The contribution of myostatin (MSTN) and additional modifying genetic loci to race distance aptitude in Thoroughbred horses racing in different geographic regions

Article *in* Equine Veterinary Journal · January 2019

DOI: 10.1111/evj.13058

CITATIONS
7

READS
174

6 authors, including:

Beatrice A Mcgivney
University College Dublin
55 PUBLICATIONS **1,173** CITATIONS
SEE PROFILE

Mary Rooney
Trinity College Dublin
24 PUBLICATIONS **40** CITATIONS
SEE PROFILE

Lisa M Katz
University College Dublin
110 PUBLICATIONS **1,207** CITATIONS
SEE PROFILE

Andrew C Parnell
Maynooth University
128 PUBLICATIONS **7,376** CITATIONS
SEE PROFILE

Equine Veterinary Journal

Equine Veterinary Journal ISSN 0425-1644
DOI: 10.1111/evj.13058

The contribution of myostatin (*MSTN*) and additional modifying genetic loci to race distance aptitude in Thoroughbred horses racing in different geographic regions

E. W. HILL[†‡]*, B. A. McGIVNEY[†], M. F. ROONEY[§], L. M. KATZ[#], A. PARNELL[¶] and D. E. MacHUGH[‡¥]

[†]Plusvital Ltd, Dun Laoghaire, Co. Dublin, Ireland
[‡]UCD School of Agriculture and Food Science, University College Dublin, Belfield, Dublin, Ireland
[§]School of Biochemistry and Immunology, Trinity Biomedical Sciences Institute (TBSI), Trinity College Dublin, Dublin, Ireland
[#]UCD School of Veterinary Medicine, University College Dublin, Belfield, Dublin, Ireland
[¶]UCD Insight Centre for Data Analytics, University College Dublin, Belfield, Dublin, Ireland
[¥]UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin, Ireland.

*Correspondence email: emmeline.hill@plusvital.com; Received: 14.03.18; Accepted: 14.11.18

Summary

Background: Race distance aptitude in Thoroughbred horses is highly heritable and is influenced largely by variation at the myostatin gene (*MSTN*).
Objectives: In addition to *MSTN*, we hypothesised that other modifying loci contribute to best race distance.
Study design: Using 3006 Thoroughbreds, including 835 'elite' horses, which were >3 years old, had race records and were sampled from Europe/Middle-East, Australia/New Zealand, North America and South Africa, we performed genome-wide association (GWA) tests and separately developed a genomic prediction algorithm to comprehensively catalogue additive genetic variation contributing to best race distance.
Methods: 48,896 single-nucleotide polymorphism (SNP) genotypes were generated from high-density SNP genotyping arrays. Heritability estimates, tests of GWA and genomic prediction models were derived for the phenotypes: average race distance, best race distance for elite, nonelite and all winning horses.
Results: Heritability estimates were high (h^2_m = 0.51, best race distance – elite; h^2_m = 0.42, best race distance – nonelite; h^2_m = 0.40, best race distance – all) and most of the variation was attributed to the *MSTN* gene. *MSTN* locus SNPs were the most strongly associated with the trait and included BIEC2-438999 (ECA18:66913090; P = 4.51 × 10^{-110}, average race distance; P = 2.33 × 10^{-42}, best race distance – elite). The genomic prediction algorithm enabled the inclusion of variation from all SNPs in a model that partitioned horses into short and long cohorts following assignment of *MSTN* genotype. Additional genes with minor contributions to best race distance were identified.
Main limitations: The nongenetic influence of owner/trainer decisions on placement of horses in suitable races could not be controlled.
Conclusions: *MSTN* is the single most important genetic contributor to best race distance in the Thoroughbred. Employment of genetic prediction models will lead to more accurate placing of horses in races that are best suited to their inherited genetic potential for distance aptitude.

Keywords: horse; Thoroughbred; race distance; genomics; myostatin; prediction; GWAS

Introduction

Thoroughbred horses (Flat) race in distance categories ranging from ~1000 m (5 furlongs) to 4000 m (20 furlongs) and are rarely capable of racing at a high level across the distance range. Broadly, horses are considered to be 'sprinters', 'middle-distance' or 'stayers'. Traditionally, pedigree records and conformation characteristics have been used to attempt to identify the best race distance for an individual horse at an early age. There is a premium in some regions for horses suited to shorter distance races, since the value of the races tend to be higher; in Australia 39% of Group races are <1400 m compared with 23% in Great Britain and Ireland [1–3]. Furthermore, horses that are suited to shorter distances tend to be better suited to race as 2-year-olds [4–6]. The demand for such horses has led to proposals to encourage breeding of 'stayers' to counteract the strong selection for 'sprinters' [7]. Despite a commercial trend for early speed, paradoxically, the longer distance races (e.g. Epsom Derby, Prix de L'Arc de Triomphe, Breeders' Cup Classic, Melbourne Cup) tend to have greater prestige and higher prize money.

The power of predictive tests that use genome-wide marker information is dependent on the total genetic contribution to the trait (heritability), the number of reference animals with accurate phenotypes, and the development of robust and reliable algorithms to best estimate genetic values for animals. Variation at the myostatin gene (*MSTN*) has been shown to be a major contributor to optimum race distance in Thoroughbreds [5,8–10] and it is widely referred to as the "*Speed Gene*". Genomic selection [11], which depends on large numbers of genetic markers, became technically feasible with the availability of high-density single-nucleotide polymorphism (SNP) arrays. Genomic selection breeding programmes, on which the genomic prediction approach is predicated, have been highly successful in livestock populations [12–14].

We have evaluated the effect of *MSTN* on distance aptitude in a considerably larger population than previously reported and assessed the variable effect in different racing regions. Furthermore, we performed a genome-wide association study (GWAS) to test the hypothesis that other additive genetic variants may also be contributing to distance aptitude. Finally, we developed a genomic prediction model to catalogue the suite of genetic variants contributing to the distance trait that may be applied in the improved management of horses in-training.

Methods

Samples and phenotypes

DNA samples (n = 3006) were collected with owners' consent and approved for use in research. The sex, month of birth (corrected for hemisphere), region with greatest number of starts and race distance (m) for each race start were recorded. Regions were defined as Australia and New Zealand, Europe and Middle-East, North America and South Africa (Supplementary Item 1).

Average race distance was calculated for all horses (n = 3006). For horses that had won at least one race, the best race distance (best race

distance – all, n = 2371) was defined as the distance of the highest-grade race won. For each region the number of horses that had won at least one race was Australia and New Zealand (n = 426), Europe and Middle-East (n = 1073), North America (n = 508) and South Africa (n = 225). The best race distance was analysed separately for 'elite' (won at least one Stakes/Group/Listed race; i.e. won a Black Type race) (best race distance – elite, n = 829) and nonelite winners (best race distance – nonelite, n = 1536) and for all winning horses (best race distance – all). The criterion for best race distance – elite was as previously described [10,15]. Details are provided in Supplementary Item 2.

DNA, genotyping and quality control

DNA was isolated from blood or hair samples and genotyped using the Illumina EquineSNP50 BeadChip (SNP50), the Illumina EquineSNP70 BeadChip (SNP70) or the Affymetrix Axiom™ Equine 670K SNP genotyping array (SNP670). Only individuals and SNPs with a genotyping rate >95% were included with a minor allele frequency threshold >0.05 applied. This resulted in a set of n = 48,896 SNPs derived from the three platforms. Details of this SNP set and concordance across genotyping platforms are provided in the Supplementary Item 3.

Two polymerase chain reaction (PCR)-based assays were used to genotype a subset of n = 143 Thoroughbred horses for the *MSTN* 227 bp ERE-1 SINE insertion polymorphism (hereafter referred to as the 'SINE') to determine concordance with the SNPs (g.66493737C/T and g.66913090). Primer and assay details are provided in Supplementary Item 4.

Heritability estimates

Chip or marker heritability (h^2_m) was estimated using the genomic-relatedness-based restricted maximum-likelihood (GREML) method within Genome-wide Complex Trait Analysis (GCTA) [version 1.24.2] [16,17] for an additive model with each of the phenotypes considered as a continuous trait. Sex, *MSTN* genotype at SNP g.66493737C/T (representing the *MSTN* locus), month of birth (corrected for hemisphere) and region were included as covariates in the estimation of h^2_m, each separately, and together in a final model.

Genome-wide association study (GWAS)

A series of GWAS were performed for each of the phenotypes using a linear model implemented through the egscore function in the R software package GenABEL. An identity-by-state matrix was calculated for all samples and principal components derived from this matrix were used to correct for population stratification. Sex, month of birth (corrected for hemisphere) and region were included as co-variates in the analyses. The Benjamini-Hochberg correction factor [18] for probability values of ≤0.05 was applied to control the false discovery rate and the Bonferroni correction for multiple testing calculated a modified P-value threshold ($P<1.02 \times 10^{-6}$).

Genomic Prediction using Random Forests with mixed effects (RFME) modelling

A prediction model was fitted for best race distance – elite using a standard two-step approach [19]. We first fitted a fixed effects model to account for SNP effects and subsequently a univariate random effects model to account for residual polygenic effects constrained by genetic relatedness. To obtain the best possible prediction model we used a Random Forests [20] machine learning method to account for the fixed effects. All calculations were performed using the *R* programming language. Since the data set contained <1000 horses (n = 835) we used fivefold cross validation [21] to determine the out-of-sample performance of the model (Supplementary Item 5). To generate predictions for future (i.e. unknown) horses the model run was repeated on the full data set (i.e. no cross validation) to produce estimated important SNPs (via the Random Forest variable importance score). The procedure corresponds to a standard GBLUP run (fitted via REML) with the linear fixed effects portion replaced by a more flexible random forests approach.

The inbreeding co-efficient was included in the Random Forest feature as there is evidence (e.g. in Norwegian trotters) to indicate that inbreeding may influence performance and the level of pedigree-based inbreeding is often a factor which Thoroughbred breeders take into consideration when making mating decisions. For the calculation of the inbreeding coefficient, the SNP dataset was pruned based on the variance inflation factor (VIF) to 9659 SNPs using PLINK [version 1.07] [22].

Bioinformatics and gene mining

Gene clusters were extracted from the top 100 SNPs from the GWAS. A locus was defined as one or multiple consecutively associated SNPs within a chromosomal region with all distances <1 Mb between two adjacent associated loci. Genes within 500 kb up and downstream of the flanking SNPs for each locus were extracted from Ensembl BioMart.

Results

Heritability

Marker or chip heritability estimates ranged from h^2_m = 0.25–0.51 (Table 1). When no co-variate was included heritability estimates were highest for average race distance (h^2_m = 0.50). The highest estimate was for best race distance – elite with sex included as a co-variate (h^2_m = 0.51). The estimates were similar for best race distance – nonelite (h^2_m = 0.42) and best race distance – all (h^2_m = 0.40). Considering the inclusion and exclusion of co-variates, sex, region and month of birth contributed little to the heritability estimates. However, *MSTN* genetic variation (represented by genotypes at the g.66493737C/T SNP) contributed considerably to the heritability estimates; the highest proportion of the variability explained by the *MSTN* SNP was for best race distance – elite (0.25) and the lowest proportion of the heritable variation for best race distance – nonelite (0.11).

Genome-wide association study (GWAS)

We performed a GWAS for each phenotype, including region, sex and month of birth as co-variates and a GWAS for BRD-E where the *MSTN* SNP (g.66493737C/T) was also included as a co-variate. For each of the phenotypes SNPs on ECA18 at the *MSTN* gene region were the highest ranked: 64 (average race distance), 42 (best race distance – elite), 44 (best race distance – nonelite) and 70 (best race distance – all) consecutive top-ranked SNPs (Supplementary Item 6); the SNPs spanned regions of 7.2 Mb (best race distance – elite) to 16.5 Mb (best race distance – all). In all cases, the highest ranked SNP was BIEC2-438999 (ECA18:66913090; $P = 4.51 \times 10^{-110}$, average race distance; $P = 2.33 \times 10^{-42}$, best race distance – elite) (Supplementary Item 6). The extreme peak on ECA18 can be visualised on GWAS plots for best race distance – elite and average race distance (Fig 1a, b) and best race distance – nonelite and best race distance – all (Supplementary Item 7). Genes were identified within the regions containing the top 100 SNPs in the GWAS for best race distance – elite and are provided in Supplementary Item 8. The QQ plots (Supplementary Item 9) indicated a large contribution to the trait from SNPs that did not reach the threshold for significance in the GWAS.

Prediction model development and performance using RFME

A prediction model was developed for best race distance – elite since the estimates for heritability were highest for this trait (when sex was included as a covariate) and the heritable contribution from *MSTN* was also highest. The top SNP contributing to the prediction model was also the top SNP in the GWAS and contributed 9.3% of the total variation for the phenotype. The previously described SNP (g.66493737C/T) contributed 7.5% of the total variation. The top 100 SNPs accounted for 30% of the variation in best race distance (Supplementary Item 10). Among these, 23 SNPs spanning 2.8 Mb at the *MSTN* locus on ECA18 were responsible for 27% of the variation in the trait.

Using the prediction model for best race distance – elite, the correlation between the actual and predicted phenotype was 0.59 and R^2 = 34.8%. When corrected for heritability (h^2_m = 0.50) the model identified a set of SNPs that contributed to 69.6% of the heritable variation in the trait i.e. 35% of the total phenotypic variance. Using the best race distance prediction model, the correlation between the predicted distance and the actual best

TABLE 1: Predicted heritability (h^2_m) of each phenotype with co-variates. MOBc refers to the month of birth corrected for hemisphere

Phenotype	Co-variates	h^2_m	s.e.	P value	n
Average race distance	None	0.50	0.03	<1E-17	3006
Average race distance	MSTN genotype	0.38	0.03	<1E-17	3006
Average race distance	Region	0.44	0.03	<1E-17	3006
Average race distance	Sex	0.50	0.03	<1E-17	3006
Average race distance	MOBc	0.50	0.03	<1E-17	3006
Average race distance	Final	0.37	0.03	<1E-17	3006
Best race distance – all	None	0.40	0.03	<1E-17	2371
Best race distance – all	MSTN genotype	0.26	0.03	<1E-17	2371
Best race distance – all	Region	0.38	0.03	<1E-17	2371
Best race distance – all	Sex	0.41	0.03	<1E-17	2371
Best race distance – all	MOBc	0.40	0.03	<1E-17	2371
Best race distance – all	Final	0.26	0.03	<1E-17	2371
Best race distance – elite	None	0.49	0.06	<1E-17	835
Best race distance – elite	MSTN genotype	0.25	0.06	2.12E-08	835
Best race distance – elite	Region	0.48	0.06	<1E-17	835
Best race distance – elite	Sex	0.51	0.06	<1E-17	835
Best race distance – elite	MOBc	0.49	0.06	<1E-17	835
Best race distance – elite	Final	0.26	0.06	1.14E-08	835
Best race distance – nonelite	None	0.42	0.04	<1E-17	1536
Best race distance – nonelite	MSTN genotype	0.30	0.04	<1E-17	1536
Best race distance – nonelite	Region	0.37	0.04	<1E-17	1536
Best race distance – nonelite	Sex	0.42	0.04	<1E-17	1536
Best race distance – nonelite	MOBc	0.42	0.04	<1E-17	1536
Best race distance – nonelite	Final	0.29	0.04	<1E-17	1536

win distance for an independent set of horses (best race distance – nonelite, n = 1536) was 0.46 and R^2 = 21.3%. When corrected for heritability (h^2_m = 0.42) the predictor SNPs accounted for 51% of the heritable variation in the trait.

Phenotypic variation among regions and *MSTN* genotypes

The data were analysed separately for each *MSTN* genotype defined by the g.66493737C/T SNP and separately in each of the four main racing regions represented in the data set (n = 835; Europe and Middle East, Australia and New Zealand, North America, South Africa). Regional variation in the distribution of *MSTN* SNP genotypes was observed (Table 2, Fig 2). For instance, there were almost twice as many C:C horses among elite race winners in Australia and New Zealand (0.46) compared with Europe and Middle East (0.26) and North America (0.28). There were almost seven times as many C:Cs as T:Ts within the Australia and New Zealand population.

In each region best race distances were partitioned among the genotype cohorts (Table 3, Supplementary Item 11). In Australia and New Zealand >88% of C:Cs had a best race distance ≤1600 m and in Europe and Middle East, North America and South Africa a best race distance ≤1600 m was observed for 83, 64 and 85% of C:C horses, respectively. In Australia and New Zealand 4% of C:Cs had a best race distance >2000 m and there was no observation of C:Cs with best race distance >2200 m in Europe and Middle East or North America. Conversely, >85% of T:Ts had a best race distance >1600 m; in Australia and New Zealand (77%), Europe and Middle East (89%) and North America (80%). It also appears that the race pattern in different racing jurisdictions has a major impact on *MSTN* genetic variation in a population; for example, in Australia and New Zealand >54% of Group races are competed at <1400 m, and >45% of the population are C:C, whereas in Europe and Middle East approximately 25% of the population are C:C to meet the demand of <15% of Group 1 races competed at <1600 m.

To further refine the distance predictions, the genomic prediction model was used to separate the horses into 'short' and 'long' cohorts within each region. Using the median distance for each genotype as the cut-off for each cohort, the distance ranges were examined for each of the six cohorts (i.e. C:C-short, C:C-long, C:T-short, C:T-long, T:T-short and T:T-long). The mean best race distance for elite horses and all horses (elite and nonelite winners) split by region and predicted C:C/C:T/T:T long/short cohort are provided in Tables 4 and 5 and Figure 3. When all horses were included there was a significant difference (P<0.05) in mean best race distance between each long/short cohort apart from the North America C:C and C:T cohorts. This may be explained by the narrow range of race distances in these cohorts ranging from 1000 to 2414 m with mean distance of 1691 ± 303 m. This reflects the race pattern in North America where >58% of Group 1 races are competed at the 1700–2000 m distance range. The most significant difference was observed for the Europe and Middle East C:T-short and C:T-long cohorts with C:T-long horses having on average a best race distance 241 m longer than C:T-short horses ($P = 2.24 \times 10^{-09}$).

In Australia and New Zealand 96% of C:C-short and 86% of C:C-long had a best race distance ≤1600 m (P = 0.0007). The greatest difference in C:C-short and C:C-long distance was observed in Europe and Middle East, where 84% of C:C-short and 71% of C:C-long had best race distance ≤1600 m ($P = 7.23 \times 10^{-05}$). In Australia and New Zealand there was no C:C-short horse with a best race distance >2000 m compared with 5% C:C-long. Variation was also observed among the short and long T:T cohorts. In Australia and New Zealand 24% of T:T-short had a best race distance >2000 m compared with 63% of T:T-long (P = 0.01) and in Europe and Middle East 51% of T:T-short had a best race distance >2000 m compared with 65% of T:T-long (P = 0.04) (Table 6).

SINE, SNP concordance

Complete (100%) concordance between the SINE and the g.66493737C/T SNP was observed (n = 143). A single individual was homozygous for the BIEC2-438999 SNP and heterozygous for the SINE. Linkage disequilibrium between the g.66493737C/T and BIEC2-438999 SNPs was r^2 = 0.93. The greatest discordance was among C:Cs, where 8% were heterozygous at BIEC2-438999 and among C:Ts 2% were homozygous for the 'long' BIEC2-438999 allele. All other differences were ≤1% (Supplementary Item 12).

Discussion

We have performed an investigation of the genetic contribution to distance phenotypes in the largest cohort of racing Thoroughbreds reported to-date. In this study, the sample size for Group/Listed race winners (best race distance – elite) was tenfold larger than the original study describing the contribution of *MSTN* genetic variation to optimum race distance [9] and was expanded to include horses that had raced in the major race regions of the world and winners of non-Group/Stakes races to increase the sample size by over 25-fold.

The highest heritability was observed for average race distance (0.50), when no co-variate was included, which may simply be explained by the largest sample size for this phenotype. Notably, the smallest contribution to the heritable variation explained by the *MSTN* SNP (0.12) was observed for the average race distance phenotype when it was included as a covariate, suggesting that other factors may be incorrectly influencing trainers regarding race distance.

Heritability estimates were highest for best race distance – elite when sex was included as a co-variate (0.51), which may be explained by the subtleties of variation in race patterns for colts and fillies previously observed [10]. The contribution of the *MSTN* SNP (g.66493737C/T) to best race distance – elite heritability was estimated to be 0.24, indicating that it is responsible for almost half the variation in genetic potential for distance. When all horses were included the heritability dropped to 0.40 and the

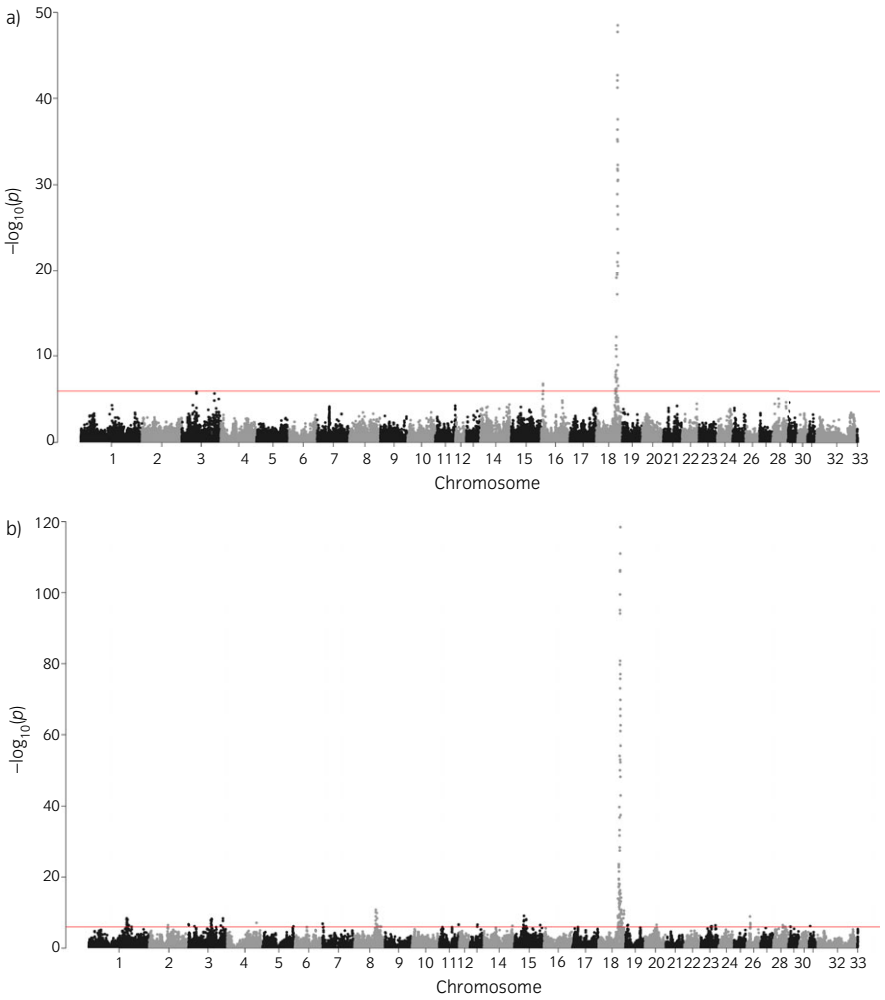

Fig 1: a) Manhattan plot showing GWAS results for best winning race distance for elite horses. The red line indicates the P-value cut-off for genome wide significance using the Benjamini Hochberg corrections for multiple testing. b) Manhattan plot showing GWAS results for average race distance for all horses. The red line indicates the P-value cut-off for genome wide significance using the Benjamini Hochberg corrections for multiple testing.

contribution of *MSTN* was estimated to be 0.14. While the sample size was smaller for best race distance – elite, these data suggest that horses competing at Group/Stakes level are generally being placed in races that are most suitable to their *MSTN* genotype.

The estimates of contribution to variance in the trait established by the genomic prediction model are in good agreement with the heritability estimates. The 24 SNPs representing the *MSTN* locus on ECA18 that were among the top 100 SNPs in the genomic prediction model were responsible for 27.1% of the variation in the trait. This is similar to the proportion of variance attributed to the *MSTN* SNP (g.66493737C/T) under the heritability model.

The BIEC2-438999 and g.66493737C/T SNPs were the top ranked SNPs in the GWAS (best race distance – elite, $P = 2.33 \times 10^{-42}$ and $P = 2.29 \times 10^{-41}$ respectively). When all horses were included, BIEC2-438999 had a significance value ($P = 4.51 \times 10^{-110}$, average race distance) equivalent to values for the strongest genetic associations that have been observed in the top 1% of human GWAS studies that included in some cases >100,000 samples. These human traits include eye colour traits [23,24], blood cell phenotypes [25,26] and male-pattern baldness [27], traits that have been clearly established to have well defined genetic contributions with major gene effects, thus lending support to the notion of a major genetic contribution of *MSTN* to the distance trait in Thoroughbreds.

TABLE 2: Regional distribution of *MSTN* genotype among elite race winners. Percentages are rounded to the nearest whole number

MSTN genotype	Australia and New Zealand	Europe and Middle East	North America	South Africa	Australia and New Zealand (%)	Europe and Middle East (%)	North America (%)	South Africa (%)
CC	118	93	45	20	46	26	28	32
CT	123	200	101	32	48	57	63	52
TT	17	61	15	10	7	17	9	16

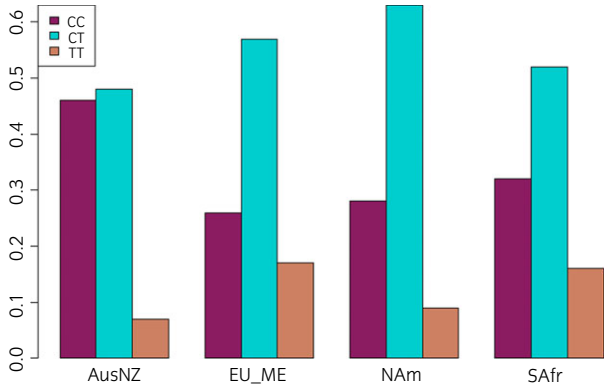

Fig 2: Variation in MSTN genotype distribution among elite horses in different racing regions. The proportion of elite horses for each of the *MSTN* genotypes in each race region is shown (Australia and New Zealand, n = 258; Europe and Middle East, n = 354; North America, n = 161; South Africa n = 62).

BIEC2-438999 was observed to make the greatest contribution to variation in the distance trait. Consequently, the BIEC2-438999 SNP may be a key influence in the genomic prediction model developed here to distinguish short and long C:Cs, C:Ts and T:Ts (g.66493737C/T), because while the LD was high (r^2 = 0.93) there was not complete concordance. Notwithstanding this, the SINE located in the *MSTN* promoter region, modulates the expression of a reporter gene suggesting that *MSTN* gene expression is under-expressed when the SINE is present [28]. This evidence supports the hypothesis that the SINE is the functional variant affecting best race distance and that LD accounts for the association between SNPs in the *MSTN* region and best race distance. Although Santagostino *et al.* [28] reported incomplete concordance between the SINE and the g.66493737C/T SNP here complete (100%) concordance was observed for all n = 143 horses genotyped. However, for a single individual discordance was observed between the SINE and the BIEC2-438999 SNP. Therefore the g.66493737C/T SNP may be a better indicator of the presence of the SINE. A separate study using a sample set of n = 31 Thoroughbred horses and 270 horses from 13 other breeds [29] found LD (r^2) of 0.93 in the Thoroughbred but a much lower LD (0.41) across all breeds. While multiple haplotypes were identified containing the C-allele without the SINE

TABLE 3: Regional distribution of *MSTN* genotype by best race distance (m) among elite race winners. Percentages are rounded to the nearest whole number

	1000–1200 m (%)	1201–1400 m (%)	1401–1600 m (%)	1601–1800 m (%)	1801–2000 m (%)	2001–2200 m (%)	2201–2400 m (%)	2400+ m (%)	N
Australia and New Zealand									
C:C	59	17	12	6	3	1	2	1	118
C:T	11	10	28	10	11	10	8	14	123
T:T	0	6	18	6	0	12	24	35	17
Europe and Middle East									
C:C	41	19	23	9	5	3	0	0	93
C:T	8	14	20	10	16	12	12	10	200
T:T	2	5	5	7	20	11	23	28	61
North America									
C:C	24	11	29	27	9	0	0	0	45
C:T	3	8	17	40	22	3	4	4	101
T:T	0	7	13	40	7	0	20	13	15
South Africa									
CC	55	10	20	0	10	0	5	0	20
C:T	9	12	25	16	28	0	6	3	32
T:T	0	0	10	20	20	0	20	30	10

TABLE 4: Mean best race distance (m) among elite race winners split by region and predicted distance categories

Region	CC-short			CC-long			CT-short			CT-long			TT-short			TT-long		
	Mean	s.e.	n	Mean	s.e.	n	Mean	s.e.	n	Mean	s.e.	n	Mean	s.e.	n	Mean	s.e.	n
Australia and New Zealand	1285	262	26	1342	363	68	1716	432	37	1989	632	44	1817	257	3	2740	521	6
Europe and Middle East	1286	258	47	1422	296	59	1786	444	134	2012	520	95	2107	587	28	2436	469	37
North America	1555	209	34	1541	255	13	1721	198	43	1728	246	58	1705	213	9	2073	380	10
South Africa	1339	342	14	1379	415	15	1724	388	21	1804	387	24	2133	602	6	2163	828	4

TABLE 5: Mean best race distance (m) among all winners split by region and predicted distance categories

Region	CC-short			CC-long			CT-short			CT-long			TT-short			TT-long		
	Mean	s.d.	n	Mean	s.d.	n	Mean	s.d.	n	Mean	s.d.	n	Mean	s.d.	n	Mean	s.d.	n
Australia and New Zealand	1236	244	65	1374	341	168	1592	358	120	1820	577	121	1788	546	16	2306	572	19
Europe and Middle East	1362	302	120	1517	357	175	1714	445	446	1955	504	229	2107	599	108	2262	528	110
North America	1454	241	48	1436	238	64	1630	245	95	1668	280	134	1643	211	28	1869	505	26
South Africa	1246	233	56	1446	417	29	1465	294	98	1657	373	65	1674	387	17	2085	600	14

insertion there was no instance of the SINE insertion and the T-allele within the same haplotype indicating that the original SINE insertion event occurred within a haplotype containing the g.66493737C/T SNP C-allele and recent intense selection for this haplotype in the Thoroughbred has resulted in reduced haplotypic diversity in the presence of the SINE insertion and the C-allele [30]. LD patterns are likely to vary across populations dependent on the proportion of C- and T-alleles in the population. The observation of high LD in this region, and the extremely long haplotypes, indicates that multiple genetic markers may be indicative of the presence of the SINE. Studies have indicated that the presence of

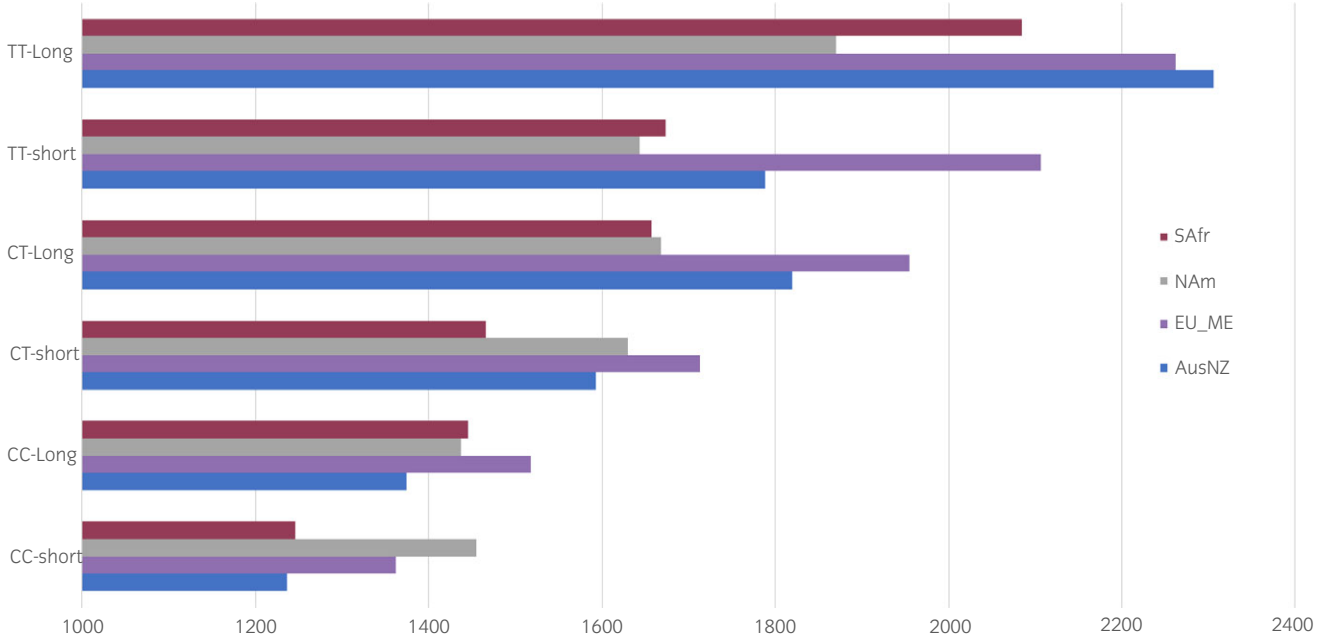

Fig 3: The average best race distances (m) for all winners in each long and short cohort in each region.

TABLE 6: Regional distribution of predicted distance cohort by best race distance among all race winners

	1000–1200 m (%)	1201–1400 m (%)	1401–1600 m (%)	1601–1800 m (%)	1801–2000 m (%)	2001–2200 m (%)	2201–2400 m (%)	2400+ m (%)	n
Australia and New Zealand									
CC-short	68	11	17	3	2	0	0	0	65
CC-long	48	22	16	6	2	3	2	1	168
CT-short	17	23	28	9	9	9	2	2	120
CT-long	17	10	25	12	9	9	7	12	121
TT-short	12	19	25	12	6	6	6	12	16
TT-long	0	5	11	0	21	16	21	26	19
Europe and Middle East									
CC-short	41	22	21	8	5	1	2	0	120
CC-long	25	21	25	10	9	6	4	1	175
CT-short	13	19	22	11	14	9	6	6	446
CT-long	8	10	15	9	17	12	14	16	229
TT-short	4	11	9	11	14	12	18	21	108
TT-long	1	5	8	6	15	9	24	32	110
North America									
CC-short	27	17	25	27	4	0	0	0	48
CC-long	30	22	23	20	5	0	0	0	64
CT-short	11	11	22	41	14	1	0	1	95
CT-long	6	16	18	40	10	3	4	2	134
TT-short	4	11	21	54	7	4	0	0	28
TT-long	4	8	19	35	8	4	12	12	26
South Africa									
CC-short	68	9	20	2	2	0	0	0	56
CC-long	48	14	21	0	7	3	3	3	29
CT-short	32	16	30	13	8	0	0	1	98
CT-long	14	18	26	15	18	0	3	5	65
TT-short	24	6	18	12	35	0	6	0	17
TT-long	0	7	21	21	7	0	14	29	14

the SINE, directly assayed and indicated by SNPs in the region, is responsible for variation in *MSTN* gene expression [28,31] resulting in muscle fibre type variation [29,32], precocity [4] and rates of differential growth [5,33] in Thoroughbreds, which consequently influences measured speed variables [8] and distance aptitude at various stages of racing.

We have shown here, in agreement with previous work discussed above, that although genomic variation at the *MSTN* locus is the main determinant of best race distance, there are other genetic markers that also contribute to the trait. Metabolic requirements and physiological responses differ considerably for endurance and shorter distance intense exercise. Endurance is associated with increased mitochondrial abundance and a shift in substrate preference from carbohydrates to fatty acids; conversely, during intense sprint exercise there is an increased reliance on anaerobic metabolism and localised hypoxia may occur in muscle. Based on the GWAS analysis we identified three candidate genes that may contribute to optimal race distance through modulation of substrate availability and the response to exercise induced hypoxia. The solute carrier family 7 member 5 (*SLC7A5*) gene, the ETS proto-oncogene 1, transcription factor (*ETS1*) gene and the peroxisome proliferator activated receptor gamma (*PPARG*) gene are located on ECA3, ECA7 and ECA16, respectively—genomic regions that contained two or more of the top 100 SNPs in the GWAS analysis. These genes are linked through a mechanistic interaction with the mammalian/mechanistic target of rapamycin complex 1 (mTORC1) complex which is a master regulator of anabolic processes including the rate of protein synthesis and the anabolic response to resistance exercise [34,35]. The expression of *SLC7A5* increases following exercise [36,37] and is modulated by insulin concentrations and hypoxia [38,39]. The *SLC7A5* gene product is central to the regulation of protein translation and cell proliferation via the activation of the mTORC1 signalling pathway [40,41]. *PPARG* is a downstream target of mTORC1, modulates fat metabolism and can inhibit *ETS1* gene expression [42]. In racehorses *PPARG* gene expression is significantly differentially expressed in skeletal muscle in response to training [43]. *ETS1* knockdown and over expression studies in human cell lines identified differential expression of genes in metabolic and oxidative stress response pathways. The *ETS1* knockdown model had increased oxygen consumption as measured by high resolution respirometry while overexpression of the ETS1 transcription factor resulted in increased expression of genes involved in glycolysis [44].

Conclusion

The training and preparation of Thoroughbreds for racing is a multifaceted process and includes selection of the most suitable race for an individual horse. Race distance aptitude is generally based on evaluation of the recorded race performance of relatives in the horse's pedigree, assessment of conformation and other physical characteristics, trainer observations and exercise rider feedback during training, jockey opinion following a race and opportunity arising from the race pattern that varies temporally due to the restriction of distances for 2-year-old races and spatially due to higher value races at certain distances specific to geographic racing region. We have provided here substantial evidence that genetic variation at the *MSTN* locus is the major determinant of race distance aptitude and have described the additional genomic variants that have a minor but measurable contribution to the distance trait. Based on results from the present study and previous research work, we suggest that the body of knowledge within the breeding and racing industry can now be augmented with evidence-based genomic prediction tools, leading to more accurate placing of horses in races that are best suited to their genetic potential. It would also be prudent to monitor genetic variation in the global population and refine practice towards the future sustainability of the Thoroughbred.

Authors' declaration of interests

The "Speed Gene" (MSTN) is the subject of multiple granted (NZ591711, EP2352850, JP5667057, US8771943, US9249470, AU2009290452) and pending (US2016215335, EP17190252) patents. E.W. Hill, L.M. Katz and D.E. MacHugh are inventors on the patents and receive royalties from a licence agreement between University College Dublin and Plusvital Ltd from commercial genetic tests for the "Speed Gene". Plusvital Ltd offers a test ("Distance Plus") for other genetic markers identified in this manuscript. E.W. Hill is Chief Science Officer, E.W. Hill and D.E. MacHugh are shareholders and B.A. McGivney is employed by Plusvital Ltd. A. Parnell received payment for work on the project to develop the genomic prediction model.

Ethical animal research

Research ethics committee oversight not currently required by this journal: the study was performed on excess material from samples that were originally collected for horse management purposes. DNA samples were collected with owners' consent and horse owners gave their consent that they could be used in research.

Sources of funding

The research was funded by Plusvital Ltd and Science Foundation Ireland under Grant Number 11/PI/1166.

Acknowledgements

We acknowledge the agreement of horse owners for use of the genetic data in research.

Authorship

E.W. Hill designed the study, performed the study, interpreted the data and prepared the manuscript. B.A. McGivney performed data analysis and interpretation and prepared the manuscript. M.F. Rooney performed data analysis. L.M. Katz interpreted the data. A. Parnell performed data analysis and prepared the manuscript. D.E. MacHugh interpreted the data. All authors approved the final manuscript.

References

1. Racing Australia. (2016) Australian Group Races.
2. British Horse Racing Authority. (2017) UK Flat Pattern 2017.
3. Horse Racing Ireland. (2017) Irish Pattern Races, Listed Races and Premier Handicaps 2017.
4. Farries, G., McGettigan, P.A., Gough, K.F., McGivney, B.A., MacHugh, D.E., Katz, L.M. and Hill, E.W. (2018) Genetic contributions to precocity traits in racing Thoroughbreds. *Anim. Genet.* **49**, 193-204.
5. Hill, E.W., Gu, J., Eivers, S.S., Fonseca, R.G., McGivney, B.A., Govindarajan, P., Orr, N., Katz, L.M. and MacHugh, D. (2010) A sequence polymorphism in MSTN predicts sprinting ability and racing stamina in thoroughbred horses. *PLoS ONE* **5**, e8645.
6. Hill, E.W., Ryan, D.P. and MacHugh, D.E. (2012) Horses for courses: a DNA-based test for race distance aptitude in thoroughbred racehorses. *Recent Pat. DNA Gene Seq.* **6**, 203-208.
7. Webb-Carter, C. (2015) A Study into British Stayers and Staying Races, The Thoroughbred Breeders' Association. Available at: https://www.thetba.co.uk/wp-content/uploads/2015/04/A-TBA-Study-into-the-Future-of-British-Stayers-and-Staying-Races.pdf (Accessed 1 August 2017).
8. Hill, E.W., Fonseca, R.G., McGivney, B.A., Gu, J., MacHugh, D.E. and Katz, L.M. (2012) MSTN genotype (g.66493737C/T) association with speed indices in Thoroughbred racehorses. *J. Appl. Physiol.* **112**, 86-90.
9. Hill, E.W., McGivney, B.A., Gu, J., Whiston, R. and MacHugh, D.E. (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. *BMC Genom.* **11**, 552.
10. Tozaki, T., Hill, E.W., Hirota, K., Kakoi, H., Gawahara, H., Miyake, T., Sugita, S., Hasegawa, T., Ishida, N., Nakano, Y. and Kurosawa, M. (2012) A cohort study of racing performance in Japanese Thoroughbred racehorses using genome information on ECA18. *Anim. Genet.* **43**, 42-52.
11. Meuwissen, T.H., Hayes, B.J. and Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-1829.

12. Corbin, L.J., Blott, S.C., Swinburne, J.E., Vaudin, M., Bishop, S.C. and Woolliams, J.A. (2010) Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Anim. Genet.* **41**, *Suppl.* **2**, 8-15.
13. Kim, E.S. and Kirkpatrick, B.W. (2009) Linkage disequilibrium in the North American Holstein population. *Anim. Genet.* **40**, 279-288.
14. Qanbari, S., Pimentel, E.C., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R. and Simianer, H. (2010) The pattern of linkage disequilibrium in German Holstein cattle. *Anim. Genet.* **41**, 346-356.
15. Hill, E.W., McGivney, B.A., Gu, J.J., Whiston, R. and MacHugh, D.E. (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. *BMC Genom.* **11**, 9.
16. Visscher, P.M., Hemani, G., Vinkhuyzen, A.A., Chen, G.B., Lee, S.H., Wray, N.R., Goddard, M.E. and Yang, J. (2014) Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10**, e1004269.
17. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Human Genet.* **88**, 76-82.
18. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289-300.
19. Aulchenko, Y.S., de Koning, D.J. and Haley, C. (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577-585.
20. Breiman, L. (2001) Random forests. *Mach. Learn.* **45**, 5-32.
21. Hastie, T.T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*, Springer-Verlag New York Inc., New York, NY.
22. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. and Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Human Genet.* **81**, 559-575.
23. Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, L.S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I. and Mountain, J. (2010) Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6**, e1000993.
24. Liu, F., Wollstein, A., Hysi, P.G., Ankra-Badu, G.A., Spector, T.D., Park, D., Zhu, G., Larsson, M., Duffy, D.L., Montgomery, G.W., Mackey, D.A., Walsh, S., Lao, O., Hofman, A., Rivadeneira, F., Vingerling, J.R., Uitterlinden, A.G., Martin, N.G., Hammond, C.J. and Kayser, M. (2010) Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet.* **6**, e1000934.
25. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., Lambourne, J.J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J.R., Daugherty, L.C., Delaneau, O., Freson, K., Garner, S.F., Grassi, L., Guerrero, J., Haimel, M., Janssen-Megens, E.M., Kaan, A., Kamat, M., Kim, B., Mandoli, A., Marchini, J., Martens, J.H.A., Meacham, S., Megy, K., O'Connell, J., Petersen, R., Sharifi, N., Sheard, S.M., Staley, J.R., Tuna, S., van der Ent, M., Walter, K., Wang, S.Y., Wheeler, E., Wilder, S.P., Iotchkova, V., Moore, C., Sambrook, J., Stunnenberg, H.G., Di Angelantonio, E., Kaptoge, S., Kuijpers, T.W., Carrillo-de-Santa-Pau, E., Juan, D., Rico, D., Valencia, A., Chen, L., Ge, B., Vasquez, L., Kwan, T., Garrido-Martin, D., Watt, S., Yang, Y., Guigo, R., Beck, S., Paul, D.S., Pastinen, T., Bujold, D., Bourque, G., Frontini, M., Danesh, J., Roberts, D.J., Ouwehand, W.H., Butterworth, A.S. and Soranzo, N. (2016) The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415-1429 e1419.
26. Pickrell, J.K., Berisa, T., Liu, J.Z., Segurel, L., Tung, J.Y. and Hinds, D.A. (2016) Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709-717.
27. Hagenaars, S.P., Hill, W.D., Harris, S.E., Ritchie, S.J., Davies, G., Liewald, D.C., Gale, C.R., Porteous, D.J., Deary, I.J. and Marioni, R.E. (2017) Genetic prediction of male pattern baldness. *PLoS Genet.* **13**, e1006594.
28. Santagostino, M., Khoriauli, L., Gamba, R., Bonuglia, M., Klipstein, O., Piras, F.M., Vella, F., Russo, A., Badiale, C., Mazzagatti, A., Raimondi, E., Nergadze, S.G. and Giulotto, E. (2015) Genome-wide evolutionary and functional analysis of the Equine Repetitive Element 1: an insertion in the myostatin promoter affects gene expression. *BMC Genet.* **16**, 126.
29. Petersen, J.L., Valberg, S.J., Mickelson, J.R. and McCue, M.E. (2014) Haplotype diversity in the equine myostatin gene with focus on variants associated with race distance propensity and muscle fiber type proportions. *Anim. Genet.* **45**, 827-835.
30. Bower, M.A., McGivney, B.A., Campana, M.G., Gu, J., Andersson, L.S., Barrett, E., Davis, C.R., Mikko, S., Stock, F., Voronkova, V., Bradley, D.G., Fahey, A.G., Lindgren, G., MacHugh, D.E., Sulimova, G. and Hill, E.W. (2012) The genetic origin and history of speed in the Thoroughbred racehorse. *Nat. Commun.* **3**, 643.
31. McGivney, B.A., Browne, J.A., Fonseca, R.G., Katz, L.M., Machugh, D.E., Whiston, R. and Hill, E.W. (2012) MSTN genotypes in Thoroughbred horses influence skeletal muscle gene expression and racetrack performance. *Anim. Genet.* **43**, 810-812.
32. Rooney, M.F., Porter, R.K., Katz, L.M. and Hill, E.W. (2017) Skeletal muscle mitochondrial bioenergetics and associations with myostatin genotypes in the Thoroughbred horse. *PLoS ONE* **12**, e0186247.
33. Tozaki, T., Sato, F., Hill, E.W., Miyake, T., Endo, Y., Kakoi, H., Gawahara, H., Hirota, K., Nakano, Y., Nambo, Y. and Kurosawa, M. (2011) Sequence variants at the myostatin gene locus influence the body composition of Thoroughbred horses. *J. Vet. Med. Sci.* **73**, 1617-1624.
34. Bodine, S.C., Stitt, T.N., Gonzalez, M., Kline, W.O., Stover, G.L., Bauerlein, R., Zlotchenko, E., Scrimgeour, A., Lawrence, J.C., Glass, D.J. and Yancopoulos, G.D. (2001) Akt/mTOR pathway is a crucial regulator of skeletal muscle hypertrophy and can prevent muscle atrophy in vivo. *Nat. Cell Biol.* **3**, 1014-1019.
35. Goodman, C.A., Frey, J.W., Mabrey, D.M., Jacobs, B.L., Lincoln, H.C., You, J.S. and Hornberger, T.A. (2011) The role of skeletal muscle mTOR in the regulation of mechanical load-induced growth. *J. Physiol.* **589**, 5485-5501.
36. Mitchell, C.J., Zeng, N., D'Souza, R.F., Mitchell, S.M., Aasen, K., Fanning, A.C., Poppitt, S.D. and Cameron-Smith, D. (2017) Minimal dose of milk protein concentrate to enhance the anabolic signalling response to a single bout of resistance exercise; a randomised controlled trial. *J. Int. Soc. Sports Nutr.* **14**, 17.
37. Murakami, T. and Yoshinaga, M. (2013) Induction of amino acid transporters expression by endurance exercise in rat skeletal muscle. *Biochem. Biophys. Res. Commun.* **439**, 449-452.
38. Elorza, A., Soro-Arnaiz, I., Melendez-Rodriguez, F., Rodriguez-Vaello, V., Marsboom, G., de Carcer, G., Acosta-Iborra, B., Albacete-Albacete, L., Ordonez, A., Serrano-Oviedo, L., Gimenez-Bachs, J.M., Vara-Vega, A., Salinas, A., Sanchez-Prieto, R., Martin del Rio, R., Sanchez-Madrid, F., Malumbres, M., Landazuri, M.O. and Aragones, J. (2012) HIF2alpha acts as an mTORC1 activator through the amino acid carrier SLC7A5. *Mol. Cell* **48**, 681-691.
39. Walker, D.K., Drummond, M.J., Dickinson, J.M., Borack, M.S., Jennings, K., Volpi, E. and Rasmussen, B.B. (2014) Insulin increases mRNA abundance of the amino acid transporter SLC7A5/LAT1 via an mTORC1-dependent mechanism in skeletal muscle cells. *Physiol. Rep.* **2**, e00238.
40. Cormerais, Y., Giuliano, S., LeFloch, R., Front, B., Durivault, J., Tambutte, E., Massard, P.A., de la Ballina, L.R., Endou, H., Wempe, M.F., Palacin, M., Parks, S.K. and Pouyssegur, J. (2016) Genetic disruption of the multifunctional CD98/LAT1 complex demonstrates the key role of essential amino acid transport in the control of mTORC1 and tumor growth. *Cancer Res.* **76**, 4481-4492.
41. Milkereit, R., Persaud, A., Vanoaica, L., Guetg, A., Verrey, F. and Rotin, D. (2015) LAPTM4b recruits the LAT1-4F2hc Leu transporter to lysosomes and promotes mTORC1 activation. *Nat. Commun.* **6**, 7250.
42. Ogawa, D., Nomiyama, T., Nakamachi, T., Heywood, E.B., Stone, J.F., Berger, J.P., Law, R.E. and Bruemmer, D. (2006) Activation of peroxisome proliferator-activated receptor gamma suppresses telomerase activity in vascular smooth muscle cells. *Circ. Res.* **98**, e50-e59.
43. Bryan, K., McGivney, B.A., Farries, G., McGettigan, P.A., McGivney, C.L., Gough, K.F., MacHugh, D.E., Katz, L.M. and Hill, E.W. (2017) Equine skeletal muscle adaptations to exercise and training: evidence of differential regulation of autophagosomal and mitochondrial components. *BMC Genom.* **18**, 595.
44. Verschoor, M.L., Verschoor, C.P. and Singh, G. (2013) Ets-1 global gene expression profile reveals associations with metabolism and oxidative stress in ovarian and breast cancers. *Cancer Metab.* **1**, 17.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Supplementary Item 1: Samples and phenotypes.
Supplementary Item 2: Phenotype summary.
Supplementary Item 3: DNA, genotyping and quality control
Supplementary Item 4: Primer and assay details
Supplementary Item 5: Genomic prediction using Random Effects Random Forests (RERF) modelling.
Supplementary Item 6: SNP data.
Supplementary Item 7: Manhattan plots: best race distance for nonelite winners and average race distance for all raced horses.
Supplementary Item 8: Genes: best race distance for elite winners.
Supplementary Item 9: QQ plots showing the distribution of P-values: best race distance for elite winners; best race distance for nonelite winners; best race distance for all winners; average race distance for all raced horses.
Supplementary Item 10: Top ranking SNPs.
Supplementary Item 11: Boxplots showing the distribution of the best race distance for elite horses split by genotype and region.
Supplementary Item 12: Genotype discordance between the g.66493737C/T SNP and SNP BIEC2-438999 (0, 1, 2), n = 2371.