History and integrity of thoroughbred dam lines revealed in equine mtDNA variation

E. W. Hill*, D. G. Bradley*, M. Al-Barody†, O. Ertugrul‡, R. K. Splan§, I. Zakharov¶ and E. P. Cunningham*

*Department of Genetics, Smurfit Institute of Genetics, Trinity College, Dublin, Ireland. †Animal Production Department, Faculty of Agriculture, Minia University, Minia, Egypt. ‡Veterinary Faculty, Ankara University, Ankara, Turkey. §Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. ¶Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

Summary

Mitochondrial DNA (mtDNA) D-loop sequences (381 bp) from 100 thoroughbreds in 19 of the most common matrilineal female families were used to reconstruct a founder female population for the thoroughbred (~1650–1750 AD). Seventeen haplotypes were found to have contributed to the 19 female lineages. In order to place the reconstructed founder population in wider historical context, we examined, using both single strand conformation polymorphism and direct sequence analysis, variation in a 343 bp mtDNA fragment in that population and 13 other horse populations of disparate provenance. Interpopulation diversity analyses revealed no significant difference in variation between the thoroughbred founder population and the 13 other diverse horse populations and suggested a non-random partitioning of diversity among geographically diverse horse populations. Within thoroughbreds, almost half of the female families, which are each considered from pedigrees to have matrilines converging to one of >30 historically recognized female ancestors, contained sequences which were inconsistent with common descent. Examination of the anomalies in the context of pedigree records suggests the majority might be best explained as confusion of mares at the foundation stages, although some could have some bearing on more recent (19th century – 1980) lineages. We have used this combination of molecular and historical information to identify some of the founder dams and to make new interpretations about the early history of the thoroughbred.

Keywords *Equus caballus*, horse, mitochondrial DNA, pedigree, population genetics, sequence, thoroughbred founder.

Introduction

The genesis of the thoroughbred in the 17th and 18th centuries owes much to the close association and enthusiasm of the Tudor and Stuart kings for horse racing. These monarchs, along with associated gentry, endeavoured to breed improved racehorses by crossing local running horses with imported Eastern stock (Willett 1975). The development of the thoroughbred was greatly encouraged by the appointment to the court of James I of George Villiers as Master of the Horse (1616 AD) and later of James Darcy as Master of the Royal Stud (1660 AD) under Charles II (Prior 1935). Following some minor importations to Britain of exotic horses in the early 1600s, significant numbers of horses with Arabian, Turk and Barb ancestry were introduced towards the end of the century (Wentworth 1938; Willett 1975). Although a General Stud Book (GSB) was

Address for correspondence

E. P. Cunningham, Department of Genetics,
Smurfit Institute of Genetics, Trinity College, Dublin 2, Ireland.
E-mail: epcnnghm@tcd.ie

Accepted for publication 17 February 2002

established in 1791 AD (Weatherby & Sons 1791), the precise origins of the thoroughbred still remain largely unclear mainly because of limited or inaccurate recording of pedigrees in the early foundation stages. Today, however, this stud book contains the oldest, most comprehensive record of all domestic animal pedigrees and is maintained under a strict registration process requiring, since the mid-1980s, verification of parentage by blood-typing and DNA analysis.

While only three predominant male thoroughbred lineages are recognized (Darley Arabian, Byerley Turk and Godolphin Arabian) pedigree analyses indicate that 30 founder mares have contributed to 94% of modern maternal lineages (Cunningham *et al.* 2001). The GSB listed 78 of the earliest known mares but acknowledges the presence of numerous errors and duplications and there remains much speculation about thoroughbred maternal ancestry.

A system (Bruce Lowe's Family Figure System) for the classification of thoroughbred female lines was devised by the end of the 19th century that traced all mares in the GSB at that time as far back as possible in direct maternal descent to one of 43 founder mares, descendants of which are considered a *family*. These are ranked in order of the number of Classic race winners prior to 1890 that were members of that family and named according to that rank, i.e. *Family 1* had the most Classic race winners, *Family 2* the second most, etc. (Lowe 1913). Today, *family* assignment is often considered an important indicator of genetic value within the multibillion dollar thoroughbred industry. Also, specific mating combinations between *families* are sometimes considered more or less auspicious than others. We have used this system to identify descendants of 19 of the most common female *families* (Table 1) in a large resource of thoroughbreds.

Because of organelle maternal inheritance (Hutchinson *et al.* 1974), mitochondrial DNA (mtDNA) haplotypes should be shared by all individuals within a *family*. Stability of maternal inheritance within documented horse pedigrees has been demonstrated in both Lippizan (> 200 years) (Kavar *et al.* 1999) and Arabian (~100 years) (Bowling *et al.* 2000) horses.

Materials and methods

Mitochondrial DNA sequences were determined by direct sequencing or inferred by comparative single strand conformation polymorphism (SSCP) analysis for 100 thoroughbreds (blood samples) representing 19 female *families*. A further 81 sequences were generated in the same way for individuals (hair samples) from seven other horse populations [Far East: Tuva ($n = 11$); Near East: Anatolian ($n = 13$), Cukorova ($n = 12$), Egyptian ($n = 8$), Fulani

Table 1 Thoroughbred founder females considered by historical pedigree records and by genetics illustrating the extent of sequence sharing among families and the occurrence of anomalous sequences within families. In families with more than one sequence, the founder haplotype is indicated by an asterisk.

Family number	Founder mare	Approximate date	*n*	Haplotypes	Type of anomaly
1	Tregonwell's Natural Barb Mare	1657–1670	9	F, H*	MOD
2	Burton's Barb Mare	1660–1685	7	F	–
3	Dam of the Two True Blues	1690	6	E	–
4	Layton Barb Mare	1650	10	J	–
5	Massey Mare	1714	4	L*, M	DR
6	Old Morocco Mare	1656	3	C*, N	DR
7	Lord Darcy's Blacklegged Royal Mare	1710	5	F	–
8	Bustler Mare	1680	6	F	–
9	Old Spot Mare	1700	10	A*, G	DR
10	Grey Childers Mare	1741	1	B	–
11	The Pet Mare	1697	4	J*, L, P	MOD & MUT
12	Royal Mare	1700	3	G*, Q	MUT
13	Sedbury Royal Mare	1665	6	J	–
14	Oldfield Mare	1695	7	D	–
16	Hutton's Old Spot Mare	1695	8	F*, H	MOD
17	Byerley Turk Mare	1700–1710	2	F	–
19	Davill's Woodcock Mare	1690	7	K*, O	MOD
22	Belgrade Turk Mare	1718	1	F	–
25	Brimmer Mare	1699	1	I	–

MOD: Relatively recent anomaly in modern pedigree; DR: deep rooted anomaly, possible foundation stage confusion; MUT: possible *de novo* mutation; *Founder haplotype.

($n = 11$); Europe: Connemara ($n = 12$), Shetland ($n = 14$)]. DNA extractions were carried out using a standard phenol/chloroform extraction protocol.

Polymerase chain reaction (PCR) primers for the mitochondrial hypervariable region were designed based on published horse sequence (Xu & Arnason 1994). The entire mtDNA D-loop region between nucleotides 15 351 and 30 was amplified in one representative (family reference) from each thoroughbred *family* ($n = 19$) and all individuals ($n = 81$) in the other breeds using primers FP 5′-CAC TGA AAA TGC CTA GAT GA-3′ and RP 5′-ACA CCA GTC TTG TAA ACC AG-3′. PCR was performed in a 20-μl reaction: 1× Platinum® *Taq* PCR buffer (20 mM Tris–HCl, 50 mM KCl), 1.5 mM $MgCl_2$, 1 μM of each primer, 0.2 mM of each dNTP, 2 U Platinum® *Taq* DNA polymerase (GibcoBRL®, Invitrogen Corp., Calsbad, CA, USA), distilled H_2O. Thermal cycling was carried out on an M J Research PTC100 thermal cycler at an annealing temperature of 55 °C.

DNA sequencing of the PCR product between nt 15 437 and 15 847 was performed using a SequiTherm EXCEL™ II DNA Sequencing Kit (Epicentre Technologies, Madison, WI, USA) with infrared dye (IRD) labelled internal sequencing primers: IRD700 5′-CTA GCT CCA CCA TCA ACA CC-3′ and IRD800 5′-ATG GCC CTG AAG AAA GAA CC-3′. The IRD labelled chain-terminated fragments were separated according to size on an acrylamide gel on a LI-COR® 4200 automated sequencer (LI-COR Inc., Lincoln, NE, USA) and read using the LI-COR® Base ImagIR software (LI-COR Inc.). In total, 381 bp sequence (nt 15 456–15 837) was generated and analysed in thoroughbreds. A shorter 343 bp (nt 15 476–15 818) stretch was generated for other populations. When inter-population comparisons were performed, the shorter fragment was used. Sequences have been deposited in GenBank (accession numbers AF481232–AF481334).

SSCP analysis was carried out, for thoroughbreds only, using primers FP 5′-TAG CTC CAC CAT CAA CAC C–3′ and RP 5′ -GCT GAT TTC CCG CGG CTT GG-3′ which amplified nt 15 437–15 743. PCR was performed in a 10-μl reaction: 1× Platinum® *Taq* PCR buffer (20 mM Tris–HCl, 50 mM KCl), 1.5 mM $MgCl_2$, 1 μM of each primer, 0.2 mM of each dNTP, 0.5 μCI α-^{32}P dCTP, 2 U Platinum® *Taq* DNA polymerase (GibcoBRL®), distilled H_2O. Thermal cycling was carried out on an M J Research PTC100 thermal cycler at an annealing temperature of 55 °C. PCR products were run on an acrylamide gel containing glycerol at 4 °C at 8 W for 14 h. Individuals from the same *families* were loaded beside each other, including the family reference. All gels were analysed by direct observation following exposure to X-ray film. When an SSCP pattern different to the family reference was observed within a *family*, direct sequencing of the sample was performed.

For greater geographical reference, a further 28 sequences, from six additional populations [Far East: Mongolian ($n = 4$), Cheju ($n = 7$), Tsushima ($n = 2$), Yunnan ($n = 2$); Europe: Lippizan ($n = 10$), Belgian ($n = 3$)] were taken from GenBank (accession numbers – Belgian: AF064630–2; Cheju: AF014405–8, AF014410–12; Lippizan: AF168689–98; Mongolian: AF014413–15, AF056071; Tsushima: AF169009–10; Yunnan: AF014416–17). Complete data for all 14 populations were analysed for 326 nucleotide sites in 343 bp (missing data and indels were ignored) between nt 15 476 and 15 818 (Xu & Arnason 1994).

Sequences were aligned using the ClustalX Multiple Sequence Alignment Program (version 1.81) (Thompson *et al.* 1997). Ignoring gaps in the sequence a neighbour-joining phylogeny was created in ClustalX and viewed in TreeView. The probability of identity (PI), the probability that two individuals in a population share an identical haplotype, was estimated as the sum of the product of the frequency of each haplotype in that population ($PI = \sum a_k^2$ where a is the frequency of the kth haplotype) (Bowling *et al.* 2000). All other diversity statistics were calculated in Arlequin version 2.0 (Schneider *et al.* 2000).

Results

In the thoroughbred, 39 polymorphic sites (three indels, 35 transitions, one transversion) in a 381 bp mtDNA D-loop fragment defined 19 haplotypes in the 19 female lineage *families*. When insertion/deletion events were ignored, 17 haplotypes were found to have contributed to the 19 female lineages (Table 2). However, only 11 *families* conserved a single haplotype, i.e. more than one haplotype was detected within eight of the 19 *families* (Table 1). In six of these eight mismatched *families* there was one predominant sequence with a single mismatched sample. Anomalous sequences in two families (*Families 11* and *12*) differed from the numerically predominant (majority) sequence in the family by only one nucleotide substitution and it is possible that these two could be a result of *de novo* mutation. We suggest that the extent of nucleotide differences (>2–14 nt) between the majority and anomalous sequences in the other heterogenous families (*Families 1, 5, 6, 9, 11, 16* and *19*) are best explained by confusion between horses from either another family sharing the anomalous sequence, a family not represented in this sample, or a non-thoroughbred. One *family* (*Family 11*) had two aberrant sequences and in another *family* (*Family 9*) two sequences were represented approximately equally (11 nt divergent). Pedigree analysis has allowed some determination of the anomalies in time, indicating some possible foundation stage confusions where these occur at a potentially deep root in a pedigree

Table 2 Variable nucleotide positions in a 381 bp mtDNA D-loop fragment in 19 of the most common thoroughbred female families. Haplotypes A–L were considered founder haplotypes and were used to reconstruct a thoroughbred founder population. Haplotypes M–Q may have arisen by mutation or derive from a family founder not represented in this study. Reference sequence from GenBank X79547.

Haplotype	Nucleotide position	15465	15478	15494	15496	15521	15526	15534	15538	15540	15542	15584	15585	15596	15597	15601	15602	15603	15604	15617	15635	15649	15650	15659	15666	15684	15703	15709	15718	15737	15770	15771	15807	15810	15825	15826	15827
RefSeq		C	C	T	A	G	T	C	A	A	C	T	G	A	A	T	C	T	G	T	C	A	A	T	G	C	T	C	C	T	C	C	C	A	C	A	A
A		.	.	.	.	.	.	.	.	.	T	C	.	.	G	.	T	.	.	.	T	.	.	.	A	.	C	.	.	.	.	.	.	.	.	.	.	.
B		.	.	C	G	.	.	T	.	.	.	C	A	.	.	.	.	C	.	.	.	G	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.
C		.	.	C	G	.	.	T	.	.	.	C	A	.	.	.	.	C	.	.	.	G	.	.	.	T	.	.	.	.	.	T	.	.	.	.	G
D		.	.	C	G	.	.	T	.	.	.	C	A	.	.	.	.	C	.	.	.	G	.	.	.	.	.	.	.	.	.	T	.	.	.	.	G
E		.	T	C	G	.	.	T	.	.	.	C	A	.	.	.	.	C	.	.	.	G	.	.	.	.	.	.	.	.	.	T	.	.	.	.	G
F		.	.	C	G	.	.	T	.	.	.	C	A	.	.	.	T	C	A	.	.	G	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.
G		.	.	C	G	.	.	T	.	.	.	C	.	.	.	.	T	C	.	.	.	G	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.
H		.	.	.	.	.	.	.	.	.	.	C	A	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	G
I		.	.	.	.	.	.	.	.	.	.	C	.	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	G
J		.	.	.	.	.	.	G	.	.	.	.	A	.	.	.	T	.	.	.	.	.	G	.	.	.	.	T	.	.	.	T	.	.	.	G	.
K		.	.	.	.	C	.	.	G	.	C	A	.	.	.	T	.	.	.	.	.	G	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.
L		.	.	.	A	.	.	.	.	.	C	.	G	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	C	T	.	.	G	.	.	.
M		.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	G	.	G	.
N		.	.	.	.	.	.	.	.	.	C	.	.	.	.	T	.	.	C	.	.	.	C	.	.	.	.	.	.	.	.	T	T	.	.	.	G
O		.	.	.	.	.	.	G	.	.	C	.	G	.	.	T	.	.	.	.	.	G	.	.	.	.	T	.	.	.	.	T	.	.	.	G	.
P		T	.	.	.	.	.	G	.	.	.	A	.	.	.	T	.	.	.	.	.	G	.	.	.	.	T	.	.	.	.	T	.	.	.	G	.
Q		.	.	C	G	.	.	T	.	.	.	C	.	.	.	.	T	C	.	.	.	G	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.

(*Families* 5, *6* and *9*) or alternatively, more recent (19th century – 1980) confusions when placed within a relatively modern lineage (*Families 1, 11, 16* and *19*) (Fig. 1; Table 1).

In order to reconstruct a founder population (~1650–1750 AD) for the thoroughbred we assumed the majority haplotype within each maternal lineage *family* to represent the founder sequence. In *Family 9*, where two sequences were detected in approximately equal proportions, the haplotype that was not shared with another family was inferred (A). Following this assumption, we have identified 12 distinct founder haplotypes (A–L) in a reconstruction of the founder population. Ten of these founder haplotypes (A–E, G–I, K–L) were unique to single families and two (F, J) were shared by more than one family. The remaining five haplotypes may have arisen by mutation or derive from a family founder not represented in this study (M–Q). The most common thoroughbred haplotype (F) shared by six (*Families 2, 7, 8, 16, 17* and *22*) of the 19 founders (frequency: 0.32), was found twice as often as the most common haplotypes (mean = 0.15; range = 0.13–0.17) in four horse populations which approximate as representatives of ancestral source populations (Connemara, Anatolian, Cukorova, Egyptian) for the thoroughbred. The estimated PI of haplotypes in contemporary populations ranged from 0.09 to 0.24 and was 0.15 in the thoroughbred founder population.

In a wider equine comparison, 88 polymorphic sites (24 indels, 56 transitions, eight transversions) in a 343 bp fragment defined 92 horse sequences in the 19 thoroughbred families and 109 individuals from 13 geographically diverse populations. When indels were ignored, 64 polymorphic sites defined 74 distinct haplotypes. Using this mtDNA fragment, the phylogenetic relationships between modern horse sequences and thoroughbred founder haplotypes were reconstructed in a neighbour-joining phylogeny (Fig. 2). The labelled clades in this unrooted tree corresponded to those determined by Vila *et al.* (2001) although within this sequence no distinction was made between clades A and E. No clear geographic affiliation of clades was apparent except clade F, which almost exclusively contained sequences from the Far East (frequency: 0.54) and the Near East (0.46) with the exception of one European (0.08) haplotype. No thoroughbred sequences were found within this clade.

The mean number of pairwise differences observed among all horse sequences was estimated as 7.51 or 2.30%. If the average rate of equid mtDNA divergence is assumed between 4.1 and 8.1% per million years (Vila *et al.* 2001) then the average coalescence for modern horse mtDNA sequences is estimated between 280 000 (lower limit) and 560 000 (upper limit) years before present (YBP), dates consistent with previous estimates for the horse (Vila *et al.* 2001).

An ancient divergence of horse lineages is supported in an analysis of population structure. When populations were grouped according to geographical origin (Far East: Cheju, Mongolian, Tsushima, Tuva, Yunnan; Near East: Anatolian, Cukorova, Egyptian, Fulani; Europe: Belgian, Connemara, Lippizan, Shetland) in an analysis of molecular variance, the greatest partitioning of genetic variation was detected among individuals within populations (92.4%). Although a significant proportion of the variation (7.9%) is

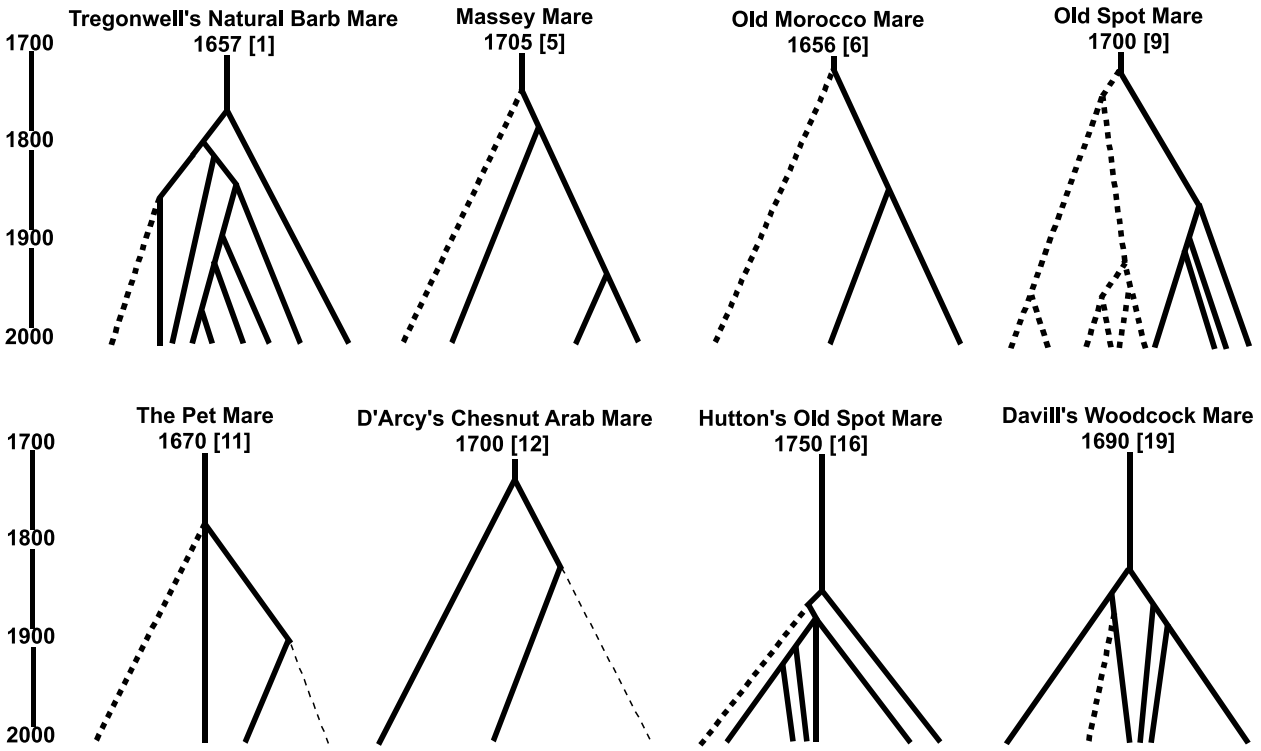

Figure 1 Pedigree trees for thoroughbred female families. The end of each branch represents a contemporary thoroughbred tracing in direct maternal lineage in the General Stud Book (GSB) to one the founder dams (name and approximate date of birth given), descendants of which are considered a *family*. Pedigrees are illustrated relative to a linear time scale. The anomalous lineages (thin broken line) in *Families 11 and 12* are possible *de novo* mutations. All other anomalous dam lines are indicated by thick broken lines to include all descendants of the most recent common ancestor between the anomalous and typical haplotypes. Descendants (lineages shown by thick broken lines) of Maid of the Glen 1858 (*1*), Hag 1744 (*5*), the dam(s) of Betty Percival 1715 and Cream Cheeks 1695 (*6*), a Curwen Bay Barb Mare 1708/1709 (*9*), Young Camilla 1787 (*11*), Lady Alice 1855 (*16*) and Violet 1858 (*19*) may contain a genetic heritage different to that which pedigree information suggests.

partitioned non-randomly among regional populations, no difference was detected among the three regionally distinct groups, despite a tentative eastern-specific clade in the phylogeny. No indication of regional clustering of variation was detected from pairwise Fst genetic distances between populations (Table 3).

No difference in nucleotide diversity was detected between the thoroughbred founder and contemporary horse populations. In the horse, a range between 1.81% sequence divergence in the Fulani and 3.10% in the Tuva was observed. Mean sequence diversity in the thoroughbred founder population (2.25%) was similar to estimates for the Anatolian (2.09%), Cheju (2.22%), Connemara (2.24%) and Cukorova (2.43%) populations.

Discussion

The examination of detailed pedigree records documenting the history of the thoroughbred, coupled with mtDNA sequence analysis, has facilitated the first in-depth investigation of the founder mares of the thoroughbred. Genetic diversity estimates in the thoroughbred female founders are not dissimilar to observed estimates in contemporary horse populations (Table 3). However, the examination of horse population diversity reveals a consistent absence of geographical structure and a lack of phylogenetic sorting of haplotypes into divergent but inwardly invariant groups as seen in other large domesticates (MacHugh & Bradley 2001). Such heterogenous genetic origins of the horse prevent us from making any firm statement about thoroughbred origins *per se* – thoroughbreds might equally descend from a single diverse source population as they may have evolved from several populations, though historical records suggest the latter (Willett 1975).

The uncertainty of the genetic origins of the thoroughbred is a consequence of the domestication history of the horse. High diversity estimates, limited definitive haplotype clustering within populations and random distribution of diversity among horse populations is consistent with the capture and exploitation of genetically diverse wild

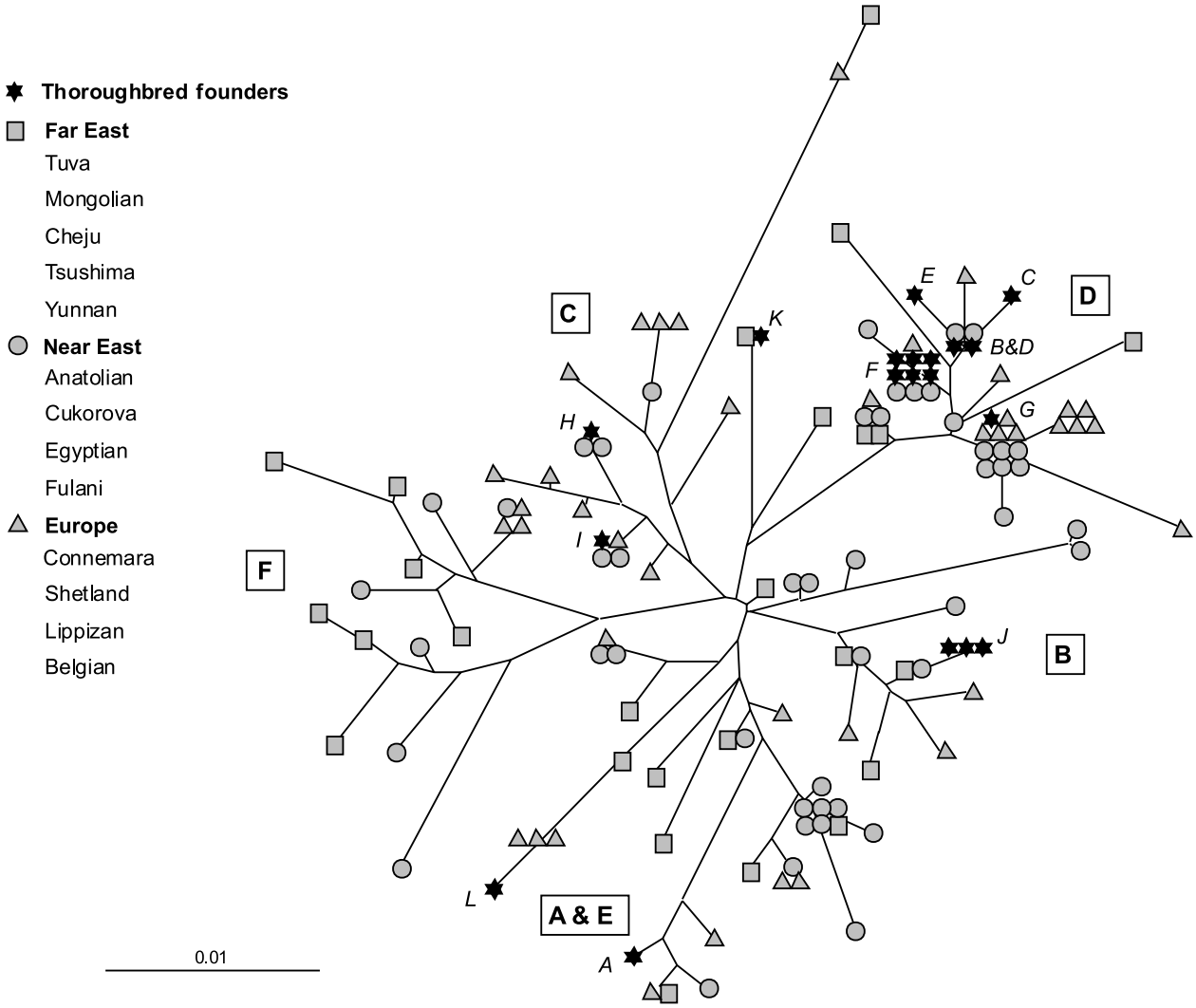

Figure 2 Neighbour-joining phylogenetic tree relating mitochondrial DNA (mtDNA) haplotypes in horse populations from three geographically widespread regions. A founder population for the thoroughbred was reconstructed from sequence and pedigree information and the haplotypes therein are represented in the tree as stars (Haplotypes A–L). Although considered separate founders in 381 bp *Families 10 and 14* (Haplotypes B and D) share an identical sequence in 343 bp and therefore share the same node in this tree. Haplotypes cluster in five distinct clades similar to those determined by Vila *et al*. (2001), although in this sequence, clades A and E are indistinguishable.

progenitors having taken place in multiple locations, possibly over a broad time span throughout the >6000 years association between humans and the horse (Brown & Anthony 1998; Vila *et al*. 2001). The high mobility of the horse, enabled by the nature of its domestic roles will also have led to an obscuring of the genetic structure within the species through post-domestic migration.

Haplotype sharing among thoroughbred founders is much higher than observed in other horse populations. The estimated PI in the thoroughbred founder population (0.15) suggests that at least 15 in 100 randomly sampled thoroughbreds (given that all founder lineages are represented equally) share an identical sequence. This is three times the observed PI in Arabian horses (0.05) (Bowling *et al*. 2000) and higher than in all other populations examined here except the Fulani (0.24) and Shetland (0.23), though these estimates may be biased by unintentional sampling of relatives. In the thoroughbred, because PI estimates are much higher in founders than expected we propose that some thoroughbred lineages may descend from common dams.

For example, the coupling of this genetic information with pedigree records strongly indicates that *Families 4* (Layton Barb Mare 1670), *11* (The Pet Mare 1697) and *13* (Sedbury Royal Mare 1665) may descend from a single

Table 3 Population pairwise Fst genetic distances among populations and mean number of pairwise differences within populations of geographically disparate horses.

	Cheju ($n = 7$)	Tuva ($n = 11$)	Anatolian ($n = 13$)	Cukorova ($n = 12$)	Egyptian ($n = 8$)	Fulani ($n = 12$)	Shetland ($n = 14$)	Lippizan ($n = 10$)	Connemara ($n = 12$)	TB founders ($n = 19$)
Cheju	7.238 (4.59)	–	–	–	–	+	–	–	–	+
Tuva	0.000	10.109 (14.84)	–	–	–	–	–	–	–	+
Anatolian	0.010	0.002	6.800 (8.29)	–	–	–	–	–	–	–
Cukorova	0.033	0.000	0.005	7.924 (11.67)	–	–	–	–	–	–
Egyptian	0.000	0.000	0.000	0.008	9.250 (10.86)	–	–	–	–	–
Fulani	0.145*	0.042	0.051	0.047	0.064	5.909 (14.39)	+	–	–	–
Shetland	0.075	0.021	0.065	0.000	0.073	0.100*	8.363 (14.72)	–	–	+
Lippizan	0.000	0.000	0.000	0.040	0.000	0.061	0.072	9.267 (11.47)	–	+
Connemara	0.104	0.035	0.000	0.000	0.036	0.022	0.037	0.052	7.288 (9.808)	–
TB founders	0.173*	0.100*	0.056	0.065	0.077	0.085	0.131	0.120*	0.002*	7.322 (19.49)

Above diagonal (and *): significant Fst *P*-values in 110 permutations at significance level of 0.05; diagonal elements: mean number of observed pairwise differences (and variance) within populations; below diagonal: population pairwise Fst genetic distances.

common founder. The founder haplotype in each is identical (Haplotype J), all three were owned by James Darcy, were kept at Sedbury Stud and lived at about the same time. Further, most historical literature, including GSB entries, entertains this notion, though none is conclusive (Lowe 1913; Wentworth 1938; Prior 1935). The GSB records that a daughter of the Layton Barb Mare 1670 (*Family 4*) produced two foals. One was unnamed but, bred by James Darcy's daughter, may have been The Pet Mare 1697 (*Family 11*). Also, some sources argue that The Pet Mare 1697 may have been a synonym for Grey Royal 1697, granddaughter of the Sedbury Royal Mare 1665 (*Family 13*).

It has also been persuasively argued that the *Family 7* founder, Lord Darcy's Blacklegged Royal Mare 1710, shared a common ancestress with these three families (Prior 1935), but our results give two different haplotypes (F and J, 11 bp different) and indicate this cannot be the case. However, haplotype matches suggest possible matrilineal relationship between *Family 7* and one or more of the founders of *Families 2, 8, 16, 17* and *22*. Some historical literature infers these five families descend from a single common founder, but no historical records suggest any link with the *Family 7* founder.

We propose that as few as 12 founders may have contributed to the major lineages within the 19 thoroughbred families included in this study. However, if we also consider the deep-rooted anomalies, which probably result from confusion at the foundation stages, then *Families* 5 and *6* both have a contribution from an additional founder (Haplotypes M and N). The deep-rooted *Family 9* anomaly (Hapolype G) may best be explained by confusion with the founder of *Family 12*. In fact, only one anomalous haplotype (Haplotype O, *Family 19*) in a relatively modern pedigree (19th century – 1980) is not accounted for by a match with another sampled family.

Although each family is expected to have only one founder and this founder is considered to contribute to one family only we have uncovered a web of founder sharing. Female founders, as they are currently understood, may have contributed differently to these 19 families than previously thought. Further, descendants of Maid of the Glen 1858 (*1*), Hag 1744 (5), the dam(s) of Betty Percival 1715 and Cream Cheeks 1695 (*6*), a Curwen Bay Barb Mare 1708/1709 (*9*), Young Camilla 1787 (*11*), Lady Alice 1855 (*16*) and Violet 1858 (*19*) (Fig. 1) may contain a genetic heritage different to that which pedigree information suggests.

The coupling of genetic and historical data provides a powerful tool to identify and correct errors that may be present in contemporary thoroughbred pedigrees. This is vital for thoroughbred breeders who rely on the accuracy of

stud books, as million-dollar decisions are frequently made based on the integrity of the pedigrees they record. Also, such parallel analyses lend new perspective to the interpretation of the early history of the thoroughbred and the contribution of the founder mares to the present day thoroughbred gene pool.

Acknowledgements

The authors wish to thank the Connemara Pony Breeders Society, John Dooley, Gainsborough Stud UK, Gainsborough Farm US, Woodpark Stud and Monksgrange Stud for access to samples. This work was financed in full by Gainsborough Stud Management Ltd.

References

Bowling A., Del Valle A. & Bowling M. (2000) A pedigree-based study of mitochondrial D-loop DNA sequence variation among Arabian horses. *Animal Genetics* **31,** 1–7.

Brown D.R. & Anthony D.W. (1998) Bit wear, horseback riding and the Botai site in Kazakhstan. *Journal of Archaeological Science* **25,** 331–47.

Cunningham E.P., Dooley J.J., Splan R.K. & Bradley D.G. (2001) Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Animal Genetics* **32,** 360–4.

Hutchinson C.A., Newbold J.E., Potter S.S. & Hall E.M. (1974) Maternal inheritance of mammalian mitochondrial DNA. *Nature* **251,** 536–8.

Kavar T., Habe F., Brem G. & Dovc P. (1999) Mitochondrial D-loop sequence variation among the 16 maternal lines of the Lipizzan horse breed. *Animal Genetics* **30,** 423–30.

Lowe C.B. (1913) *Breeding racehorses by the figure system*. (ed. by W. Allison). The Field and Queen (Horace Cox) Ltd, UK.

MacHugh D.E. & Bradley D.G. (2001) Livestock genetic origins: goats buck the trend. *Proceedings of the National Academy of Sciences of the USA* **98,** 5382–4.

Prior C.M. (1935) *The Royal Studs of the Sixteenth and Seventeenth Centuries*. Horse and Hound Publications Ltd, London.

Schneider S., Roessli D. & Excoffier L. (2000) *Arlequin: a Software for Population Genetics Data Analysis* (Version 2.0). Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F. & Higgins D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **24,** 4876–4882.

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,** 474–7.

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

Wentworth (1938) *Thoroughbred Racing Stock*. George Allen & Unwin Ltd, UK.

Willett P. (1975) *An Introduction to the Thoroughbred*. Stanley Paul Ltd, London.

Xu X. & Arnason U. (1994) The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene* **148,** 357–62.