ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

doi:10.1111/j.1365-2052.2010.02104.x

# Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance

**E. W. Hill*, J. Gu*, B. A. McGivney* and D. E. MacHugh*,†**

*Animal Genomics Laboratory, UCD School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland. †UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland

**Summary**

Athletic performance is influenced by a complex interplay among the environment and a suite of genes, which contributes to system-wide structure and function. In a panel of elite and non-elite Thoroughbred horses ($n$ = 148), we genotyped 68 SNPs in 17 putative exercise-relevant genes chosen from a genome scan for selection. We performed a series of case–control and quantitative association tests for relationships with racecourse performance. Thirteen SNPs in nine genes were significantly ($P < 0.05$) associated with a performance phenotype. We selected five SNPs in four genes (*ACSS1, ACN9, COX4I1, PDK4*) for validation in an independent sample set of elite and non-elite Thoroughbreds ($n$ = 130). Two SNPs in the *PDK4* gene were validated ($P < 0.01$) for associations with elite racing performance. When all samples were considered together ($n$ = 278), the PDK4_ 38973231 SNP was strongly associated ($P < 0.0005$) with elite racing performance. Individuals with the A:A and A:G genotypes had a 16.2–16.6 lb advantage over G:G individuals in terms of handicap rating. Re-sequencing of the *PDK4* gene and further genotyping will be required to identify the causative variant that is likely influencing exercise-induced variation in expression of the gene. Notwithstanding, this information may be employed as a marker for the selection of racehorses with the genetic potential for superior racing ability.

**Keywords** Thoroughbred, exercise, SNP, association test, gene.

## Introduction

Intense selection for athletic phenotypes in the Thoroughbred horse (*Equus caballus*) during the past 300 years has resulted in structural and functional system-wide adaptations that have significantly enhanced the physiological characteristics that enable elite athletic performance (Constantinopol *et al.* 1989; Jones *et al.* 1989; Evans *et al.* 1993). The athleticism of the Thoroughbred is attributed to a range of extreme physiological characteristics including a large muscle mass to body weight ratio, high skeletal muscle mitochondrial density and oxidative enzyme activity and considerable intramuscular stores of energy substrates (Hinchcliff *et al.* 2008).

While the phenotypic adaptations to elite athleticism in Thoroughbred horses are well described, the understanding of the molecular contributions to such exquisitely adapted exercise-related phenotypes is still in its infancy (Harrison & Turrion-Gomez 2006; Eivers *et al.* 2009; Gu *et al.* 2009; McGivney *et al.* 2009; Hill *et al.* 2010). Genetic contributions to athletic performance phenotypes in humans are well documented and more than 220 genes have been described (Bray *et al.* 2009). Although it is likely that Thoroughbred racing performance is also influenced by a large number of genes, only one performance-associated sequence variant in an exercise-relevant gene has previously been reported for the horse (Hill *et al.* 2010). The recent and strong selection for exercise-related traits has left signatures in the genome of the Thoroughbred. In a population genetics-based genome scan, positively selected loci have been identified in the extreme tail-ends of the distributions for statistics ($F_{ST}$ and Ewens–Watterson test) that identify departures from patterns of genetic variation expected under neutral genetic drift (Gu *et al.* 2009). Such outlier approaches have led to an understanding of the selective forces that have shaped the recent evolution of human populations (Akey 2009; Oleksyk *et al.* 2010; Pritchard

Address for correspondence

E. W. Hill, UCD School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland.
E-mail: emmeline.hill@ucd.ie



 

*et al.* 2010). Within the positively selected genomic regions, enrichments for genes involved in insulin signalling, fat substrate utilization and muscle strength have been identified. Genes in these functional categories likely play key roles in contributing to the lean, muscular, athletic phenotype that is typical for Thoroughbreds.

Genomic regions that have been targets for selection represent the most likely regions to contain structural genetic variation contributing to functional and phenotypic variance in exercise-relevant traits. Therefore, to identify genes that represent the most likely targets for selection, we investigated whether genes within these regions may contain sequence variants that contribute to the genetic variation in racetrack performance in the Thoroughbred population (Gaffney & Cunningham 1988).

We interrogated the EquCab2.0 SNP database for Thoroughbred SNPs located within the genomic sequence of 20 putative exercise-relevant genes located within the top-ranked outlier genomic regions previously described (Gu *et al.* 2009). A panel of 68 SNPs in 17 genes was selected for genotyping. To investigate associations between the sequence variants and racing phenotypes, we genotyped a group of Thoroughbred horses ($n$ = 148) and performed a series of population-based case–control genetic association investigations by separating the samples on the basis of retrospective racecourse performance. In addition, we performed quantitative trait association tests using best race distance and handicap rating (Racing Post Rating, RPR) as phenotypes.

## Materials and methods

### Genomic DNA samples

More than 1400 registered Thoroughbred horse tissue samples (hair or fresh blood) were collected from stud farms, racing yards and sales establishments in Ireland and New Zealand between 1997 and 2009. Genomic DNA was extracted from either fresh whole blood or hair samples using a modified version of a standard phenol/chloroform method (Sambrook & Russell 2001).

Horses were categorized based on retrospective racecourse performance records as 'Elite' Thoroughbreds (TBE) or 'Other' Thoroughbreds (TBO). To minimize confounding effects of racing over obstacles, only horses with performance records in Flat races were considered for inclusion in the study cohorts. Elite Thoroughbreds were Flat racehorses that had won at least one Group race (Group 1, Group 2 or Group 3) or a Listed race – the highest standard and most valuable elite Flat races are known as Group races and Listed races are the next in status. Other Thoroughbreds had competed in at least one race, but had never won a race and had handicap (Racing Post Rating, RPR) ratings <80. Race records were derived from three sources. Europe race records were derived from the Racing Post on-line database (http://www.racingpost.com); Australasia and South East Asia race records were derived from Arion Pedigrees (http://www.arion.co.nz); and North America race records were derived from the Pedigree Online Thoroughbred database (http://www.pedigreequery.com). In all cases, pedigree information was used to control for genetic background by attempting to exclude samples sharing relatives. Over-representation of popular sires within the pedigrees was avoided where possible.

### Study sample set

A panel of Thoroughbred samples ($n$ = 148) was selected from the repository and separated into two distinct performance cohorts; elite (TBE, $n$ = 86; mean RPR = 115) and other (TBO, $n$ = 62; mean RPR = 59) (Table 1). In each cohort, there was no sharing of sires or dams. The elite performer group contained horses (of which 84 were Group race winners) that competed in a total of 1170 races and won 425, including 215 Group races, of which 91 were at Group 1 level. The other (non-elite) cohort competed in 537 races and won 15, none of which was a Group race.

### Validation sample set

An additional $n$ = 130 Thoroughbred samples were genotyped for six SNPs. The samples were subdivided into elite ($n$ = 97; mean RPR = 113) and other ($n$ = 33; mean RPR = 49) cohorts. There was some sharing of sires within and among the validation set cohorts (Table 1).

**Table 1** Thoroughbred sample sets.

| | *n* | No. sires | No. Group race winners | Mean RPR | Range RPR |
|---|---|---|---|---|---|
| Study set | | | | | |
| TB | 148 | 136 | | | |
| TBE | 86 | 86 | 84 | 115 | 87–134 |
| TBO | 62 | 62 | 0 | 59 | 21–79 |
| Validation set | | | | | |
| TB | 130 | 81 | | | |
| TBE | 97 | 63 | 70 | 113 | 90–138 |
| TBO | 33 | 24 | 0 | 49 | 17–73 |
| All | | | | | |
| TB | 278 | 186 | | | |
| TBE | 183 | 128 | 175 | 113 | 84–138 |
| TBO | 95 | 77 | 0 | 53 | 17–79 |

The original sample set contained unrelated individuals that were categorized based on retrospective racing performance as elite (TBE) or other (TBO). The validation sample set contained some sharing of sires within each performance cohort.
RPR, Racing Post Rating handicap rating.

**Table 2** Candidate gene details.

| Gene symbol | Gene description | Gene ontology: Biological Process | KEGG pathway | Chr | Distance from marker (Mb) | Dh/sd[1] | *P* | $P_{ST}$[2] | *P* |
|---|---|---|---|---|---|---|---|---|---|
| *ACN9* | ACN9 homolog (*S. cerevisiae*) | GO:0005996~ monosaccharide metabolic process | | 4 | 1.66 | −6.12 | <0.001 | 0.45 | <0.01 |
| *ACSS1* | Acyl-CoA synthetase short-chain family member 1 | GO:0006732~ coenzyme metabolic process | hsa00010:Glycolysis/ Gluconeogenesis | 22 | 0.55 | NS | NS | 0.42 | <0.01 |
| *ACTA1* | Actin, alpha 1, skeletal muscle | GO:0006936~ muscle contraction | hsa05110:Vibrio cholerae infection | 1 | 2.83 | NS | NS | 0.45 | <0.01 |
| *ACTN2* | Actinin, alpha 2 | GO:0006936~ muscle contraction | hsa04510:Focal adhesion | 1 | 3.67 | NS | NS | 0.45 | <0.01 |
| *ADHFE1* | Alcohol dehydrogenase iron-containing protein 1 | GO:0055114 oxidation reduction | hsa00010:Glycolysis/ Gluconeogenesis, hsa00071: Fatty acid metabolism, | 9 | 0.11 | −9.44 | <0.001 | NS | NS |
| *AGT* | Angiotensinogen (serpin peptidase inhibitor, clade A, member 8) | GO:0001944~ vasculature development | hsa04614:Renin-angiotensin system | 1 | 4.05 | NS | NS | 0.45 | <0.01 |
| *COX4I1* | Cytochrome c oxidase subunit IV isoform 1 | GO:0004129 cytochrome-c oxidase activity | hsa00190:Oxidative phosphorylation | 3 | 3.53 | −4.752 | 0.002 | NS | NS |
| *CYP51A1* | Cytochrome P450, family 51, subfamily A, polypeptide 1 | GO:0055114 oxidation reduction | hsa01100 Metabolic pathways | 4 | 2.71 | −6.12 | <0.001 | 0.45 | <0.01 |
| *GGPS1* | Geranylgeranyl diphosphate synthase 1 | GO:0044255~ cellular lipid metabolic process | hsa00100:Biosynthesis of steroids | 1 | 4.77 | NS | NS | 0.45 | <0.01 |
| *GSN* | Gelsolin (amyloidosis, Finnish type) | GO:0007015~ actin filament organization | hsa04810:Regulation of actin cytoskeleton | 25 | 0.69 | −6.78 | <0.001 | 0.35 | <0.05 |
| *MTFR1* | Mitochondrial fission regulator 1 | CC_GO:0005739 mitochondrion | | 9 | 0.52 | −9.44 | <0.001 | NS | NS |
| *MUSK* | Muscle, skeletal, receptor tyrosine kinase | GO:0007517~ muscle organ development | | 25 | 9.86 | −6.78 | <0.001 | 0.35 | <0.05 |
| *NDUFA8* | NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 8, 19 kDa | GO:0006119~ oxidative phosphorylation | hsa00190:Oxidative phosphorylation, | 25 | 0.06 | −6.78 | <0.001 | 0.35 | <0.05 |
| *PDK4* | Pyruvate dehydrogenase kinase, isozyme 4 | GO:0005996~ monosaccharide metabolic process | | 4 | 0.38 | −6.12 | <0.001 | 0.45 | <0.01 |
| *PON1* | Paraoxonase 1 | GO:0006644~ phospholipid metabolic process | hsa00361:gamma-Hexachlorocyclohexane degradation | 4 | 0.09 | −6.12 | <0.001 | 0.45 | <0.01 |
| *PTGS1* | Prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase) | GO:0006631~ fatty acid metabolic process | hsa00590:Arachidonic acid metabolism | 25 | 0.26 | −6.78 | <0.001 | 0.35 | <0.05 |
| *SGCE* | Sarcoglycan, epsilon | GO:0007517~ muscle organ development | | 4 | 0.36 | −6.12 | <0.001 | 0.45 | <0.01 |
| *TNC* | Tenascin c | GO:0007165~ signal transduction | hsa04510:Focal adhesion | 25 | 5.96 | −6.78 | <0.001 | 0.35 | <0.05 |

**Table 2** (Continued)

| Gene symbol | Gene description | Gene ontology: Biological Process | KEGG pathway | Chr | Distance from marker (Mb) | Dh/sd[1] | *P* | $F_{ST}^2$ | *P* |
|---|---|---|---|---|---|---|---|---|---|
| *TOMM20* | Translocase of outer mitochondrial membrane 20 homolog (yeast) | GO:0006886~ intracellular protein transport | hsa00600:Sphingolipid metabolism | 1 | 4.96 | NS | NS | 0.45 | <0.01 |
| *UGCG* | UDP-glucose ceramide glucosyltransferase | GO:0006643~ membrane lipid metabolic process | hsa00600 Sphingolipid metabolism | 25 | 9.01 | −6.78 | <0.001 | 0.35 | <0.05 |

Twenty candidate genes were selected on the basis of gene ontology (GO biological process, GO cellular compartment or KEGG pathway) and their presence in one of the top ranked regions with a signature of selection in the Thoroughbred. The chromosome, distance from microsatellite marker and selection statistics are shown.

[1]Dh/sd = Ewens-Watterson test statistic for deviation from expected heterozygosity.

[2]$F_{ST}$ = among population differentiation.

### Selection of candidate genes

A panel of 20 genes (Table 2) was selected from approximately 100 putative exercise-relevant genes among the 1202 predicted genes located within the genomic regions lying at the extremes of the distributions for departure from expected neutral variation identified in a population genetics–based genome scan (Gu *et al.* 2009). Candidate genes were selected based on gene ontology and were prioritized according to the statistical ranking of the genomic outlier. With the exception of *COX4I1*, all genes were selected from the top three ranked regions for each of the two statistics (Dh/sd and $F_{ST}$). *COX4I1* was included because of its relevance to exercise adaptation determined in previous gene expression experiments (Eivers *et al.* 2009) and its location within the region ranked fifth in the Dh/sd distribution.

### Selection of SNPs from EquCab2.0 SNP database and SNP genotyping

The EquCab2.0 SNP database was interrogated for SNPs discovered in the Horse Genome Sequencing Project (Wade *et al.* 2009) that were located within the genomic sequence of each of the 20 candidate genes. No SNPs were identified in the regions containing the genomic sequence for the *AGT* and *CYP51A1* genes, and these genes were therefore excluded. *MUSK* was excluded as it contained no Thoroughbred SNPs and was located >9 Mb from the microsatellite marker that displayed a signature of selection (Gu *et al.* 2009). Of the 17 genes that were included, 13 contained genomic sequence variants that had been identified in Thoroughbred. For four genes (*UCGC, SGCE, TOMM20, ACSS1* and *COX4I1*), SNPs which were initially identified in other equine populations were included. In total, 66 SNPs located within the genomic sequence of 17 putative exercise-relevant genes (including six within protein-coding regions) and two SNPs located in flanking sequences (*ACTA1*) were selected from the EquCab2.0 SNP database for genotyping (Table S1). SNP genotyping was carried out using iPLEX® technology (Jurinke *et al.* 2004) by Sequenom Inc. at their facilities in San Diego, USA.

### Statistical analysis

SNP-phenotype associations were investigated for 57 SNPs in a series of case–control tests: TBE *vs* TBO; TBE ≤ 8 f *vs* TBE > 8 f; TBE ≤ 7 f *vs* TBE > 8 f; TBE > 8 f *vs* TBO; TBE ≤ 8 f *vs* TBO; TBE ≤ 7 f *vs* TBO; and TBE males *vs* TBO males. The sample sizes for each set were as follows: TBE, $n = 86$; TBO, $n = 62$; TBE > 8 f, $n = 36$; TBE ≤ 8 f, $n = 51$; TBE ≤ 7 f, $n = 41$; TBE males, $n = 41$; TBO males, $n = 18$.

All statistical analyses, including tests of association were performed using PLINK Version 1.05 (http://pngu.mgh.harvard.edu/purcell/plink) (Purcell *et al.* 2007). Quality control analyses included computation of sample allele frequencies and percentage of missing genotypes. A series of case–control association tests were performed for all loci. Statistical significance was assessed using the Cochran-Armitage test for trend and an unconditioned genotypic model. The linear regression model was used to evaluate quantitative trait association using best race distance (furlongs) and highest lifetime Racing Post Rating (RPR) as the phenotypes. Best race distance was defined as the distance (furlongs) of the highest grade Group race won by an individual. In cases where multiple races of the same grade were won, the distance of the race in which the most prize money was won was used.

## Results

### Genotyping assay performance

Fifty-eight SNPs had call rates of >90%. Four SNPs had call rates of <15% and were excluded (Table S2). Of the

successful genotyping assays, seven displayed a minor allele frequency (MAF) <0.02 and were excluded. An exact test for deviation from Hardy–Weinberg proportions was applied at each locus (Wigginton *et al.* 2005). Six SNPs in four genes [*ACSS1* (3 SNPs), *TNC* (1 SNP), *ACTN2* (1 SNP) and *PTGS1* (1 SNP)] showed significant departures from Hardy–Weinberg equilibrium (HWE) ($P < 0.05$). However, we did not exclude SNPs that deviated significantly from Hardy–Weinberg proportions from subsequent analyses as the Thoroughbred population does not meet many of the requirements for HWE. In addition, the tests for association remain valid under departure from Hardy–Weinberg proportions, albeit with a potential loss in power if they reflect systematic genotyping errors. Therefore, 57 SNPs were included in the genetic association tests.

### Case–control association tests

We performed a series of population-based case–control association tests to investigate SNP associations with retrospective racecourse performance phenotypes in Thoroughbred horses. Results for all case–control tests are available in Table S3.

### Significant association with elite race winning performance

The three *PDK4* SNPs (PDK4_38968139, PDK4_38969307 and PDK4_38973231) were significantly ($P < 0.01$) associated with elite race winning performance (TBE). The *PDK4* SNPs were also significantly associated with performance in four of the other case–control association tests (TBE > 8 f *vs* TBO; TBE ≤ 8 f *vs* TBO; TBE ≤ 7 f *vs* TBO and TBE *vs* TBO males), and in each case the PDK4_38973231 SNP had the strongest association. The strongest association was with elite racing performance (i.e. TBE *vs* TBO) ($P = 0.0017$; odds ratio = 2.20). A full set of results is available in Table S3.

### Significant association with short distance elite race winning performance

When the elite cohort was separated into subgroups of individuals that had won their best race over short distances (TBE ≤ 8 f) and long distances (TBE > 8 f), the PON1_38697145 SNP was significantly associated with elite short distance racing ($P = 0.0358$, odds ratio = 3.47). Further subdivision of the short distance subgroup into individuals that won their best race over even shorter distances (TBE ≤ 7 f) revealed a significant association with the ACN9_40279726 SNP ($P = 0.0448$, odds ratio = 2.07). When the frequencies of alleles among the short distance elite race winning cohort (TBE ≤ 8 f and TBE ≤ 7 f) was compared with the non-winning cohort, the ADHFE1_18802749 ($P = 0.0494$, odds ratio = 4.20), GSN_25024464 ($P = 0.0354$, odds ratio = 3.18) and GSN_25028755 ($P = 0.0454$, odds ratio = 3.03) SNPs had significantly different allele frequency distributions.

### Significant association with elite race winning performance among males

Male selection in the Thoroughbred population is particularly pronounced, with an approximately 1:45 ratio of breeding males to females (Indecon 2004). Therefore, we investigated SNP associations with performance in males only. In addition, two of the *PDK4* SNPs (PDK4_38969307 and PDK4_38973231), the ACTN2_74842283 ($P = 0.0437$, odds ratio = 3.25), PTGS1_25991437 ($P = 0.0197$, odds ratio = 3.76), PTGS1_26007699 ($P = 0.0051$, odds ratio = 4.60) and COX4I1_32772871 ($P = 0.0442$, odds ratio = 2.23) SNPs were significantly associated with elite race winning performance among males.

### Quantitative trait association tests

We performed a quantitative trait association test using the best race distance (furlongs) for each individual as the phenotype. The ACN9_40279726 ($P = 0.0321$) and PON1_38697145 ($P = 0.0350$) SNPs were significantly associated with best race distance. We also performed a quantitative trait association test using handicap rating (Racing Post Rating) for each individual as the phenotype (Table S4). Two of the *PDK4* SNPs were significantly associated with handicap rating (PDK4_38969307, $P = 0.0369$; and PDK4_38973231, $P = 0.0252$).

### Validation of associations

Five SNPs in four genes were selected for validation in an independent sample set (Table 3): ACN9_40279726; ACSS1_759076; COX4I1_32772871; PDK4_38969307; and PDK4_38973231. The two *PDK4* SNPs were significantly associated with elite racing performance in the validation set (PDK4_38969307, $P = 0.0255$; PDK4_38973231, $P = 0.0150$). The PDK4_38973231 SNP consistently had the strongest association, and when all samples ($n = 278$) were considered, the significance of association became stronger (PDK4_ 38973231, $P = 0.0004$, odds ratio = 1.97 C.I. (95) = 1.35–2.87).

We then attempted to determine the most parsimonious genetic model for PDK4_38973231 by repeating the analysis with coding variables for additive, recessive and overdominant models (Table 4). For PDK4_38973231, a dominant model in which the A:A and A:G genotypes were favourable provided the best explanation for the data ($P = 0.0003$), with the A:A and A:G genotypes more common among elite (70%) than non-elite (47%) racehorses.

Both *PDK4* SNPs were also validated for association with RPR (Validation set: PDK4_38969307, $P = 0.0369$;

**Table 3** Case-control association test results.

| | | | Study | | | Validation | | | All | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | Gene SNP | Location | $\chi^2$ | *P* | OR | $\chi^2$ | *P* | OR | $\chi^2$ | *P* | OR |
| 3 | *COX4I1* | 32 772 871 | 3.46 | 0.0628 | 1.64 | 0.04 | 0.8494 | 1.057 | 2.56 | 0.1097 | 1.36 |
| 4 | *PDK4* | 38 969 307 | 7.41 | 0.0065 | 2.03 | 4.99 | 0.0255 | 1.994 | 11.86 | 0.0006 | 1.94 |
| 4 | *PDK4* | 38 973 231 | 9.87 | 0.0017 | 2.20 | 5.92 | 0.0150 | 2.118 | 12.68 | 0.0004 | 1.97 |
| 22 | *ACSS1* | 759 076 | 3.84 | 0.0501 | 1.88 | 0.12 | 0.7304 | 1.139 | 2.39 | 0.1218 | 1.45 |
| 4 | *ACN9*[1] | 40 279 726 | 6.21 | 0.0130 | 2.48 | 0.05 | 0.8268 | 1.07 | 3.83 | 0.0500 | 1.61 |

Four SNPs in three genes (*ACSS1*, *COX4I1* and *PDK4*) were validated for association with elite racing performance (TBE *vs* TBO). $\chi^2$, *P*-value (*P*) and odds ratios (OR) are shown for the study sample set, the validation sample set and a cohort containing all genotyped samples.
[1]This comparison was between elite horses that won their best race ≤7 f and elite horses that won their best race >8 f.

**Table 4** Genetic model for the PDK4_38973231 A/G SNP.

| Model | TBE | TBO | $\chi^2$ | DF | *P*-value |
|---|---|---|---|---|---|
| Genotype | 35/92/55 | 10/35/50 | 13.74 | 2 | 0.0010 |
| Trend | 162/202 | 55/135 | 12.22 | 1 | 0.0005 |
| Allelic | 162/202 | 55/135 | 12.68 | 1 | 0.0004 |
| Dominant | 127/55 | 45/50 | 13.32 | 1 | 0.0003 |
| Recessive | 35/147 | 10/85 | 3.476 | 1 | 0.0623 |

A dominant model in which the A:A and A:G genotypes were favourable provided the best explanation for the association of the PDK4_38973231 SNP.
TBE, elite Thoroughbred; TBO, ordinary Thoroughbred; DF, degrees of freedom.

PDK4_38973231, $P = 0.0252$; All: PDK4_38969307, $P = 0.0017$; PDK4_ 38973231, $P = 0.0008$) (Table S4). At locus PDK4_ 38973231, A:A and A:G horses had on average a significantly higher handicap rating (A:A 98.3 ± 32.8; A:G 97.9 ± 29.0) than individuals with the G:G (81.7 ± 31.4) genotype (Table S5).

## Discussion

In a series of case–control and quantitative association tests, we have identified SNPs associated with racing performance phenotypes located within genomic regions that have been targets of selection during the development of the Thoroughbred. It is likely that these SNPs are linked to functional variants in genes or regulatory elements that are associated with physiological adaptations that enable superior racing performance in Thoroughbreds.

For the first time, we report here the association of a SNP with elite race winning performance. When all individuals with a RPR record ($n = 228$) were considered, the A:A and A:G genotypes (PDK4_ 38973231) had on average a 16.2–16.6 lb handicap advantage over G:G horses. The expression of *PDK4* is co-ordinated by the transcriptional co-activator PGC-1α (Wende *et al.* 2005), which has been identified as one of the critical control factors in the adaptation to exercise (Arany 2008). Specifically, PGC-1α is a key regulator of energy metabolism that regulates insulin sensitivity by controlling glucose transport, mediates exercise-induced angiogenesis (Chinsomboon *et al.* 2009) and co-ordinates mitochondrial biogenesis via its interaction with nuclear encoded mitochondrial protein genes (Scarpulla 2008). The regulation of glucose utilization is tightly controlled by the uptake of glucose by glucose transporters, the rate of glycolytic flux and the conversion of pyruvate to acetyl-CoA in mitochondria via the catalytic function of the pyruvate dehydrogenase complex (PDC). The critical rate-limiting step in the oxidation of glucose is the regulation of assembly of the PDC, which is controlled by pyruvate dehydrogenase kinase (PDK). PDK blocks the formation of the PDC, resulting in the beta-oxidation of fatty acids to acetyl-CoA as the substrate for oxidative phosphorylation. The oxidation of fatty acids is highly efficient in the generation of ATP and is controlled by the expression of *PDK4* in skeletal muscle during and after exercise (Pilegaard & Neufer 2004).

Structural and functional genomics approaches represent powerful strategies to dissect key components of the molecular contribution to performance phenotypes and the biology of the equine athlete. We have previously identified a significant increase (+7.4-fold) in equine skeletal muscle *PDK4* mRNA during recovery from exercise (Eivers *et al.* 2009). This observation is consistent with prolonged inhibition of the PDC to decrease glucose oxidation and increase mitochondrial fatty acid oxidation (Wende *et al.* 2005). Variation in gene expression is strongly influenced by structural genetic variation, and therefore the integration of functional data in the interpretation of structural variation is highly relevant. Although it is unlikely that the *PDK4* SNP (PDK4_38973231) described in this study is directly influencing the control of gene expression, it is likely to be in linkage disequilibrium with a SNP that is affecting the expression of the *PDK4* gene. Re-sequencing efforts and further genotyping will be required to determine the functional variant in the *PDK4* gene associated with elite performance. Notwithstanding, this information may be employed as a marker for the selection of racehorses with the genetic potential for superior racing ability.

## Acknowledgements

We thank the numerous contributors of Thoroughbred horse samples. This work was financed by a Science Foundation Ireland President of Ireland Young Researcher Award (04/YI1/B539) to EH.

## Conflicts of interest

EH, JG and DM are named in patent applications. The remaining authors have declared no conflicts of interest.

## References

Akey J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research* **19,** 711–22.

Arany Z. (2008) PGC-1 coactivators and skeletal muscle adaptations in health and disease. *Current Opinion in Genetics and Development* **18,** 426–34.

Bray M.S., Hagberg J.M., Perusse L., Rankinen T., Roth S.M., Wolfarth B. & Bouchard C. (2009) The human gene map for performance and health-related fitness phenotypes: the 2006–2007 update. *Medicine and Science in Sports and Exercise* **41,** 35–73.

Chinsomboon J., Ruas J., Gupta R.K., Thom R., Shoag J., Rowe G.C., Sawada N., Raghuram S. & Arany Z. (2009) The transcriptional coactivator PGC-1alpha mediates exercise-induced angiogenesis in skeletal muscle. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 21401–6.

Constantinopol M., Jones J.H., Weibel E.R., Taylor C.R., Lindholm A. & Karas R.H. (1989) Oxygen transport during exercise in large mammals. II. Oxygen uptake by the pulmonary gas exchanger. *Journal of Applied Physiology* **67,** 871–8.

Eivers S.S., McGivney B.A., Fonseca R.G., MacHugh D.E., Menson K., Park S.D., Rivero J.L., Taylor C.T., Katz L.M. & Hill E.W. (2009) Alterations in oxidative gene expression in equine skeletal muscle following exercise and training. *Physiological Genomics* **40,** 83–93.

Evans D.L., Harris R.C. & Snow D.H. (1993) Correlation of racing performance with blood lactate and heart rate after exercise in thoroughbred horses. *Equine Veterinary Journal* **25,** 441–5.

Gaffney B. & Cunningham E.P. (1988) Estimation of genetic trend in racing performance of thoroughbred horses. *Nature* **332,** 722–4.

Gu J., Orr N., Park S.D., Katz L.M., Sulimova G., MacHugh D.E. & Hill E.W. (2009) A genome scan for positive selection in thoroughbred horses. *PLoS ONE* **4,** e5767.

Harrison S.P. & Turrion-Gomez J.L. (2006) Mitochondrial DNA: an important female contribution to thoroughbred racehorse performance. *Mitochondrion* **6,** 53–63.

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.

Hinchcliff K.W., Kaneps A.J. & Geor R.J. (2008) *Equine Exercise Physiology: The Science of Exercise in the Athletic Horse*. Elsevier Saunders, Edinburgh.

Indecon (2004) An assessment of the economic contribution of the Thoroughbred breeding and horse racing industry in Ireland (A final report for the Irish Thoroughbred Breeders' Association, European Breeders' Fund, and Horse Racing Ireland). Available at: http://www.goracing.ie/AssetLibrary/Files/HRI/Info/indeconReportJULY2004.pdf.

Jones J.H., Longworth K.E., Lindholm A., Conley K.E., Karas R.H., Kayar S.R. & Taylor C.R. (1989) Oxygen transport during exercise in large mammals. I. Adaptive variation in oxygen demand. *Journal of Applied Physiology* **67,** 862–70.

Jurinke C., Oeth P. & van den Boom D. (2004) MALDI-TOF mass spectrometry: a versatile tool for high-performance DNA analysis. *Molecular Biotechnology* **26,** 147–64.

McGivney B.A., Eivers S.S., MacHugh D.E., MacLeod J.N., O'Gorman G.M., Park S.D., Katz L.M. & Hill E.W. (2009) Transcriptional adaptations following exercise in thoroughbred horse skeletal muscle highlights molecular mechanisms that lead to muscle hypertrophy. *BMC Genomics* **10,** 638.

Oleksyk T.K., Smith M.W. & O'Brien S.J. (2010) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **365,** 185–205.

Pilegaard H. & Neufer P.D. (2004) Transcriptional regulation of pyruvate dehydrogenase kinase 4 in skeletal muscle during and after exercise. *Proceedings of the Nutrition Society* **63,** 221–6.

Pritchard J.K., Pickrell J.K. & Coop G. (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* **20,** R208–15.

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.

Sambrook J. & Russell D.W. (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Scarpulla R.C. (2008) Nuclear control of respiratory chain expression by nuclear respiratory factors and PGC-1-related coactivator. *Annals of the New York Academy of Sciences* **1147,** 321–34.

Wade C.M., Giulotto E., Sigurdsson S. *et al.* (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326,** 865–7.

Wende A.R., Huss J.M., Schaeffer P.J., Giguere V. & Kelly D.P. (2005) PGC-1alpha coactivates *PDK4* gene expression via the orphan nuclear receptor ERRalpha: a mechanism for transcriptional control of muscle glucose metabolism. *Molecular and Cellular Biology* **25,** 10684–94.

Wigginton J.E., Cutler D.J. & Abecasis G.R. (2005) A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76,** 887–93.

## Supporting information

Additional supporting information may be found in the online version of this article.

**Table S1** SNP details: Locus ID/Local SNPID, gene name, chromosome location and SNP assay.

**Table S2** Genotyping assay results: Assay ID, genotyping call rate, total number of individuals genotyped and allele frequencies.

**Table S3** Case–control association test results: $\chi^2$, *P*-value (*P*) and odds ratios (OR) for the study sample set for a series of case–control association tests.
**Table S4** Quantitative association test results.
**Table S5** Quantitative trait means for PDK4_38973231: Means for Racing Post Ratings for each of the three PDK4_38973231 SNP genotypes.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer reviewed and may be re-organized for online delivery, but are not copy edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.