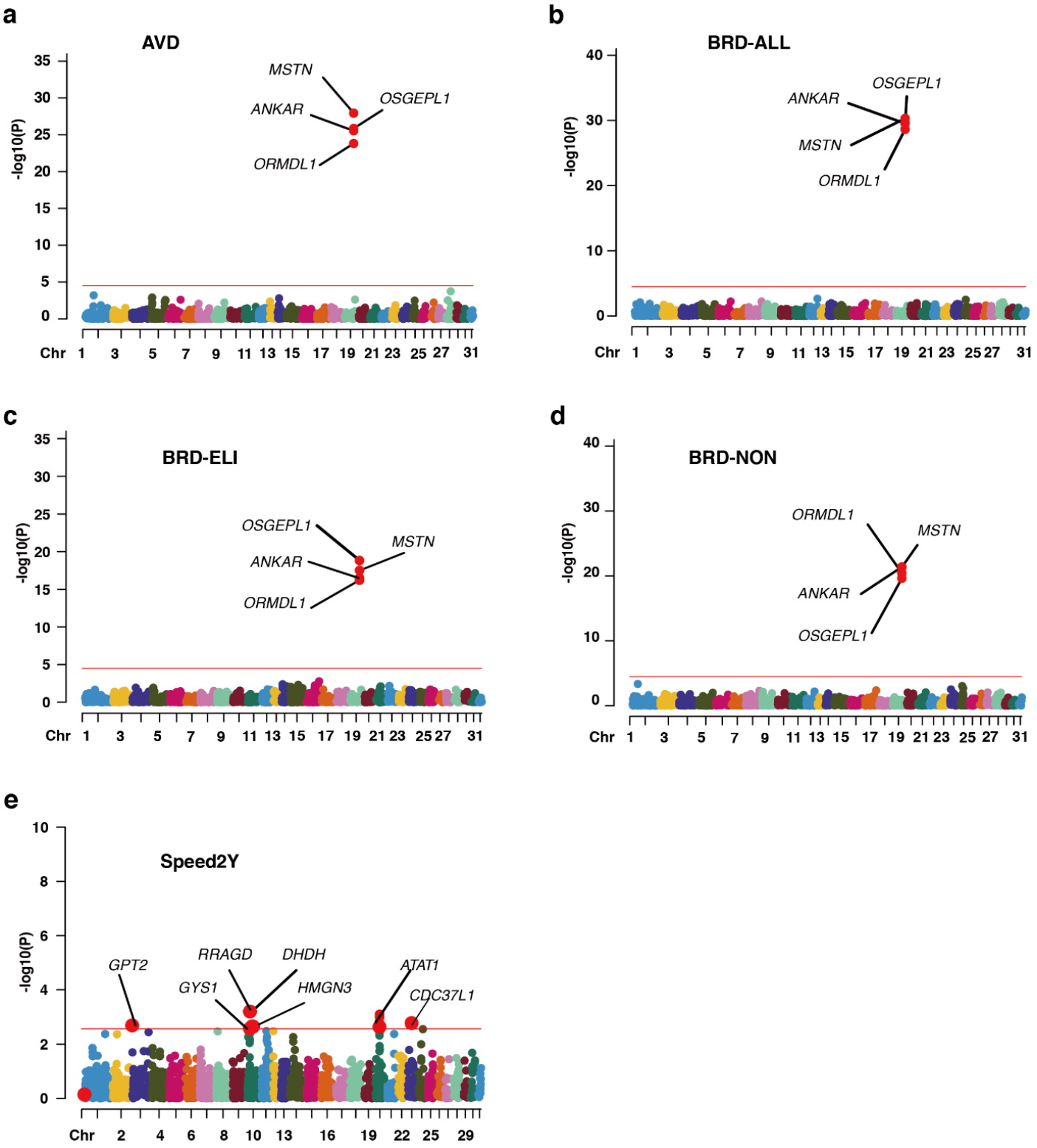

Figure 4

Manhattan plot depicting results from TWAS. (a, b, c, and d) present the TWAS outcomes for average distance for all horses (AVD), best racing distance for all horses (BRD-ALL), best racing distance for elite horses (BRD-ELI), and best racing distance for nonelite horses (BRD-NON), respectively. (e) presents the TWAS results for speed of horses at two years of age (Speed2Y). Chromosomes have been

distinguished with different colours in x axis. The y axis indicates the significant P value of TWAS. The red horizontal line indicates the *P*-value threshold, using the Bonferroni corrections for multiple testing.

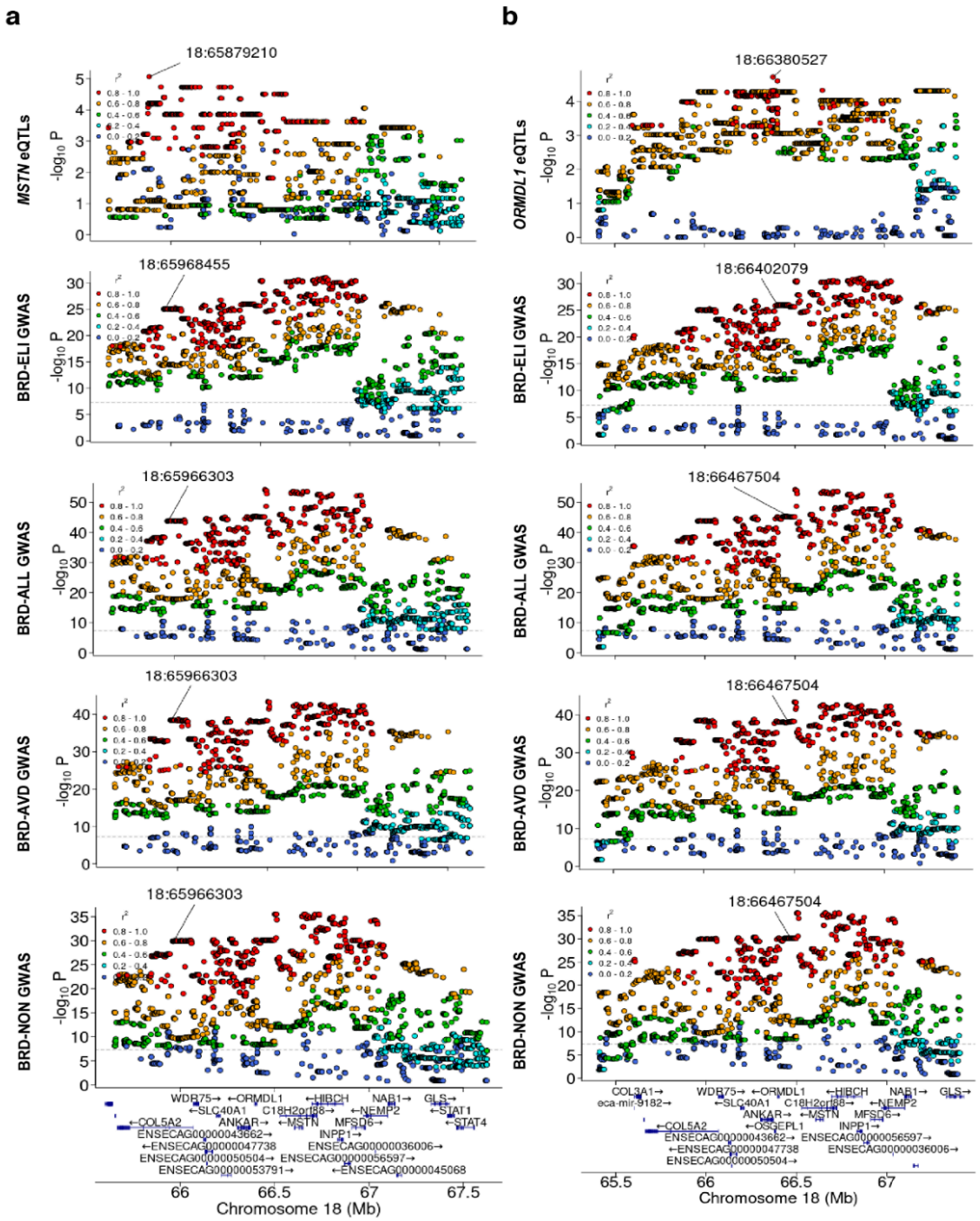

Figure 5

The colocalization of eQTLs and GWAS loci. (a) the colocalization of cis-eQTLs of *MSTN* gene and GWAS loci of BRD-ELI, BRD-ALL, BRD-AVD, and BRD-NON on chromosome 18. (b) the colocalization of

cis-eQTLs of *ORMDL1* gene and GWAS loci of BRD-ELI, BRD-ALL, BRD-AVD, and BRD-NON on chromosome 18. The colours of variants are based on their LD with the most significant variant.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.docx
- TWASsupplementarydata.xlsx
- GA.png