Mitochondrial DNA: An important female contribution to thoroughbred racehorse performance

Stephen Paul Harrison [a,*], Juan Luis Turrion-Gomez [a,b]

[a] *Thoroughbred Genetics Ltd., Kent Science Park, Sittingbourne, Kent ME9 8AZ, UK*
[b] *Departamento de Microbiologia y Genetica, Universidad de Salamanca, Edificio Departamental, Salamanca 37007, Spain*

Abstract

The mitochondrial DNA (mtDNA) molecule, carrying genes encoding for respiratory chain enzymes, is a primary candidate for demonstrating associations between genotype and athletic performance in mammalian species. In humans, variation at seven protein encoding mitochondrial loci has been implicated in influencing fitness and performance characteristics. Although thoroughbred horses are selected for racing ability, there have not been any previous reported associations between genotypes and racecourse performance. The multi-factorial nature of the inheritance of racing ability is an obvious complicating factor. However, mitochondrial gene variation may represent a measurable component contributing to performance variability. Previous population studies of thoroughbreds have shown the existence of D-loop variation. Importantly, we have observed that there is also independent and extensive functional mitochondrial gene variation in the current thoroughbred racehorse population and that significant associations exist between mtDNA haplotype, as defined by functional genes, and aspects of racing performance.
© 2006 Elsevier B.V. and Mitochondria Research Society. All rights reserved.

Keywords: Mitochondrial DNA; Stamina; Racehorse; Thoroughbred; Performance; Genes

1. Introduction

There is no formal or established method of anatomical or pre-competitive performance assessment for thoroughbred racehorses. Nor are there any precise genetic criteria applied in the selection of these horses for breeding. Information about genetic inheritance of performance traits and potential breeding value is based primarily on unquantified and frequently anecdotal, anatomical observations about the progeny of stallions and mares or, naturally, on the racecourse performances of their descendents over a number of generations.

Although scientifically unproven, a traditional belief of many thoroughbred breeders is that the mother (dam), rather than the father (sire), contributes more strongly to a horse's racing stamina (Leicester, 1983). Varying degrees of importance have also been attached to differing perceived endurance and

---

* Corresponding author. Tel.: +44 1795 411544; fax: +44 1795 411543.
 E-mail address: sph@thoroughbredgenetics.com (S. P. Harrison).

performance abilities of certain female breeding lines. Historically, efforts have been made to categorise maternal lines in order of importance defined by the racing quality of horses descended from them (Lowe, 1898; Bobinski, 1953), though this information is now considered aged and incomplete.

Naturally, dated pedigree-based comparisons describing family origins and performance characteristics do not take into account shared maternal genetic origins prior to formal studbook recordings (Weatherby and Sons, 1791) and overlook potential common genotype-mediated performance characteristics. Additionally, they do not provide any indication of functional gene similarities and associated performance trends in different female lines. However, these traditional observations lend support to the postulation that variation between horses in the mitochondrial protein-encoding genes may affect relative performance characteristics.

In humans, a total of sixteen mitochondrial genes have been reported as having an effect on metabolism, which may help determine athletic or physiological potential (Wolfarth et al., 2005). Moreover, variation in seven genes encoding for respiratory chain enzymes has been implicated in influencing fitness and performance characteristics. Of these, two are constituents of the cytochrome oxidase complex IV, one is cytochrome b and four are members of the NADH dehydrogenase complex I. Specifically, variants of genes coding for NADH hydrogenase subunits ND2 and ND4 affect initial fitness and responses to training (Dionne et al., 1993). Further to this, mtDNA variation defined by haplotype incorporating functional DNA regions, has also been suggested to have a role in determining stamina and power capabilities (Niemi and Majamaa, 2005). There is also evidence that aerobic performance is influenced with a bias towards maternal inheritance (Lesage et al., 1985; Perusse et al., 2001).

Surprisingly, for a well established athletic animal, there has been a paucity of studies applied to the area of substrate utilization during exercise of thoroughbreds. However, differential respiratory and energetic requirements for horses competing over varying distances has been indicated.

Harkins et al. (1993) showed that correlations between running speeds and a range of relevant physiological variables, including VO$_2$max and blood/plasma lactate levels, were stronger when horses ran over 2000 m in comparison with 1200 m. Similarly, Eaton et al. (1992), incorporating calculations based on VO$_2$max in horses exercised over differing simulated distances, postulated that the balance of energy partitioning between aerobic and anaerobic pathways was dependent on race distance. They suggested that a bias towards the aerobic pathway in fulfilling the larger proportion of a racehorse's enery requirement becomes stronger as racing distance increases from 1000 to 3200 m.

VO$_2$max has long been considered a useful measurement of exercise capacity in humans and variation between individuals has been linked with specific mtDNA variants (Dionne et al., 1993). Differences in VO$_2$max measurements of horses running over varying distances suggest that it is feasible to suspect that mtDNA variation may also play a role in stamina and exercise potential of thoroughbreds. Studies by Harris et al. (1987) lend further support for the potential role of mitochondrial genes by showing that post-exercise depletion of muscle ATP content is greater in horses exercised over 2000 m compared with those tested over 800 m.

The pedigree and racing records kept for the thoroughbred are unparalleled in detail by any other breed or domestic species. The population is genetically finite and the restricted studbook has been employed since 1791 (Weatherby and Sons, 1791). This makes the study of the thoroughbred useful as a unique model system with application to other species, for instance humans, where extended historical, genealogical and pedigree data relating to athletic performance does not exist.

Previous population studies restricted to examinations of D-loop SSCP and sequence variation in relatively small samples of thoroughbreds (Ishida, et al.,1994; Marklund et al., 1995; Hill et al., 2002) have shown that extensive mtDNA variation exists within the breed. However, the primary objectives of our study were to determine functional mitochondrial gene variation in the current thoroughbred population and to investigate potential associations, whether between individual gene variants or combined gene variation, with racing performance and stamina potential.

2. Materials and methods

2.1. DNA samples

DNA samples, which had been extracted from white blood cells using a range of non-phenol based techniques, were taken from our collection. The DNA had been stored in 10 mM Tris–HCL/1 mM EDTA buffer at −20 °C.

One thousand thoroughbred samples were analyzed, representing 33 thoroughbred, globally occurring, maternal lines. Reference to the thoroughbred General Stud Book (Weatherby and Sons, 1791) showed that the majority of currently existing European female lines, traceable to original stud book members, were represented within this selection, thereby heightening the chances of achieving coverage of all available mtDNA variation within the breed.

A diverse group of non-thoroughbreds consisting of 21 Polish Tarpans, 23 Tibetan riding horses, 22 Peruvian Paso horses and 56 Irish Draught horses was also analyzed.

2.2. Primers and PCR amplification conditions

Primer design (Table 1) was based on a previously reported sequence (Xu and Arnason, 1994). 16 gene loci were covered, including those of 13 protein-encoding genes, the D-loop, 12S and 16S rRNA. A number of primer sets were evaluated for each gene. For ease of presentation, the primer positions detailed are those found to provide the greatest coverage at each locus without diminishing the effectiveness of SSCP analysis. As a confirmatory step, where the regions covered exceeded 800 bp, further primer pairs were used to amplify shorter, overlapping nested areas in samples from all female families and sub-families. For haplotyping purposes, variation at these sub-positions was combined. However, in every case, there was agreement between the longer and shorter, combined amplicons in the degree of allelic variation or absence of variation detected at each locus.

PCR reactions were carried out in 0.2 ml reaction tubes in a volume of 20 µl containing reaction buffer (10 mM Tris–HCL, pH 8.0, 50 mM KCL, 3.75 mM MgCl), 0.2 mM dNTPs, 0.5 pmol µl^{-1} of each primer, 0.025 U µl^{-1} *Taq* polymerase and 5 ng µl^{-1} DNA template. Amplification cycles are described in Table 1. Thirty amplification cycles were carried out for each primer pair. These varied in annealment and extension conditions but commonly started with 1 cycle of denaturation at 94 °C for 1 min and constant 30 s denaturation phases at 94 °C.

2.3. SSCP analysis

SSCP analysis was carried out using the method of Kukita et al. (1997), designed to detect sequence variation in DNA fragments of around 800 bp in low pH conditions. PCR products were denatured by heating at 95 °C for 10 min, followed by immediate placement on ice. Products were assessed for polymorphisms by running the denatured DNA on 10% polyacrylamide, 5% glycerol, 0.5×TBE gels. The gels were run at 0.6 V h^{-1}×100 bp and stained by silver staining (Caetano-Anolles et al., 1991). Clear differential SSCP patterns were produced for all detectable variants. Subject to confirmation by sequencing, newly detected SSCP variants were named alphabetically in order of detection.

2.4. DNA sequencing

Representatives of all thoroughbred functional gene SSCP variants were selected for DNA sequencing. DNA from representatives of all known thoroughbred female lines represented by our sample was also sequenced. Sequencing was performed using an Applied Biosystems ABI PRISM 377 sequencer/genotyper using the recommended protocol and labelling. PCR amplification products were cleaned for sequencing using ChargeSwitch PCR clean-up kits (Invitrogen). Two microliters of PCR product was added to 18 µl of water and sequencing of the regions was carried out using the forward and reverse primers in both directions. Sequences were analysed using BioEdit (Hall, 1999). DNA sequences for the translatable loci are held on the GenBank database and have been assigned the accession numbers DQ312329-DQ312360. Variants specific to non-thoroughbreds were scored only from SSCP data, as were those for the D-loop.

2.5. Pedigree, haplotype and racing data

Using the General Stud Book (Weatherby and Sons, 1791) we identified the pedigrees of randomly

Table 1
Loci and primers[a] used in the investigation

Gene	Locus	Forward primer 5′to3′	Reverse primer 5′to3′	mtDNA position (bp)	length (bp)	Ann. 30 s (°C)	Ext. 72 °C (min)
NADH dehy-drogenase 1	*MTND1*	TGTTCATAAT-TAACGTCCTC	CTATGTTTGTGGTG-GGATG	2772–3727	955	47.5	3
NADH dehy-drogenase 2	*MTND2*	CCCTTATCTTCA-CAACTATTC	GGGAGGATATAA-CAATTAACG	3944–4936	992	48.5	3
NADH dehy-drogenase 3	*MTND3*	ATAAACCTCA-TACTGACACTCC	TTTGGGTTCATTCG-TAGG	9498–9822	324	50.5	1.5
NADH dehy-drogenase 4	*MTND4*	CAATAGCC-TAAACTTCTCAC	GAATAGCTCTC-CAATTAGG	10345–11344	999	46.5	3
NADH dehy-drogenase 4L	*MTND4L*	ATATCTTCCTAG-CATTCACAG	TAGCATTGGAG-GAGGTTAAG	9934–10210	276	49.0	1.5
NADH dehy-drogenase 5	*MTND5*	TTTCCAACTGTT-CATCGG	GTTGGAGATGAA-GAATCCG	12193–13192	999	51.0	3
NADH dehy-drogenase 6	*MTND6*	AAACCTTCACC-TATTTATGG	TTAATCTCCAC-GAGTAACC	13587–14048	461	47.0	2
Cytochrome c oxidase 1	*MTCO1*	ACATCGG-CACTCTGTACC	AAGAAGAT-GAAGCCTAGAGC	5402–6401	999	50.0	3
Cytochrome c oxidase 2	*MTCO2*	CCCTTCCAACTAG-GATTC	ATTGATGCAGAT-CATTCTTC	7057–7724	667	48.5	2.5
Cytochrome c oxidase 3	*MTCO3*	CACCAAACC-CACGCTTAC	TCCTCATCAA-TAAATAGAGACG	8651–9424	773	52.0	2.5
ATP synthase 6	*MTATP6*	AAATC-TATTCGCCTCTTTC	AGGTGTTGTCGTG-TAAGTAAAG	7973–8643	671	49.0	2.5
ATP synthase 8	*MTATP8*	ATGCCACAGTTG-GATACATC	GTAGCGAAA-GAGGCGAATAG	7804–7996	192	49.0	1.5
Cytochrome b	*MTCYB*	CCTAATCCTC-CAAATCTTAAC	CTAAGAGTCAGAA-TACGCATTG	14307–15175	869	52.0	3
12S RNA	*MTRNR1*	AGAATTACA-CATGCAAGTATCC	CAAGTA-CACCTTCCGGTA-TAC	108–1039	931	52.0	3
16S RNA	*MTRNR2*	CTAAAGCTAGCC-CAAACAATAC	GTTTGTGTTTGCCG-AGTTC	1114–1939	825	50.5	3
16S RNA	*MTRNR2*	TGTTAACCCAACA-CAGGC	GGCGGTAGAAGT-TATAAATTAG	1868–2679	811	52.0	3
D-loop	*DLOOP*	GCTCCACCATCAA-CACCCAAAG	TGAAGAAAGAAC-CAGATGCCAG	15420–15860	440	52.0	2

[a] A number of primer sets were evaluated for each gene. Nested sequences were also used but for ease of presentation, the primer positions listed are those found to provide the greatest coverage at each locus without diminishing the effectiveness of SSCP analysis.

selected batches of 1000 thoroughbred horses known to have been racing or breeding in the years 1953 and 2003 in the UK. Where possible, we assigned mtDNA types to each of them and made corrections for anomalous haplotype possession when necessary. There was no discernible shift in haplotype distri-bution in the UK during this period.

Separate data sets were also collected randomly from Raceform records (Raceform, 2003) which detail the annual racing careers of all thoroughbreds running in the UK. From this we identified the pedigrees of 1000 horses shown to have been racing at three years old (3 yo) in 2003. We were able to assign mtDNA haplotypes to 99.8% of horses and to estimate the distribution of identifiable haplotypes in the current population at 3 yo.

European racing governing bodies describe the selected races as 'Group' races. In these high value races, where the entry fees and the kudos of winning are correspondingly higher, horses are likely to be racing to their true potential. A consistent test of stamina or speed should, therefore, be applied.

The races were chosen on the basis that they have been run consistently over the same distance and age group during the years covered by the study. In total, 21 different races for 3 yo horses were scored over at least 44 years ($n = 1035$) between 1953 and 2003. These races were the: Fred Darling, Greenham, Nell Gwyn and Jersey Stakes, all 1400 m; 1000 Guineas, 2000 Guineas, St James's Palace, Coronation and Craven Stakes, all 1600 m; Sandown Classic, Musidora and Dante Stakes, all 2000 m; Lingfield Derby Trial, Derby, Oaks, Ribblesdale, King Edward VII, Great Voltigeur, Chester Vase and Gordon Stakes, all 2400 m; St Leger Stakes, 2800 m.

2.6. Statistics notes

For each haplotype, correlation coefficients and probabilities were calculated for RI of each race versus racing distance. In Table 4 only average RI for each distance bracket is shown for ease of presentation. Similarly, to determine the significance of differences between haplotypes in performance at specific distance brackets, ANOVAs used the RI for all races within each distance grouping. In the text, only average RI is shown for ease of presentation. In these ANOVAs, the data from the one race at 2800 m was omitted.

To assess the difference in haplotype distribution between various horse breeds, a Fisher's exact test was applied using the frequency of occurrence of each haplotype in the DNA samples from each breed.

2.7. Haplotype relationship analysis

Relationship assessments between haplotypes were carried out using sequence data for the variable, translatable mitochondrial genes. A relationship tree based on differences between the sequences of variants was constructed by Unweighted Pair Group Method with Arithmetic Mean (UPGMA) using the MEGA version 3.1 program (Kumar et al., 2004). The algorithm for UPGMA is discussed in detail in Nei and Kumar (2000) and this method assumes that the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages or haplotypes.

3. Results and discussion

In the thoroughbred, allelic variation was found in eight translatable genes and in the 16S and 12S RNA genes (Table 2). Horses were assigned as one of 17 genetic 'haplotypes' based on the combined variation at each of these genes and in the D-loop. The genes coding for NADH hydrogenase 1, 3 and 6, ATP synthase 8 and cytochrome oxidase 3 showed no allelic variation. Corresponding absence of variation at these loci in a diverse group of non-thoroughbreds, not selected for racing, indicated that this conservation was unlikely to be due to variant selection at these loci for enhanced racing ability.

The non-thoroughbreds could be categorised into two types: geographically distinct, old breeds that existed prior to the formal instigation of the thoroughbred consisting of Polish tarpans, Tibetan riding horses and Peruvian paso horses; and the relatively new breed of Irish draught horse which has developed partially as a result of cross breeding with the thoroughbred.

In both cases one would expect that, sufficiently different selection pressures to the thoroughbred have been applied and therefore, observations of common variant restriction between these animals, may help to eliminate, from consideration, loci and gene variants that do not obviously contribute to racing performance. Conversely, it is more difficult to state categorically that haplotype distribution variance between the breeds is a result of selection for racing ability in the thoroughbred since variability in haplotype distribution may be due to differences in the initial population structure of the respective breeds. Additionally, the range of racing distances employed for thoroughbreds could, theoretically, give rise to positive selection pressure for a number of haplotypes. A summary of the percentage distribution of haplotypes in the DNA samples from the different breeds is shown in Tables 3a and b.

Haplotype relationships between various breeds have already been described in previous studies (Vila, et al., 2001; Hill et al., 2002; Jansen et al., 2002) and it is not an objective of this paper to further discussion of breed classification and evolution. However, there are some observations that may be of relevance in relation to selection for racing performance and these are discussed in greater detail below.

Table 2
Haplotypes and their constituent variants

	%age Pop[a]	Loci										
		D-loop	CYB	ATP 6	ND2	ND4	ND4L	ND5	CO1	CO2	16S	12S
Haplotype												
I	19.2	D	A	A	A	C	B	B	B	D	A	A
II	11.8	B	B	C	A	B	A	D	A	B	A	A
III	8.4	B	B	B	A	B	A	A	A	B	A	A
IV	16.6	A	A	A	C	A	A	F	D	A	A	A
V	1.4	C	A	A	A	C	B	C	C	D	A	A
VI	0.2	F	B	C	A	B	A	A	A	B	A	A
VII	2.2	K	A	A	A	C	B	C	C	C	A	B
VIII	2.0	E	A	A	A	C	B	C	C	C	A	B
IX	2.2	L	A	A	A	D	A	F	D	A	A	A
X	6.4	M	A	E	A	C	B	A	A	D	A	C
XI	2.8	I	A	A	A	C	B	E	C	D	A	A
XII	2.4	H	C	D	B	C	B	A	D	D	C	A
XIII	0.8	C	A	A	A	E	B	C	C	D	A	A
XIV	<0.01	J	A	A	A	C	B	A	A	D	A	C
XV	7.4	B	B	C	A	B	A	A	A	B	A	A
XVI	13.0	G	B	C	A	B	A	A	A	B	B	A
XVII	2.0	N	B	C	A	B	A	A	A	B	A	A
No. of variants		14	3	5	3	5	2	6	4	4	3	3

[a] This population percentage is estimated from the pedigrees of members of the 3 yo racing population in 2003 and is not provided by the haplotype proportions derived from analysis of our DNA collection.

Previously, SSCP and sequencing analysis of the D-loop region has been used as a definitive means of assessing thoroughbred mtDNA variability (Marklund et al., 1995; Hill et al., 2002). Importantly, Hill et al. (2002) showed that a significant proportion of thoroughbred female lines carried a D-loop variant that did not conform to studbook expectations. They attributed these findings to inaccuracies in the recording of pedigree data or to de novo mutations. These irregularities dictate that it would not be feasible to predict common performance trends amongst thoroughbred female lines unconfirmed in origin or relationship by mtDNA analysis.

Additionally, however, it is also not possible to rely upon D-loop analysis as a sole means of describing variability. Although we identified a total of 17 haplotypes in the thoroughbred, there were only 14 D-loop variants found. We have observed that variation in the, potentially, more athletically import-ant protein encoding genes is not fully described by the, supposedly more hypervariable, D-loop region. For the purposes of this study, its isolated use as a basis for assessing potential mtDNA/racing

performance correlations or clarification of female line origins would have been limited. There are also implications for the assessment of mtDNA variation/association studies in other species when an examination is based on variability at a single locus. Possession of the same D-loop variant by horses did not guarantee that they did not vary at other loci (Table 2). In particular, the Types II, III and XV, carrying the same D-loop variant 'B', could be distinguished through variation at the MTATP6 and MTND5 loci. Similarly, horses carrying the same functional gene variants sometimes possessed differ-ent D-loop versions. This included haplotypes VII and VIII and haplotypes XV and XVII. Therefore, determination of mtDNA status solely by D-loop could lead to misinterpretation of functional gene variation in up to 34.8% of the racing population.

From analysis of the pedigrees of the horses in our sample we determined that 33 different female lines were categorised. Importantly, we were able to ascertain, through use of the greater discerning power of combined loci typing, that past inaccuracies in the recording of thoroughbred studbook registration

Table 3a
Percentage distribution of haplotypes in: TB (thoroughbred); IDH (Irish draught horse); PASO (Peruvian paso); TIBETAN (Tibetan riding horse); POLISH (Polish tarpan)

	N	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XV-II	POL 1	ID 1	TI-T8	PAS 1
TB	1000	19.7*	16.1*	5.4*	16.1*	1.0*	0.7*	1.4*	2.6*	2.8*	4.3*	1.7*	0.4*	0.4*	0.2*	9.3*	14.5*	1.4*	0.0	0.0	0.0	0.0
IDH	56	7.0*	7.0*	1.8*	11.0*	21.4*	0.0	0.0	5.3*	9.0*	0.0	1.8*	0.0	0.0	1.8*	28.6*	0.0	0.0	3.5*	1.8*	0.0	0.0
PASO	22	0.0	0.0	0.0	0.0	9.0*	0.0	27.5*	4.5*	0.0	9.0*	0.0	4.5	4.5*	0.0	41.0*	0.0	0.0	0.0	0.0	0.0	4.5*
TIBETAN	23	8.7*	0.0	0.0	4.3*	13.0*	0.0	0.0	0.0	26.0*	0.0	0.0	0.0	0.0	8.7*	4.3*	0.0	0.0	0.0	0.0	35.0*	0.0
POLISH	21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	29.0*	0.0	0.0	0.0	0.0	38.0*	0.0	0.0	0.0	33.0*	0.0	0.0	0.0

In the thoroughbreds, percentages are based on the haplotypes detected by DNA analysis not on estimates in the 3 yo population via pedigree assessment. *Shows the presence of a haplotype. **Variant structure for these haplotypes in shown in Table 3b.

are more widespread than indicated using D-loop analysis alone. Out of 19 female lines examined, Hill et al. (2002) identified eight containing some horses with unexpected mtDNA haplotypes. In the lines analysed in this study, 28 incorrect sub-branches were identified by virtue of anomalous mtDNA inheritance and these were spread over 19 of the 33 lines examined. All of the anomalous female lines possessed haplotypes shared with other lines, indicating a confounding role for studbook inaccuracy rather than mutational events.

The extent of the sample size provided near saturation coverage of all correct and incorrect female branches exhibited by the current population of horses racing in the UK (98.9%). This has made it possible to extrapolate from pedigree data the mtDNA haplotype of all horses belonging to these lines and to make corrections when necessary. A mechanism is provided whereby it is possible to investigate relative performance merits of haplotypes and to study effectiveness at different racing distances. The frequency of each haplotype varies considerably. An estimation of the percentage occurrence of each in the UK thoroughbred three years old population is shown in Table 2.

The abundance of genetic types was contributed to by the occurrence of between 2 and 6 variants at each variable, functional gene locus. The thoroughbred has been bred for over 200 years for racecourse performance and these observations might suggest a reduced likelihood of coincidental selection of any specific mtDNA haplotype or gene variant. However, different horses perform to varying degrees of competence in races run over a variety of distances and age groups that may require specific physiological attributes. Therefore, it is possible that variability at mtDNA loci may contribute to differential stamina potential.

Using pedigree information, it was possible to assign mtDNA haplotype data for all winners of major UK three years old (3 yo) horse races run between 1954 and 2003. From this we calculated the percentage winning success of each haplotype for each race during this period. Dividing the percentage of wins by the percentage occurrence of each haplotype in the general 3 yo population provided a success index (Race Index — RI) for each race. Race Indices were also calculated for each gene variant

Table 3b
Variant structure of haplotypes specific to non-thoroughbred

Type	D-loop	CYB	ATP 6	ND 2	ND 4	ND4L	ND5	CcO1	CcO2	16S	12S
POL1	N	A	A	A	D	A	F	A	E*	A	A
PAS1	O*	A	F*	A	C	B	A	C	D	A	B
ID1	C	A	A	A	E	B	A	E*	D	A	A
T1	P*	A	D	A	F*	B	A	F*	F*	A	A
T2	P*	A	D	A	G	B	A	D	D	A	A
T3	T*	A	C	A	B	A	A	C	B	A	A
T4	Q*	A	A	A	D	A	G*	B	A	A	A
T5	Q*	A	A	A	D	A	A	G*	G*	A	A
T6	R*	A	A	A	B	A	H*	C	B	A	A
T7	S*	A	A	A	C	A	A	A	D	A	A
T8	S*	A	A	A	C	B	B	A	D	A	A

*An asterisk labels variants not found in thoroughbreds.

carried by the winners of these races. Three years old races provide the most informative and measurable data compared with other age group races. They represent a range of racing distance, they are mainly single sex and are also run generally with each animal carrying equal weight.

Correlation coefficients, relating RI and race distance for haplotypes occurring at more than 2% of the 3 yo population (Table 4), showed that for five of them (accounting for 51.6% of the total 3 yo population) there was a significant relationship (Fig. 1). Haplotypes II, XV and XVI showed negative correlations, significant at $P < 0.05$. Two haplotypes were positively correlated. Type XI exhibited a correlation coefficient significant at $P < 0.05$, whereas Type IV was highly significant ($P < 0.001$).

For each of these haplotypes, ANOVAs were applied using their average RI grouped according to race distance. At two extremes of distance scored (1400–1600 and 2400 m, the one race at 2800 m was omitted) there were significant differences between haplotypes in average RI. Haplotypes for which RI negatively correlated with distance scored significantly higher at the shorter distances (XVI-1.42; II-1.27; XV-1.15; IV-0.84; XI-0.24; $P < 0.001$). However, at the longer distance, this was reversed and the positively correlated haplotypes scored higher (IV-1.34; XI-1.25; XVI-1.15; II-0.85; XV-0.54; $P = 0.002$). This indicates that there is a true order of racing merit amongst these haplotypes, which changes depending on the distance of race under consideration (Fig. 1).

Though in a minority over the relatively large number of races scored, there are instances were multiple race winners have occurred. It is feasible to suggest that this could result in a form of 'nested' replication. However, the issue of horses winning more than one race can present a complicated statistical picture unless a strict and consistent rule for analysis is employed. This is difficult to implement constructively. Removal or replacement of multiple winners in the analyses would represent a corruption of the random nature of the sampling procedure and it is not possible to apply a rule for omission that would not instigate a potential bias.

Moreover, in order to estimate the contribution of the mtDNA haplotype in relation to prevailing racing and genetic backgrounds, each race needs to be considered as a separate statistical entity and multiple winners must be included. The high number of races scored is sufficient to indicate trends within the haplotypes. Most importantly, inclusion of multiple winners does not ignore the effects of interacting chromosomal genetic backgrounds and hence artificially increase or reduce the apparent role of the mtDNA.

In a study assessing elite endurance and sprint human athletes, mitochondrial 'haplogroups' were defined by functional DNA variation (Niemi and Majamaa, 2005) and significant haplogroup/stamina correlations were demonstrated. The current study also provides evidence for similar haplotype trends in thoroughbred racehorses. It is not possible, at present, to state whether these effects are due to direct gene influence or to differential responses to training.

Similarly, without extended physiological studies, it is difficult to draw firm conclusions about the

Table 4
Average race index (RI) for each haplotype at different racing distance

Race distance (m)	Races at each distance	Total races scord	Haplotype												
			I	II	III	IV	VII	VIII	IX	X	XI	XII	XV	XVI	XVII
1400–1600	9	444	0.74	1.28	0.93	0.84	1.33	1.08	1.15	0.36	0.21	1.65	1.12	1.42	1.42
2000	3	141	0.98	0.96	0.50	1.41	0.95	1.36	0.69	0.90	1.06	0.58	0.70	1.31	0.44
2400	8	400	0.81	0.85	1.16	1.34	1.25	1.25	0.80	0.82	1.25	1.46	0.54	1.15	0.83
2800	1	50	0.73	0.85	1.67	1.57	1.82	2.00	1.82	0.31	0.00	0.83	0.54	0.62	2.00
Percentage in population			19.20	11.80	8.40	16.60	2.20	2.00	2.20	6.40	2.80	2.40	7.40	13.00	3.00
Correlation coefficient (r)[a]			0.097	0.447	0.289	0.684	0.023	0.087	0.081	0.328	0.478	0.114	0.469	0.519	0.211
Significance (P)[a]				<0.05		<0.001					<0.05		<0.05	<0.05	<0.05

[a] Based on RI for all races, not average RI of each racing distance.

relative contribution of the different loci. However, there are some important indicators regarding candidates for further investigation, which might be derived from an assessment of the molecular relationships between haplotypes.

The thoroughbred is a composite breed and the various maternal lines have diverse geographical and founder breed origins established prior to the formation of the thoroughbred studbook. Also, as demonstrated by the range of haplotypes found in non-thoroughbreds (Tables 3a and b), the breed does not include all potential mitochondrial genetic variations. Because of this, it is difficult to draw firm phylogenetic conclusions about haplotype derivations and relationships per se. However, Fig. 2 shows a UPGMA relationship tree illustrating the similarities and differences between haplotypes based on sequence data for the variable, translatable functional genes, which may be expected to have the strongest influence on athletic phenotype. The relationships derived from the sequence data corroborate the patterns shown in total variant similarity at these loci. Moreover, these data show the presence of five distinct clusters (labelled C1–C5), which provide some indication of the more important loci.

C1 contains only one haplotype (XII) and is quite distinct from the other clusters in terms of its variant constitution. However, the differences are less pronounced in regard of sequence variation and a closer relationship is seen to exist with C2–C4. Apart from this, all of the other haplotypes fall into clusters that conform to one of two broad variant 'skeletons' and include C2–C4 and C5, respectively.

Closer observation shows that cluster C5 is more molecularly distinct from the others and contains the haplotypes with negative RI/stamina correlations (XV, XVI, II). For these haplotypes, the functional gene variants are the same, except for Type II, which differs only at the *MTND5* locus. Two other types (VI and XVII) also share the same functional gene variants. Notably, they all differ from the other haplotypes through common possession of the 'C' variant at the *MTATP6* locus. Combination of data from haplotypes carrying this variant gives rise to a strong negative correlation ($r = -0.7070$) with a significance of $P < 0.001$. The distinct molecular nature of cluster C5 might suggest that the inclination towards greater success at shorter distances could be
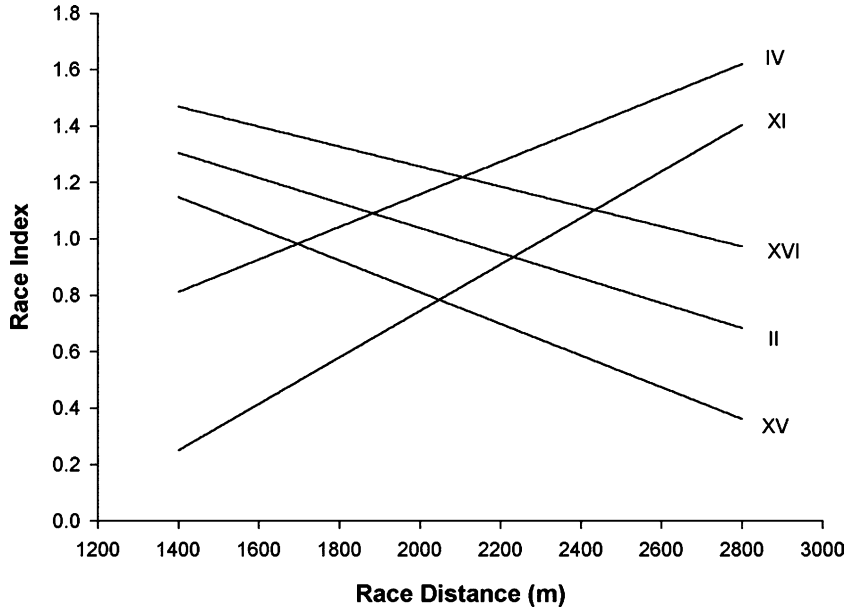
Fig. 1. Regression lines describing the relationship of RI vs. race distance for five haplotypes. Each regression is significant (II, $P<0.05$; IV, $P<0.001$; XI, $P<0.05$; XV, $P<0.05$; XVI, $P<0.05$) but also clearly illustrates the shift in their order of performance merit at different distances.
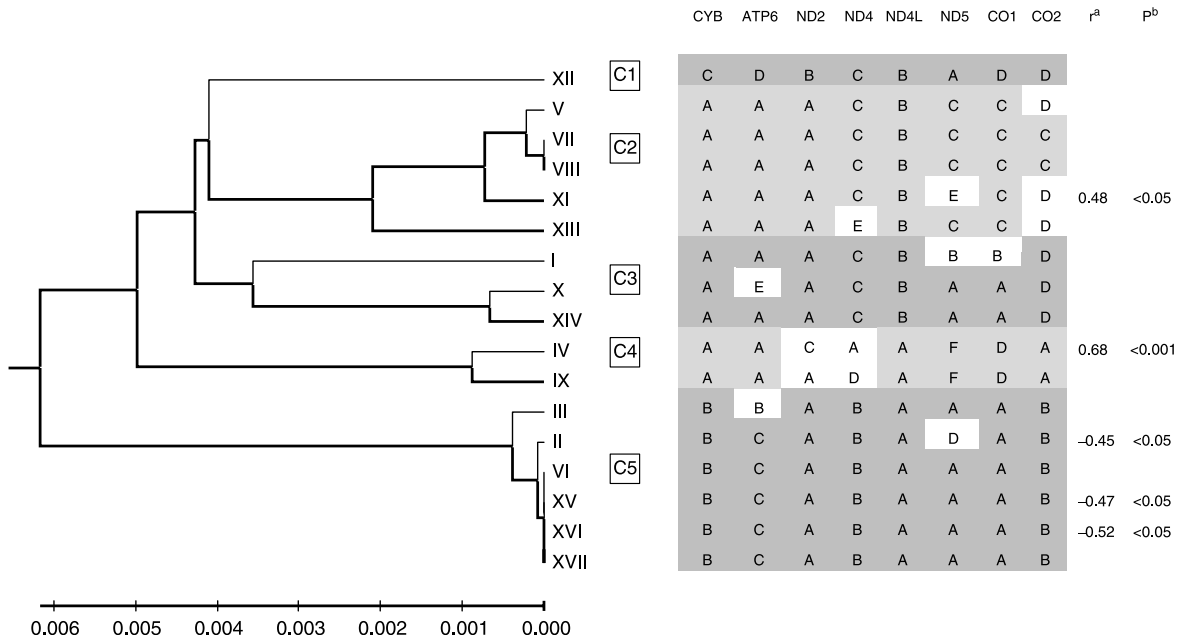

Fig. 2. UPGMA relationship tree based on sequence variation between different variants of translatable genes. Variants are provided in letter format for comparison. For both the tree and the variant profiles, the presence of five distinct clusters is indicated by shading and assignment of the labels C1–C5. r^a, correlation with racing distance; P^b, significance of the correlation.

reliant on the presence of any or a combination of specific variants at the four loci (MTCYB, MTATP6, MTND4, MTCO2) at which members of this cluster generally differ uniformly from the other haplotypes. However, haplotype III, which is unique in this cluster in having a moderate positive RI/distance correlation ($r = 0.2891$), also shares the same variants as the bulk of this combined group but importantly, differs in carrying a unique variant, 'B' at the *MTATP6* locus. There is some evidence, therefore, that variation at the MTATP6 locus may be of note and that possession of the 'C' variant, rather than acting as a hindrance to stamina potential, may, in fact, confer a positive adaptation supporting enhanced performance over shorter distances. It is possible that the product derived from variant 'C' has a beneficial effect on efficiency of energy partitioning between the aerobic and anaerobic respiratory pathways (Eaton et al., 1992) in shorter races.

Table 5 shows that the 'B' and 'C' variants have two non-synonymous nucleotide differences at positions 8035 and 8309. At the former position, variant 'B' has thymine and variant 'C' adenine, resulting in translation of isoleucine and asparagine, respectively. At the latter, variant 'B' has guanine and variant 'C' adenine, resulting in translation of methionine and isoleucine. For the purposes of this study, the limitations of restricted D-loop haplotyping is further emphasized, as the significant stamina differences existing between haplotypes II and XV (negative correlation) and Haplotype III that have the same D-loop variant, would be overlooked using D-loop analysis alone.

The molecular relationship between haplotypes exhibiting positive RI/distance correlation is less clear. Although more related to each other, by virtue of DNA sequence homology than to the members of cluster C5, the two haplotypes (IV and XI) differ from each other at six out of eight functional gene loci and have no positions at which they commonly vary from less correlated types. However, both have loci at which they carry unique variants, though they are not the same loci.

Type IV has unique variants at the *MTND2* and *MTND4* loci. At all other loci it is the same as Type IX, which has a slight negative correlation of RI vs. distance. At MTND2 Type I possesses variant 'C', which has a non-synonymous nucleotide difference with variant 'A', carried by Type IX, at position 4020 (A for G, isoleucine for methionine, Table 5) but not

variant 'B'. Despite this, as variant 'A' is carried by all other haplotypes except XII, which also varies at many other loci, variation at the MTND2 locus cannot necessarily be excluded in having an effect on extended stamina potential. At MTND4, variant 'A', specific to Type IV, carries non-synonymous substitutions at a number of positions (Table 5). At many of these positions, it differs not only in respect of variant D (carried uniquely by Type IX) but also in relation to variants possessed by other haplotypes. The combination of non-synonymous substitutions present in the 'A' variant ensure that the product from this variant is different from those of any of the others.

For Type XI, the *MTND5* locus has particular significance. It has a unique allele at this locus but is otherwise similar to haplotypes showing positive but lesser and non-significant, correlations and regressions. In terms of non-synonymous substitutions, variant 'E', carried by Type XI at MTND5, differs from all of the other variants at position 12890 (C for A, proline for threonine) and 13082 (G for C, Glutamate for Glutamine) (Table 5).

Although the primary indications are that the molecular mechanisms for the positive RI/distance correlations of Types IV and XI may be different, observations regarding the variable loci of Types IV and XI may be of importance. In humans, variants of the genes at *MTND2*, *MTND4* and *MTND5* have been shown to differentially affect VO_2 max, initial fitness and responses to training (Wolfarth et al., 2005; Dionne et al., 1993). A mutation in MTND4 has also affected exercise intolerance (Andreu et al., 1999). It is also feasible, therefore, that non-synonymous substitutions resulting in significant conformational change to the protein products of these genes may play a role in affecting similar phenotypes in horses.

Further consideration of Tables 3a and b shows that, although there are some haplotypes and gene variants that are specific to the non-thoroughbreds, for the most part they share thoroughbred haplotypes. However, application of the Fisher's exact test indicates that there is a significantly different distribution of haplotypes across the breeds ($P < 0.0001$).

Predictably, members of the Irish draught horse breed showed a greater degree of haplotype commonality to the thoroughbred, though relative distribution clearly varied. Ten of twelve haplotypes found in the Irish draught were also present in the thoroughbred

Table 5
Non-synonymous nucleotide substitutions at loci implicated in stamina determination

Variant	Position					
	8035	8047	8152	8177	8309	8491
(a) MTATP 6						
A	A	A	A	G	A	G
B	T	A	G	A	G	G
C	A	A	G	A	A	G
D	T	G	G	A	A	A
E	T	G	G	A	A	G
Amino acid	A=N	A=N	A=N	G=M	A=1	G=G
Substitution	T=1	G=S	G=S	A=1	G=M	A=E

Variant	Position													
	10404	10409	10488	10493	11066	11160	11185	11186	11188	11205	11210	11225	11238	11242
(b) MTND4														
A	T	G	A	A	G	C	C	T	C	A	T	T	C	T
B	T	A	A	A	T	C	C	T	C	A	C	A	C	T
C	A	G	G	G	G	G	A	A	T	G	T	T	T	T
D	T	A	A	A	T	C	A	T	C	A	T	A	C	T
E	T	A	A	A	T	C	C	T	C	A	C	A	A	C
Amino acid	T=V	G=A	A=Y	A=T	G=D	C=H	C=F	T=C	C=L	A=Y	T=C	T=Y	C=A	T=E
Substitution	A=E	A=T	G=C	G=A	T=Y	G=Q	A=L	A=S	T=W	G=C	C=R	A=T	T=V, A=D	C=X

Variant	Position					
	12257	12890	13082	13086	13122	13128
(c) MTND5						
A	T	A	C	C	A	T
B	C	A	C	C	A	T
C	T	A	C	C	G	T
D	T	A	C	A	A	T
E	T	C	G	C	G	T
F	T	A	C	C	A	G
Amino acid	T=Y	A=T	C=Q	C=P	A=N	T=F
Substitution	C=H	C=P	G=E	A=H	G=S	G=C

Variant	Position
	4020
(d) MTND2	
A	G
B	A
C	A
Amino acid	G=M
Subtitution	A=1

and these constituted 94.7% of the population. Interestingly, the most common haplotypes were V and XV, the former occurring in only one per cent of thoroughbred samples and the latter in nine percent. Certainly, the low proportion of Type V in the thoroughbred population indicates that it has not undergone preferential selection for racing ability but the high proportion in the Irish draught suggests that it has not been detrimental to the development of that breed. As the proportions of the haplotypes in the original founder populations of the non-thoroughbreds are not known, it is not possible to draw firm conclusions about positive haplotype selection from the current data. The non-thoroughbreds have not

been consistently selected for any clear sporting activity and therefore, presence or absence of a haplotype does not necessarily imply that selection has occurred. It is possible that discrepancy in the occurrence of Type XV between thoroughbred and Irish draught could be due to co-incidental selection of types with greater stamina potential in the thoroughbred though this evidence is inconclusive.

The other non-thoroughbred breeds also exhibit a population bias towards the haplotypes of lesser relevance in the thoroughbred. The Peruvian paso horse has a large proportion of Type XV and Type V. Type VII is also represented strongly, whilst occurring as a minority haplotype in the thoroughbred and not undergoing any obvious preferential selection for racing potential or showing any stamina bias.

The geographically distinct Tibetan riding horses and Polish tarpans have high percentages of haplotypes not found in the thoroughbred. Both exhibit a large proportion of Type IX and the latter, especially, a particularly high proportion of Type XIV. In the thoroughbred, it is clear that these haplotypes have not been selected and have no stamina bias. Like the Irish draught, Type V is relatively prominent in the Tibetan horses. Type XV, present in the Irish draught and paso, is absent from the tarpan and Tibetan group, suggesting that occurrence in the former may be as a result of closer historical relationship with the thoroughbred. In summary, it may be stated that the haplotypes occurring at higher levels, or showing stamina biases, in the thoroughbred have not been preferentially selected in non-thoroughbreds and vice versa.

Many thoroughbred breeders base breeding strategies on pedigree assessments, attempting to co-ordinate or balance the perceived stamina capabilities of parents and ancestors in the pedigree to achieve specific stamina objectives in the progeny. Over 50% of the 3 yo thoroughbred population belong to mitochondrial haplotypes that exhibit a significant leaning towards success at particular stamina extremes. These observations lend support to there being an important female component contributing to stamina optima which should be taken into account when planning thoroughbred breeding strategies.

References

Andreu, A.L., Tanji, K., Bruno, C., Hadjigeorgiou, G.M., Sue, C.M., Jay, C., Ohnishi, T., Shanske, S., Bonilla, E., DiMauro, S., 1999. Exercise intolerance due to a nonsense mutation in the mtDNA ND4 gene. *Ann. Neurol.* 45, 820–823.

Bobinski, K., 1953. *Family Tables of Racehorses*. Zamoyski, London.

Caetano-Anolles, G., Bassam, B.J., Gresshoff, P.M., 1991. DNA amplification fingerprinting using very short arbitrary oligonucleotide primers. *Biotechnology* 9, 553–557.

Dionne, F.T., Turcotte, L., Thibault, M.C., Boulay, M.R., Skinner, J.S., Bouchard, C., 1993. Mitochondrial DNA sequence polymorphism, VO2max, and response to endurance training. *Med. Sci. Sports Exerc.* 25, 766–774.

Eaton, M.D., Rose, R.J., Evans, D.L., 1992. The assessment of anaerobic capacity of thoroughbred horses using maximal accumulated oxygen deficit. *Aust. Equine Vet.* 10, 86–92.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.

Harkins, J.D., Beadle, R.E., Kamerling, S.G., 1993. The correlation of running ability and physiological variables in thoroughbred horses. *Equine Vet. J.* 25, 53–60.

Harris, R.C., Marlin, D.J., Snow, D.H., 1987. Metabolic response to maximal exercise of 800 and 2,000m in the thoroughbred horse. *J. Appl. Physiol.* 63, 12–19.

Hill, E.W., Bradley, D.G., Al-Barody, M., Ertugrul, O., Splan, R.K., Zakharov, I., Cunningham, E.P., 2002. History and integrity of thoroughbred dam lines revealed in equine mtDNA variation. *Anim. Genet.* 33, 287–294.

Ishida, N., Hasegawa, T., Takeda, K., Sakagami, M., Onishi, A., Inumaru, S., Komatsu, M., Mukoyama, H., 1994. Polymorphic sequence in the D-loop region of equine mitochondrial DNA. *Anim. Genet.* 25, 215–221.

Jansen, T., Forster, P., Levine, M.A., Oelke, H., Hurles, M., Renfrew, C., Weber, J., Olek, K., 2002. Mitochondrial DNA and the origins of the domestic horse. *Proc. Natl Acad. Sci.* 99, 10905–10910.

Kukita, Y., Tahira, T., Sommer, S.S., Hayashi, K., 1997. SSCP analysis of long DNA fragments in low pH Gel. *Human Mutat.* 10, 400–407.

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings Bioinformat.* 5, 150–163.

Leicester, C., 1983. *Bloodstock Breeding*. Revised by H. Wright., second ed. Allen and Co. Ltd., London.

Lesage, R., Simoneau, J.A., Jobin, J., Leblanc, J., Bouchard, C., 1985. Familial resemblance in maximal heart rate, blood lactate and aerobic power. *Hum. Hered.* 35, 182–189.

Lowe, C.B., 1898. In: Allison, W., Jenkins, W.R. (Eds.), *Breeding Racehorses by the Figure System*. Veterinary Publisher, New York.

Marklund, S., Chaudhary, R., Marklund, L., Sandberg, K., Andersson, L., 1995. Extensive mtDNA diversity in horses revealed by PCR-SSCP analysis. *Anim. Genet.* 26, 193–196.

Nei, M., Kumar, S., 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York, pp. 87.

Niemi, A., Majamaa, K., 2005. Mitochondrial DNA and ACTN3 genotypes in Finnish elite endurance and sprint athletes. *Eur. J. Hum. Genet.* 13, 965–969.

Perusse, L., Gagnon, J., Province, M.A., Rao, D.C., Wilmore, J.H., Leon, A.S., Bouchard, C., Skinner, J.S., 2001. Familial aggregation of submaximal aerobic performance in the Heritage Family study. *Med. Sci. Sports Exerc.* 33, 597–604.

Raceform, 2003. *Flat Annual for 2003*. Raceform Ltd. Newbury, Berkshire, UK.

Vila, C., Leonard, J.A., Gotherstrom, A., Marklund, S., Sandberg, K., Liden, K., Wayne, R.K., Ellegren, H., 2001. Widespread origins of domestic horse lineages. *Science* 291, 474–477.

Weatherby and Sons, 1791. *An Introduction to a General Stud Book*. Weatherby and Sons, (London).

Wolfarth, B., Bray, M.S., Hagberg, J.M., Perusse, L., Ruaramaa, R., Rivera, M.A., Roth, S.M., Rankinene, T., Bouchard, C., 2005. The human gene map for performance and health-related fitness phenotypes: The 2004 update. *Med. Sci. Sports Exerc.* 37, 881–903.

Xu, X., Arnason, V., 1994. The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene* 148, 357–362.