ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

doi: 10.1111/age.12798

A genomic prediction model for racecourse starts in the Thoroughbred horse

B. A. McGivney*, B. Hernandez[†‡], L. M. Katz[§], D. E. MacHugh[¶], S. P. McGovern*, A. C. Parnell[†††], H. L. Wiencko* and E. W. Hill*****

*Plusvital Ltd, The Highline, Dun Laoghaire Industrial Estate, Dun Laoghaire, Dublin, Ireland. †Prolego Scientific, Nova UCD, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland. ‡The Irish Longitudinal Study on Aging (TILDA), Trinity College Dublin, Dublin, D02 PN40, Ireland. §UCD School of Veterinary Medicine, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland. ¶UCD Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland. **UCD School of Agriculture and Food Science, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland. ††School of Mathematics and Statistics, Insight Centre for Data Analytics, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

Summary

Durability traits in Thoroughbred horses are heritable, economically valuable and may affect horse welfare. The aims of this study were to test the hypotheses that (i) durability traits are heritable and (ii) genetic data may be used to predict a horse's potential to have a racecourse start. Heritability for the phenotype 'number of 2- and 3-year-old starts' was estimated to be $h^2_m = 0.11 \pm 0.02$ ($n = 4499$). A genome-wide association study identified SNP contributions to the trait. The *neurotrimin* (*NTM*), *opioid-binding protein/cell adhesion molecule like* (*OPCML*) and *prolylcarboxypeptidase* (*PRCP*) genes were identified as candidate genes associated with the trait. *NTM* functions in brain development and has been shown to have been selected during the domestication of the horse. *PRCP* is an established expression quantitative trait locus involved in the interaction between voluntary exercise and body composition in mice. We hypothesise that variation at these loci contributes to the motivation of the horse to exercise, which may influence its response to the demands of the training and racing environment. A random forest with mixed effects (RFME) model identified a set of SNPs that contributed to 24.7% of the heritable variation in the trait. In an independent validation set ($n = 528$ horses), the cohort with high genetic potential for a racecourse start had significantly fewer unraced horses (16% unraced) than did low (27% unraced) potential horses and had more favourable race outcomes among those that raced. Therefore, the information from SNPs included in the model may be used to predict horses with a greater chance of a racecourse start.

Keywords behaviour, durability, equine, genome-wide association study, performance, random forest model, temperament

Introduction

Thoroughbred horses are bred for competitive racing and undergo intense training and conditioning regimes, commencing as early as 20 months of age. Maintenance of strength, physical soundness and desirable behavioural characteristics are critical to the longevity of a horse's racing career. In addition to injury, the high attrition rate among Thoroughbred horses in training may be due to the functional effects of performance-limiting health traits (i.e. recurrent laryngeal neuropathy, recurrent exertional rhabdomyolysis, exercise-induced pulmonary haemorrhage etc.), negative adaptive behaviour to the training environment and/or trainer/owner decision-making regarding a particular horse's perceived lack of potential racing ability. There is high variation in racing career length among Thoroughbreds with females tending to be retired to stud at a younger age than males and with only a relatively small number of males entering the breeding population.

A key milestone in the training/racing career of a Thoroughbred is making a racecourse start. Thoroughbred horses trained for flat racing may commence racing in their 2-year-old year, although there is significant variation in age at first start. Although many horses have their first

Address for correspondence

E. W. Hill, Plusvital Ltd., Dun Laoghaire Industrial Estate, Pottery Road, Dublin, Ireland.
E-mail: emmeline.hill@plusvital.com

Accepted for publication 05 March 2019

racecourse start at 2 years of age, some may not commence racing until 3–5 years of age (More 1999). Further studies indicate that the average age at first start for Thoroughbreds is two and a half years (Sobczynska 2010). Horses that make a racecourse start at a young age are described as 'precocious' with indications that early maturity may be influenced by variation at the *myostatin* gene (*MSTN*). Horses with the C/C and C/T genotypes (*MSTN:* g.6649373C>T SNP) have been shown to have better racing outcomes as 2-year-olds than do T/T horses (Hill *et al*. 2010, 2012, 2019).

Arguably, a more economically important trait is not the age at which the horse starts racing but whether the horse progresses to making a racecourse start at all, as this critically affects the opportunity for earnings and perceived value to the pedigree. A study of active Thoroughbred mares registered in the General Stud Book (United Kingdom and Ireland) in 1975 found that of the 9765 live foals born eligible for naming, 49% were actually named and commenced training and 38% had raced by the end of their fourth year of age (Jeffcott *et al*. 1982). A more recent study of a cohort of 1022 foals born in 1999 showed similar results as 2-year-olds, whereby 52% had entered training and 32% had raced. In the 1975 cohort, 8.6% of horses were exported, whereas in the 1999 study, 28% were exported and may have raced abroad. These studies indicate that less than half of the Thoroughbred population actually makes it to the racecourse (Wilsher *et al*. 2006). Although this limits the opportunity for racecourse earnings, some unraced horses, in particular females, enter the breeding population and have a reproductive career. Also, some stallions have a higher proportion of unraced progeny, suggesting that there may be a genetic contribution to whether an individual is raced or unraced.

Racing durability or resilience in the Thoroughbred population has been described as a horse's ability to resist and withstand the rigours associated with training and racing (Velie *et al*. 2016) and is a desired trait with economic value. Durability traits include persistence (number of races), longevity (length of racing career) and frequency of races. Recently, heritability estimates in an Australian population of Thoroughbreds suggest that, among durability traits, racing longevity has the highest heritability ($h^2 = 0.12$). In addition, the heritability of racing persistence was estimated as 0.10. It is also interesting to note that the heritability of harness race starts among the Spanish Trotter horse population has been estimated as 0.17 (Sole *et al*. 2016). A recent investigation of performance traits in Norwegian-Swedish Trotters identified multiple candidate genes related to neural regulation (Velie *et al*. 2018); of particular interest were the *glutamate ionotropic receptor NMDA type subunit 2B* (*GRIN2B*) and *potassium channel regulator* (*KCNRG*) genes, which are both involved in memory and learning (Tang *et al*. 1999; Zamanillo *et al*. 1999; Lovell *et al*. 2013). These results suggest that, in addition to the physiological requirement for elite sporting performance, the ability to learn and adapt to the rigours of competitive racing are also critical to success.

The primary aim of this study was to test the hypothesis that measurable genomic variation contributes to the heritability of durability traits in the Thoroughbred population and to identify candidate genes that may contribute to variation in the trait. Subsequently, in an independent analysis we developed a genomic prediction model for the potential to race as a 2- or 3-year-old and evaluated the accuracy of the prediction model in ascertaining the likelihood of a horse being raced. The two approaches are entirely separate. The genome-wide association study (GWAS) was used to identify regions of the genome that are statistically significantly associated with the trait (after multiple comparison correction) to provide functional relevance for phenotypic effects. On the other hand, the prediction model had no requirement for statistical significance (or pruning of SNPs) and was used to provide the best possible prediction of the trait. An additional benefit of also employing a prediction model approach is that SNP feature importances are created that may provide further support for (or contradiction against) the SNPs identified in a GWAS.

Methods

Samples and phenotypes

Thoroughbred horse DNA samples ($n = 4499$) were collected with owners' consent and approval for use in research. Race records were obtained from Arion Pedigrees Ltd. for horses that had been bred or trained for racing and were at least 3 years of age at the date the race records were retrieved. The year of birth of the horses ranged from 1974 to 2013 (Table S1). Samples were collected from the global Thoroughbred population of 16 different countries and included 827 sires and 3487 dams (Tables S2 & S3, Fig. S1a,b). The number of lifetime starts was recorded for PHEN01. Horses were classified in a binary assignment as unraced (no starts) or raced (≥1 start) for PHEN02. The number of days between first and last start was recorded to define longevity of racing for horses with one or more (≥1) starts (PHEN03). Unraced horses were not included in PHEN03, and for horses raced only once, the number of days was recorded as 0. The number of racing seasons was also recorded. The number of starts in the 2-year-old and 3-year-old racing seasons was recorded separately (PHEN04), as this was considered to be the most economically relevant timeframe for a horse to have at least one racecourse start (Table 1).

Validation sample panel

An independent panel of samples was collected from 528 yearling horses in advance of yearling sales in Australia in

Table 1 Description of phenotypes and number of horses in each category used in the analyses.

Phenotype	Description	Unraced	Raced	Total
PHEN01 (continuous)	Durability—number of starts	384	3095	3479
PHEN02 (binary)	Durability—raced or unraced	384	3095	3479
PHEN03 (continuous)	Longevity—number of days between first and last start	—	3044	3044
PHEN04 (continuous)	Race ready—number of starts in 2-year-old and 3-year-old seasons	664	3835	4499

2012–2013. The total number of 2- and 3-year-old starts and total earnings was recorded for the horses at the end of their 3-year-old racing year (2016), with horses categorised on this basis as 'raced' or 'unraced'.

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). Samples and SNPs were included that had a genotyping rate greater than 95% and minor allele frequency greater than 5% respectively. A set of 48 896 informative SNPs originally derived from the SNP50 and SNP70 arrays was used for the analysis. This SNP set was extracted from the genotype data from each of the three arrays. SNPs that failed quality control or were not present on one of the array platforms were imputed using the software program BEAGLE (version 3.3.2; Browning & Browning 2016). For 10 horses genotyped using both the SNP50 and SNP70 arrays and 10 different horses separately genotyped using the SNP70 and SNP670 array post-imputation concordance was greater than 99%. For cases in which the *MSTN:*g.6649373C>T SNP was not present on the array or failed quality control, horses were genotyped for this SNP using a custom Taqman® assay (Life Technologies).

Heritability estimation

The genomic-relatedness-based restricted maximum likelihood (GREML) method within GENOME-WIDE COMPLEX TRAIT ANALYSIS (GCTA, version 1.24.2; Yang *et al.* 2011) was used to estimate genomic heritability (h^2_m). Significance was calculated using the standard mixed effects regression log likelihood ratio test evaluated under the null and alternative hypothesis where the null hypothesis was that the contribution of genetic variation to the heritability of the trait is 0 (Visscher *et al.* 2014). Sex, *MSTN* (g.66493737C/T) genotype (which was considered a fixed effect) and month of birth corrected for hemisphere (MOBc) were included as covariates in the estimation of h^2_m, both separately and together in a final model for PHEN01, PHEN02, PHEN03 and PHEN04 with the number of racing seasons included as a covariate for PHEN01 and PHEN03.

Genome-wide association study

A GWAS was performed using a linear model with covariates for PHEN04. This was 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 (Price *et al.* 2006). Sex, *MSTN* (g.66493737C/T) genotype and MOBc were included as covariates in the analyses, as there is documented evidence of their effects on Thoroughbred racing performance (More 1999; Hill *et al.* 2010, 2012). The inbreeding coefficient was also included as a covariate, as it has been associated with performance and health traits in other horse breeds (Klemetsdal 1998; Gibbons 2014). 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; Purcell *et al.* 2007). The parameters used were a window size of 50, a sliding window shift of 5 and a VIF threshold of 5, resulting in the pruned dataset. This SNP set was used to calculate the estimate of genomic inbreeding (F) using the hom function in GENABEL.

Genomic prediction using random forests with mixed effects

A prediction model was fitted for PHEN04 using a standard two-step approach with minor adjustments. Following Aulchenko *et al.* (2007), we first fitted a fixed effects model to account for SNP effects and subsequently accounted for random effects via relatedness using the residuals of this model. Such an approach is commonly known as restricted maximum likelihood (REML; Pinheiro & Bates 2000), which produces unbiased estimates of the variance components.

Although most genomic prediction approaches (e.g. GBLUP) use linear models for the fixed effects, prediction performance may be vastly improved by replacing the linear fixed effects component with a non-linear model. We used random forests, which can automatically account for possible additive, dominance, and SNP interaction effects through the use of bagged regression trees. Random forest approaches are increasingly being applied in genomic studies (Chen & Ishwaran 2012; Winham *et al.* 2012; Hill *et al.* 2019).

We applied the following steps to produce predictions of PHEN04:

1 The dataset (excluding the independent validation set) was divided into a training set and a test set of 75% and 25% of horses respectively. The training and test sets were stratified to include equivalent proportions of the *MSTN*:g.6649373C and T SNP genotypes, given that *MSTN* genotype may influence precocity.
2 For the training set, a random forest model was fitted to the square root of the continuous phenotype PHEN04 with sex, genomic inbreeding coefficient, month of birth and SNP data as features in the model.
3 Residuals for the training set were calculated using the out-of-bag predictions from the random forests, which protects against over-fitting of the method.
4 A multivariate normal distribution was applied to the training residuals to estimate the components of variation corresponding to that associated with genomic relatedness and that due to pure error respectively.
5 The random forest model was used to predict PHEN04 for the test set, and the multivariate normal distribution was used to predict the leftover variation in PHEN04 based on the genomic relatedness between the training and test sets. The random forest and multivariate normal predictions were added together and subsequently squared (because PHEN04 was square-rooted in step 2) to produce an overall prediction of PHEN04.
6 The square of the correlation coefficient between the true test PHEN04 values and the predicted PHEN04 values was used as an out-of-bag R^2 estimate to judge the performance of the model.
7 The entire model run (steps 1–5) was repeated on the full dataset (i.e. no training/test split) to produce estimated important SNPs (via the standard random forest variable importance score) and to provide predictions for future horses.

For the multivariate normal distribution fit detailed in step 4, the training residuals (r_i) were used in the following model:

$$r_i = \mu + g_i + \epsilon_i,$$

where μ is the overall mean of the residuals; g_i is the genetic random effect, given am $MVN(0, \sigma_g^2\Sigma)$ distribution, with Σ the known genomic relatedness matrix; and ϵ_i is a residual term given a $N(0, \sigma_\varepsilon^2)$ distribution. The variance terms σ_g^2 and σ_ε^2 quantify the contribution of genetic and residual variation respectively. These parameters were estimated via maximum likelihood based on the training set and subsequently used to make predictions on the test set using the standard multivariate normal formula:

$$\hat{r}_{\text{test}} = \mu + \sum\nolimits_{\text{train,test}}^{T} \sum\nolimits_{\text{train}}^{-1} (r_{\text{train}} - \mu),$$

where $\hat{r}_{\text{test}}$ is the estimated residuals for the test set, $\Sigma_{\text{train,test}}$ is the genomic relatedness matrix between the training and test sets, Σ_{train} is the genomic relatedness matrix for the training set and r_{train} is the vector of training residuals.

Genotypes for 4499 horses were available for the model development, and a standard 75%/25% training/test split of 3374/1125 horses was applied. The training and test sets were stratified to be representative of the *MSTN* genotypes, as variation at this locus has been suggested to influence precocity (Tozaki *et al*. 2011). The final model run used 48 896 SNPs and covariates representing sex, *MSTN* genotype, month of birth (corrected for hemisphere) and genomic inbreeding coefficient. The inbreeding coefficient was standardised by subtracting the mean (0.012) and dividing by the standard deviation (0.041).

Score assignment

The genomic prediction scores for the test set ($n = 1125$) were divided into duodeciles with horses assigned to one of the 12 categories based on the score. The correlation between the number of 2- and 3-year-old starts and the proportion of unraced horses in each duodecile was calculated. The reason for doing this was to determine whether the model developed based on number of races could be used to inform whether a horse is likely to be raced or unraced. To investigate this further we *post-hoc* adjusted cut-off levels used to classify horses by changing the proportion of unraced to match the population level of approximately 33%. We achieved this by repeatedly discarding a random set of raced horses so that the 33% unraced proportion was obtained, estimating predicted phenotype scores using the random forest with mixed effects (RFME) model, and creating new cut-offs for this sampled dataset. After 10 000 iterations, we averaged the cut-off scores and used these for future prediction of duodeciles for new horses. For simplicity of explanation compared to the duodeciles, we collapsed the population instead into thirds to identify horses with a high, medium or low probability of having at least one racecourse start as a 2- or 3-year-old.

Bioinformatics and candidate gene mining

Gene clusters were extracted from loci identified among the top 100 SNPs from the GWAS and RFME model. A locus was defined as one or multiple jointly associated SNPs within a region with all distances less than 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 (Smedley *et al*. 2015).

Results

Heritability

Marker or chip heritability estimates were highest for PHEN01 and PHEN04 ($h_m^2 = 0.11 \pm 0.02$). The lowest heritability estimate was observed for PHEN03 ($h_m^2 = 0.01 \pm 0.02$); see Table 2. With the exception of PHEN04, the estimated heritabilities increased with the

Table 2 Estimated heritability (h^2_m) of each phenotype with co-variates.

Phenotype	Co-variates	*n*	h^2_m	*P*-value
PHEN01 (continuous)	None	3479	0.084	7.40E-12
	Sex	3479	0.110	5.55E-17
	Sex, seasons	3479	0.059	4.32E-07
	MSTN SNP	3479	0.085	2.12E-12
	MOBc	3479	0.084	8.34E-12
	Final	3479	0.110	<1E-17
PHEN02 (binary)	None	3479	0.085	8.87E-14
	Sex	3479	0.089	1.60E-14
	MSTN SNP	3479	0.087	2.70E-14
	MOBc	3479	0.084	3.25E-13
	Final	3479	0.090	1.50E-14
PHEN03 (continuous)	None	3044	0.063	7.65E-07
	Sex, seasons	3044	0.011	2.02E-01
PHEN04 (continuous)	None	4499	0.105	<1E-17
	Sex	4499	0.106	<1E-17
	MSTN SNP	4499	0.108	<1E-17
	MOBc	4499	0.104	<1E-17
	Final	4499	0.108	<1E-17

MOBc refers to the month of birth corrected for hemisphere. The standard error for the estimates ranged from 0.016679 to 0.020519; rounded to two decimal places the standard error for all estimates was 0.02.

inclusion of the covariates sex, *MSTN* genotype and month of birth. For PHEN04, month of birth had no effect on the heritability estimate. Heritability was estimated using number of seasons and sex as covariates for PHEN01 and PHEN03. The inclusion of seasons did not increase heritability estimates for these phenotypes (PHEN01, $h^2_m = 0.06 \pm 0.02$; PHEN03, $h^2_m = 0.01 \pm 0.02$). Based on heritability estimates and economic relevance, PHEN04 (number of 2- and 3-year-old starts) was selected for the GWAS analysis and prediction model development.

Phenotypic variation

For PHEN04, 664 (14.7%) of the horses were unraced (0 starts) by the end of their 3-year-old racing year and the proportion of unraced horses was similar for males (14.5%) and females (14.8%). There was a wide range in the number of starts, although the majority (36.1%) had between six and 10 starts with less than 1% having had more than 25 starts. The largest number of starts ($n = 37$) was observed for a single horse (Fig. 1).

Identification of genetic contributions in a GWAS

In a GWAS for PHEN04 (number of 2- and 3-year-old starts), 14 SNPs reached significance at the genome-wide level following Bonferroni correction for multiple testing (modified threshold $P < 1.02 \times 10^{-6}$; see Fig. 2), with 126 SNPs significantly associated with the trait following correction using the Benjamini & Hochberg (1995) false discovery rate adjustment method (see Table S4 & Fig. 2). The top 14 SNPs were located on ECA7, 13 of which were contained within a region spanning 1.6 Mb (ECA7:40 217 482–41 818 480 bp). Loci were also identified and genes extracted among the top 100 SNPs (Table S5).

The top region in the GWAS (ECA7:40 217 482–41 818 480 bp) contained two genes, *neurotrimin* (*NTM*) and *opioid binding protein/cell adhesion molecule like* (*OPCML*). Among the top 100 SNPs there was also a 190-kb region on ECA7 that was physically separate (>20 Mb away) from the top GWAS region. This region contains the *DNA damage induced apoptosis suppressor* (*DDIAS*), *prolylcarboxypeptidase* (*PRCP*) and *RAB30, member RAS oncogene family* (*RAB30*) genes. One of the significant (Bonferroni) SNPs (*BIEC2-1004566*) is located in this region, 125 kb from the *PRCP* gene, which also contained 12 SNPs that were significant using the less stringent Benjamini-Hochberg adjustment. The QQ plot (Fig. S2) indicated a large contribution to the trait from SNPs that did not reach the threshold for significance in the GWAS, supporting our decision to use a genomic prediction method that can estimate all marker effects for all loci and capture even small genetic effects for a complex trait.

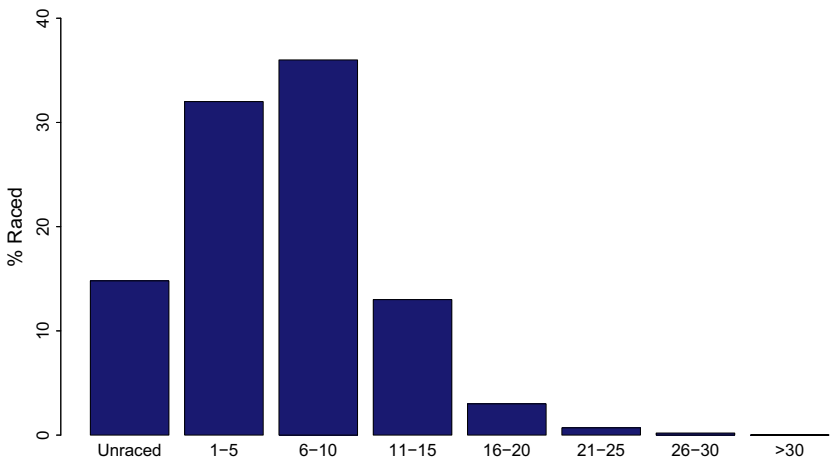

Figure 1 Proportion of total number of starts in 2-year-old and 3-year-old seasons for the study population for PHEN04.

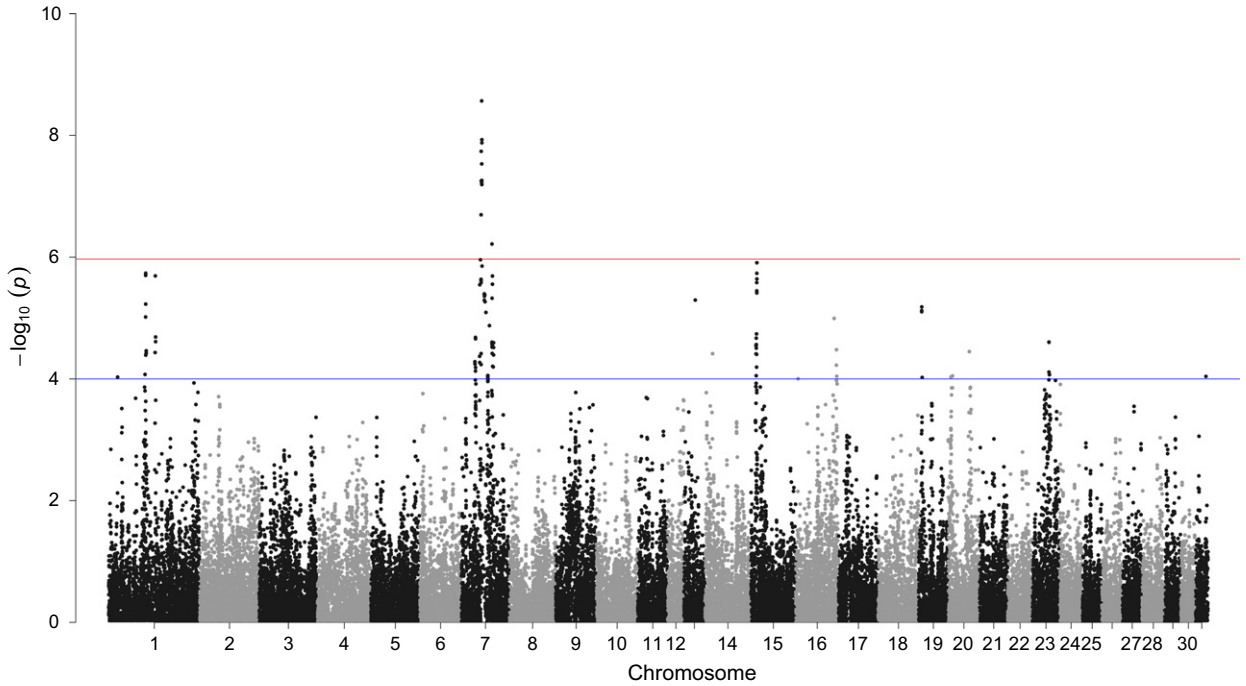

Figure 2 GWAS for PHEN04 (number of 2-year-old and 3-year-old starts). The red and blue lines indicate the *P*-value cut offs for genome-wide significance using Bonferonni and Benjamini Hochberg corrections respectively for multiple testing.

Genomic prediction using random forests with mixed effects

For the RFME model, the correlation between the actual and predicted phenotype for the test set was 0.17 with an out-of-bag R^2 of 2.7%. When corrected for heritability ($h_m^2 = 0.11$), the model identified a set of SNPs that contribute 24.7% of the heritable variation for the trait. When stratified by *MSTN* genotype, heritable variation for the trait was highest among *MSTN* C/T and C/C horses (43.0% and 25.2% respectively). A list of the top 100 SNPs in the RFME model and the genes identified within these regions are provided in Tables S6 & S7. The ECA7 regions identified among the top SNPs in the GWAS were also present among the highest ranked SNPs in the RFME method.

Score assignment

When the horses in the test set ($n = 1125$) were divided into duodeciles ranked on the basis of the genomic prediction score, there was a strong inverse relationship (−0.71, $R^2 = 0.51$) between the number of 2- and 3-year-old starts and the proportion of unraced horses in each duodecile. To accurately reflect the proportion of unraced horses in the population, cut-off levels were adjusted to reflect that approximately 33% of Thoroughbreds are unraced at the end of the 3-year-old season. For simplicity of explanation compared to the duodeciles, we collapsed the population instead into thirds, identifying horses with a high (74.1%), medium (69%) or low (56.9%) probability of having at least one racecourse start as a 2- or 3-year-old. High potential horses had a significantly higher probability of being raced than did low potential horses ($P = 3.0 \times 10^{-6}$; see Fig. 3).

Validation set

Of the 528 horses that were sampled in Australia as yearlings, 20% ($n = 105$) were unraced at the end of their 3-year-old year. By applying the prediction model for potential for a 2- or 3-year-old start to SNP genotypes for the horses, 16%, 22% and 27% of horses in each of the high, medium and low potential categories respectively were unraced (Table 3, Fig. 4). High potential horses had a significantly higher probability of being raced than did low potential horses ($P = 0.02$). In addition to running in significantly more races ($P = 5.5 \times 10^{-5}$), horses categorised as high potential also earned more than double race prize money than did low potential horses ($P = 0.006$; Figs. S3 & S4).

Discussion

In the present study, marker or SNP heritability was estimated for a number of durability traits in Thoroughbred horses. Heritability estimates were highest for PHEN01 (total number of starts) and PHEN04 (number of 2- and 3-year-old starts). Although not all the horses evaluated had completed their racing careers, all were at

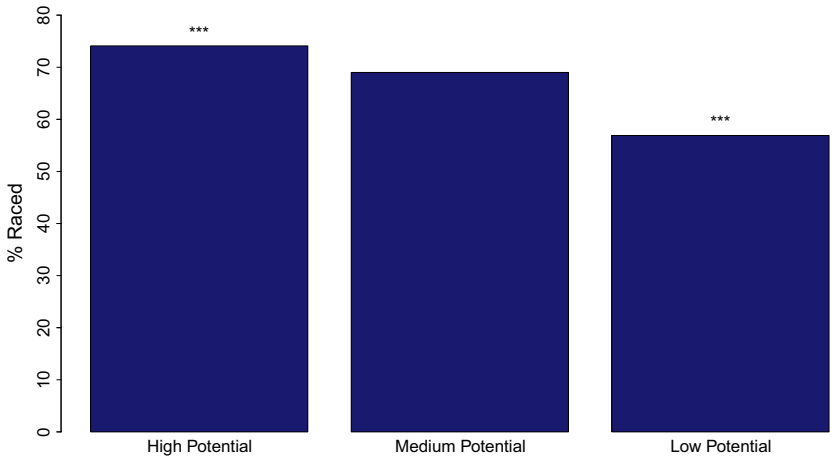

Figure 3 Proportion of raced horses in each category in the test set; $n = 352$ in each category (including unraced). There was a significant difference ($P = 3.0 \times 10^{-6}$) in the proportion of unraced low potential compared to high potential horses. $***P < 0.05$.

Table 3 Validation study indicating the percentage of horses assigned to each racing potential category and the racing performance of the horses in each category.

Racing potential	*N*	%	% Raced	% Unraced	% Wins	Ave races	Ave earnings	Ave sales price
High	249	47	84	16	48	6.1	34 769	129 140
Medium	205	39	78	22	43	5.6	26 804	117 946
Low	74	14	73	27	34	3.7	14 803	126 687
Total	528	100	80	20	44	5.6	28 878	124 548

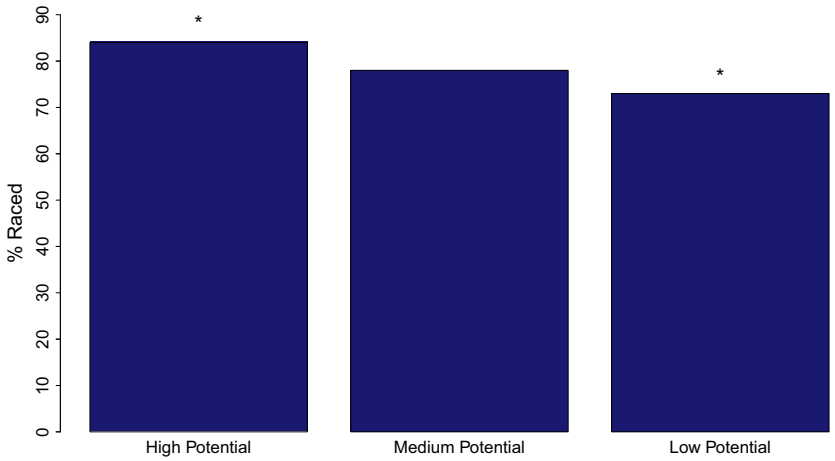

Figure 4 Proportion of raced horses in each category in the validation set. There was a significant difference ($P = 0.02$) in the proportion of unraced low potential compared to high potential horses. $*P < 0.05$.

least 3 years old, with the opportunity for racing therefore higher for the older horses. Although horses that were born earlier than 2000 were included in the analysis, the largest group of horses (20%) were born in 2009 and had four racing seasons of opportunity by 2015, when the race records were retrieved. We therefore included the number of racing seasons as a covariate, which decreased the heritability estimates. To prevent confounders as a result of number of racing seasons, we attempted to identify genetic predictors for a consistent number of seasons. We focused the analyses therefore on the number of 2- and 3-year-old starts (PHEN04) because this also is the most economically important period for flat racehorses. There was excellent agreement between the heritability estimates in this study and estimates for a range of durability traits previously generated from pedigree data (Velie *et al.* 2016). Racing persistence was the trait most similar between the two studies, with estimated heritabilities of 0.10 (Velie *et al.* 2016) and 0.11 (PHEN01 and PHEN04). These results demonstrate that racing persistence and the probability of a racecourse start are heritable in the Thoroughbred population; they also support the application of genomic prediction for genetic management of the population.

Identification of genetic contributions in a GWAS

We identified two loci on ECA7 that contained candidate genes that may contribute to the potential for racing as a 2- or 3-year-old. One locus was defined by the top 13 GWAS

SNPs and contained the *NTM* and *OPCML* genes. The SNP ranked fourth in the GWAS was *BIEC2-996632*, a *NTM* intronic C>T SNP. Although no exonic variants have been reported in this gene in the horse, insertions and copy number variants have been reported for *NTM* in the Quarter Horse (Doan *et al*. 2012). Interestingly, this region was identified as undergoing divergent selection between draft and light horses (Gurgul *et al*. 2019), and *NTM* was ranked ninth among 125 genes observed to have been positively selected during horse domestication, specifically the large subset linked to neurobiology and brain development (Schubert *et al*. 2014). Equine neurological systems perturbed by natural and artificial selection associated with domestication may therefore overlap with adaptive traits required for successful Thoroughbred racing careers. Recently, *Ntm* has been shown to influence behavioural traits in mice; specifically, *Ntm*-knockout mice were found to be deficient in emotional learning for specific tasks (Mazitov *et al*. 2017). In humans, *NTM* has been associated with lipid phenotypes (Li *et al*. 2015), childhood aggressiveness (Brevik *et al*. 2016), heart failure (Cao *et al*. 2015) and IQ (Pan *et al*. 2011). The *NTM* and *OPCML* genes are functionally related, both encoding members of the IgLON family of proteins that regulate neural growth and synapse formation. It has been suggested that they work together in complementary roles (Mazitov *et al*. 2017) and may share common promoters. Alternative promoter-driven expression of the IgLON gene family has been reported, and a complex functional relationship among the genes has been suggested (Vanaveski *et al*. 2017). In a recent transcriptomics study, using RNA-seq in equine skeletal muscle, the *NTM* gene was not differentially expressed in response to exercise or training whereas *OPCML* transcripts were significantly differentially expressed following a period of training (0.447 fold, $P = 0.0074$; Table S8; Bryan *et al*. 2017). Although the functional effects of *NTM* and *OPCML* in the horse are not fully understood, these results suggest that these genes could play a role in physiological or behavioural adaptations required for early racing and training.

The second locus identified on ECA7 contained three genes, including the *PRCP* gene. Previously, the *Prcp* gene has been implicated as an eQTL in mice, contributing to the interaction between voluntary exercise and body composition (Kelly *et al*. 2010, 2012, 2014); specifically, it has been shown to be associated with voluntary wheel running exercise in mice. The underlying basis for voluntary exercise is thought to be a complex interaction between central and peripheral nervous system components that contribute to both motivation and physical ability. Also *PRCP* deficiency has been shown to influence blood pressure and cardiac function (Tabrizian *et al*. 2015; Maier *et al*. 2017), and therefore this gene's function may have a direct effect on the physiological phenotype relevant to exercise. Gene expression data supports a role for *PRCP* in the exercise response in the horse: in a cohort of Thoroughbred horses in training, skeletal muscle RNA-seq analysis identified significant (-0.71 fold, $P = 1.41 \times 10^{-10}$) differential expression of *PRCP* transcripts following a single bout of exercise. *PRCP* gene transcripts were among the top 25% of 3241 significantly differentially expressed genes post-exercise. The expression of the gene was not significantly altered following a period of conditioning training (Bryan *et al*. 2017). Although the functional role of the *PRCP* gene in voluntary exercise is not fully understood, these results support the hypothesis that genomic variation associated with *PRCP* expression influences self-motivated exercise in Thoroughbred horses.

Temperament, including the 'will to win' on the racecourse and a horse's 'attitude' towards its exercise regime, is considered among the most important aspects to a positive outcome on the racetrack. The response to exercise training may be positive in that a horse is perceived to be enthusiastic for its work or may be negative with a horse described as being unmotivated. These responses may be interpreted as a motivational response to exercise. Thoroughbreds trained for flat racing may enter a training establishment at as young as 20 months of age and start training soon afterwards. Racehorse trainers observe the behaviour of their horses daily and tend to adapt the training and management protocol for each individual horse on the basis of their observations. There are reasons other than injury and/or clinical manifestation of a performance-limiting trait that dictate whether a horse is considered ready or not to either enter training and/or progress in a training programme, including being perceived as having a poor attitude towards exercise. For Thoroughbred racehorses in training, unlike humans, the type and intensity of exercise is not voluntary but is imposed by humans. Previously, a SNP in the *5-hydroxytryptamine receptor 1A* gene (*HTR1A*) has been shown, in a cohort of Thoroughbred racehorses, to be significantly associated with 'tractability', the ease with which animals can be trained and controlled (Hori *et al*. 2016). Therefore, with regards to the opportunity to race, we propose that behavioural aspects, underpinned by variation at genes associated with neurophysiological responses—including amongst others, the *NTM*/*OPCML* and *PRCP* gene loci—may be equally, if not more, important than physical attributes for elite athleticism.

Genomic prediction using random forests with mixed effects

In addition to identifying individual candidate genes, we have developed a prediction algorithm using 48 896 genome-wide SNPs to identify horses that are most likely to have a racecourse start in the most economically relevant period of their training/racing career. The model corresponds to a standard GBLUP run (fitted via REML) with the linear fixed effects portion replaced by a more flexible random forests approach.

When the model was used to generate predictions for both the test and validation sets, horses that were in the low potential category exhibited a 1.7-fold greater probability of being unraced than horses in the high potential category. The validation set comprised samples collected in advance of yearling sales in Australia in 2012–2013, with race records for the horses subsequently compiled in 2016 after the model had been developed using the training/test sets. Interestingly, not only did the low potential group have a higher proportion of unraced horses (27%) than the high potential cohort (16%), the horses in the low potential group that had raced also had poorer returns on the investment made. There was no significant difference in the average price paid for high potential (AUD 129K) and low potential horses (AUD 127K); however, the average earnings of the high potential horses were significantly more than the low potential cohort at AUD 35K vs. 15K respectively. This may reflect the significantly greater number of races competed by the high potential horses and the relative win percentages of 48% vs. 34% for high and low respectively.

The results presented here suggest that the traditional methods of evaluating a horse's potential for racing seem to be of limited value in discerning the population of horses with a higher potential to race, win and achieve earnings, as most horses at premier sales have already been pre-selected on the basis of pedigree and conformation. It should therefore be possible to introduce genomic assessment as an additional decision-making tool to help in identifying animals most likely to provide a return on investment. Post-purchase evaluation of a horse's potential to race may be used to modify the training environment to maximise the genetic potential of a horse by increasing motivation. Although it is not yet possible to accurately predict multi-locus genotype combinations in the next generation, breeders may use this information to favour breeding horses with the lowest probability of transmitting genomic variants that undermine the behavioural plasticity or temperament required for a successful racing career. The heritable component of this trait suggests that those mares in the breeding population that are unraced and have low genetic potential for racing are most likely to transmit unfavourable genomic variants for racing career success, regardless of the actual reason they were unraced. Therefore, selection of highly durable stallions as mate choices for mares with such a profile may be an appropriate strategy for breeders.

Conclusion

Our results indicate that genes that function in behavioural adaptations may be particularly important in the success of a racehorse. The introduction of genomic prediction tools in the Thoroughbred industry has considerable potential to improve management strategies. Not only may these results be applied in optimising the management environment for individual horses, but also these data could be readily implemented in industry-wide monitoring of the population to ensure long-term sustainability and durability. An understanding of genomic contributions to durability traits, and efficient and effective adoption of genomic prediction tools, may improve management practices towards the long-term sustainability of the breed.

Data availability

De-identified phenotypes and pre-computed univariate associations between genotype and phenotype is available in the supplementary materials or upon request via a material transfer agreement for research purposes only.

Conflicts of interest

This research was funded by Plusvital Ltd. DEM and EWH are shareholders in Plusvital Ltd. Plusvital Ltd contracted with AP and BH to perform some aspects of the analyses.

References

Aulchenko Y.S., de Koning D.J. & 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**, 577–85.

Benjamini Y. & Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

Brevik E.J., van Donkelaar M.M., Weber H. *et al.* (2016) Genome-wide analyses of aggressiveness in attention-deficit hyperactivity disorder. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* **171**, 733–47.

Browning B.L. & Browning S.R. (2016) Genotype imputation with millions of reference samples. *American Journal of Human Genetics* **98**, 116–26.

Bryan K., McGivney B.A., Farries G., McGettigan P.A., McGivney C.L., Gough K.F., MacHugh D.E., Katz L.M. & Hill E.W. (2017) Equine skeletal muscle adaptations to exercise and training: evidence of differential regulation of autophagosomal and mitochondrial components. *BMC Genomics*, **18**, 595.

Cao T.H., Quinn P.A., Sandhu J.K., Voors A.A., Lang C.C., Parry H.M., Mohan M., Jones D.J. & Ng L.L. (2015) Identification of novel biomarkers in plasma for prediction of treatment response in patients with heart failure. *Lancet* **385** (Suppl 1), S26.

Chen X. & Ishwaran H. (2012) Random forests for genomic data analysis. *Genomics* **99**, 323–9.

Doan R., Cohen N.D., Sawyer J., Ghaffari N., Johnson C.D. & Dindot S.V. (2012) Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare. *BMC Genomics* **13**, 78.

Gibbons A. (2014) Ancient DNA. The thoroughly bred horse. *Science* **346**, 1439.

Gurgul A., Jasielczuk I., Semik-Gurgul E., Pawlina-Tyszko K., Stefaniuk-Szmukier M., Szmatola T., Polak G., Tomczyk-Wrona I. & Bugno-Poniewierska M. (2019) A genome-wide scan for

diversifying selection signatures in selected horse breeds. *PLoS ONE* **14**, e0210751.

Hill E.W., Gu J., Eivers S.S., Fonseca R.G., McGivney B.A., Govindarajan P., Orr N., Katz L.M. & MacHugh D. (2010) A sequence polymorphism in *MSTN* predicts sprinting ability and racing stamina in thoroughbred horses. *PLoS ONE* **5**, e8645.

Hill E.W., Fonseca R.G., McGivney B.A., Gu J., MacHugh D.E. & Katz L.M. (2012) *MSTN* genotype (g.66493737C/T) association with speed indices in Thoroughbred racehorses. *Journal of Applied Physiology* **112**, 86–90.

Hill E.W., McGivney B.A., Rooney M.F., Katz L.M., Parnell A. & MacHugh D.E. (2019) The contribution of *myostatin* (*MSTN*) and additional modifying genetic loci to race distance aptitude in Thoroughbred horses racing in different geographic regions. *Equine Veterinary Journal*. https://doi.org/10.1111/evj.13058 [Epub ahead of print]

Hori Y., Tozaki T., Nambo Y., Sato F., Ishimaru M., Inoue-Murayama M. & Fujita K. (2016) Evidence for the effect of *serotonin receptor 1A* gene (*HTR1A*) polymorphism on tractability in Thoroughbred horses. *Animal Genetics* **47**, 62–7.

Jeffcott L.B., Rossdale P.D., Freestone J., Frank C.J. & Towers-Clark P.F. (1982) An assessment of wastage in thoroughbred racing from conception to 4 years of age. *Equine Veterinary Journal* **14**, 185–98.

\Kelly S.A., Nehrenberg D.L., Peirce J.L., Hua K., Steffy B.M., Wiltshire T., Pardo-Manuel de Villena F., Garland T. Jr & Pomp D. (2010) Genetic architecture of voluntary exercise in an advanced intercross line of mice. *Physiological Genomics* **42**, 190–200.

Kelly S.A., Nehrenberg D.L., Hua K., Garland T. Jr & Pomp D. (2012) Functional genomic architecture of predisposition to voluntary exercise in mice: expression QTL in the brain. *Genetics* **191**, 643–54.

Kelly S.A., Nehrenberg D.L., Hua K., Garland T. Jr & Pomp D. (2014) Quantitative genomics of voluntary exercise in mice: transcriptional analysis and mapping of expression QTL in muscle. *Physiological Genomics* **46**, 593–601.

Klemetsdal G. (1998) The effect of inbreeding on racing performance in Norwegian cold-blooded trotters. *Genetics Selection Evolution* **30**, 351.

Li C., Bazzano L.A., Rao D.C. *et al*. (2015) Genome-wide linkage and positional association analyses identify associations of novel *AFF3* and *NTM* genes with triglycerides: the GenSalt study. *Journal of Genetics and Genomics* **42**, 107–17.

Lovell P.V., Carleton J.B. & Mello C.V. (2013) Genomics analysis of potassium channel genes in songbirds reveals molecular specializations of brain circuits for the maintenance and production of learned vocalizations. *BMC Genomics* **14**, 470.

Maier C., Schadock I., Haber P.K. *et al*. (2017) Prolylcarboxypeptidase deficiency is associated with increased blood pressure, glomerular lesions, and cardiac dysfunction independent of altered circulating and cardiac angiotensin II. *Journal of Molecular Medicine* **95**, 473–86.

Mazitov T., Bregin A., Philips M.A., Innos J. & Vasar E. (2017) Deficit in emotional learning in neurotrimin knockout mice. *Behavioural Brain Research* **317**, 311–8.

More S.J. (1999) A longitudinal study of racing thoroughbreds: performance during the first years of racing. *Australian Veterinary Journal* **77**, 105–12.

Pan Y., Wang K.S. & Aragam N. (2011) *NTM* and *NR3C2* polymorphisms influencing intelligence: family-based association studies. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* **35**, 154–60.

Pinheiro J.C. & Bates D.M. (2000) *Mixed-Effects Models in S and S-PLUS*. Springer, New York, NY.

Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A. & Reich D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–9.

Purcell S., Neale B., Todd-Brown K. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–75.

Schubert M., Jonsson H., Chang D. *et al*. (2014) Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E5661–9.

Smedley D., Haider S., Durinck S. *et al*. (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research* **43**, W589–98.

Sobczynska M. (2010) Relationship between age at first start and racing performance in Polish Thoroughbreds and Arab horses. *Archiv Tierzucht* **53**, 639–49.

Sole M., Valera M., Gomez M.D., Solkner J., Molina A. & Meszaros G. (2016) Heritability and factors associated with number of harness race starts in the Spanish Trotter horse population. *Equine Veterinary Journal* **49**, 288–93.

Tabrizian T., Hataway F., Murray D. & Shariat-Madar Z. (2015) Prolylcarboxypeptidase gene expression in the heart and kidney: effects of obesity and diabetes. *Cardiovascular & Hematological Agents in Medicinal Chemistry* **13**, 113–23.

Tang Y.P., Shimizu E., Dube G.R., Rampon C., Kerchner G.A., Zhuo M., Liu G. & Tsien J.Z. (1999) Genetic enhancement of learning and memory in mice. *Nature* **401**, 63–9.

Tozaki T., Sato F., Hill E.W. *et al*. (2011) Sequence variants at the *myostatin* gene locus influence the body composition of Thoroughbred horses. *Journal of Veterinary Medical Science* **73**, 1617–24.

Vanaveski T., Singh K., Narvik J. *et al*. (2017) Promoter-specific expression and genomic structure of IgLON family genes in mouse. *Frontiers in Neuroscience* **11**, 38.

Velie B.D., Hamilton N.A. & Wade C.M. (2016) Heritability of racing durability traits in the Australian and Hong Kong Thoroughbred racing populations. *Equine Veterinary Journal* **48**, 275–9.

Velie B.D., Fegraeus K.J., Sole M., Rosengren M.K., Roed K.H., Ihler C.F., Strand E. & Lindgren G. (2018) A genome-wide association study for harness racing success in the Norwegian-Swedish coldblooded trotter reveals genes for learning and energy metabolism. *BMC Genetics* **19**, 80.

Visscher P.M., Hemani G., Vinkhuyzen A.A., Chen G.B., Lee S.H., Wray N.R., Goddard M.E. & Yang J. (2014) Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genetics* **10**, e1004269.

Wilsher S., Allen W.R. & Wood J.L. (2006) Factors associated with failure of Thoroughbred horses to train and race. *Equine Veterinary Journal* **38**, 113–8.

Winham S.J., Colby C.L., Freimuth R.R., Wang X., de Andrade M., Huebner M. & Biernacka J.M. (2012) SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics* **13**, 164.

Yang J., Lee S.H., Goddard M.E. & Visscher P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**, 76–82.

Zamanillo D., Sprengel R., Hvalby O. *et al.* (1999) Importance of AMPA receptors for hippocampal synaptic plasticity but not for spatial learning. *Science* **284**, 1805–11.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Shared (a) sires and (b) dams in training and validation sets.

Figure S2 QQ plot showing the distribution of *P*-values from the GWAS.

Figure S3 Number of starts in validation set split based on genomic prediction.

Figure S4 Total earnings in validation set split based on genomic prediction.

Table S1 Phenotypic data summarised by year.

Table S2 Regional distribution of samples.

Table S3 Sire representation in the dataset.

Table S4 GWAS results for PHEN04.

Table S5 Genes in regions identified from GWAS of PHEN04.

Table S6 Top SNPs contributing to random forest model.

Table S7 Genes in regions identified from top SNPs contributing to random forest model.

Table S8 Differential expression of candidate genes in response to exercise (Bryan *et al.* 2017).