

RESEARCH ARTICLE

Genetic analysis of the endangered Cleveland Bay horse: A century of breeding characterised by pedigree and microsatellite data

Andrew Dell^{1,2*}, Mark Curry¹, Kelly Yarnell³, Gareth Starbuck³, Philippe B. Wilson^{2,3*}

1 Department of Biological Sciences, University of Lincoln, Brayford Way, Brayford Pool, Lincoln, United Kingdom, **2** Rare Breeds Survival Trust, Stoneleigh Park, Stoneleigh, Warwickshire, United Kingdom, **3** School of Animal, Rural and Environmental Sciences, Nottingham Trent University, Brackenhurst Campus, Southwell, Nottinghamshire, United Kingdom

* andy.dell@btinternet.com (AD); PhilippeWilson@rbst.org.uk (PBW)



OPEN ACCESS

Citation: Dell A, Curry M, Yarnell K, Starbuck G, Wilson PB (2020) Genetic analysis of the endangered Cleveland Bay horse: A century of breeding characterised by pedigree and microsatellite data. PLoS ONE 15(10): e0240410. <https://doi.org/10.1371/journal.pone.0240410>

Editor: Chris Rogers, Massey University, NEW ZEALAND

Received: June 3, 2020

Accepted: September 27, 2020

Published: October 29, 2020

Copyright: © 2020 Dell et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant files are available from the Figshare repository (DOI: [10.6084/m9.figshare.12981614](https://doi.org/10.6084/m9.figshare.12981614)).

Funding: AD was funded by the Rare Breeds Survival Trust, and a PhD Studentship from the University of Lincoln.

Competing interests: No.

Abstract

The Cleveland Bay horse is one of the oldest equines in the United Kingdom, with pedigree data going back almost 300 years. The studbook is essentially closed and because of this, there are concerns about loss of genetic variation across generations. The breed is one of five equine breeds listed as “critical” (<300 registered adult breeding females) by the UK Rare Breeds Survival Trust in their annual Watchlist. Due to their critically endangered status, the current breadth of their genetic diversity is of concern, and assessment of this can lead to improved breed management strategies. Herein, both genealogical and molecular methods are combined in order to assess founder representation, lineage, and allelic diversity. Data from 15 microsatellite loci from a reference population of 402 individuals determined a loss of 91% and 48% of stallion and dam lines, respectively. Only 3 ancestors determine 50% of the genome in the living population, with 70% of maternal lineage being derived from 3 founder females, and all paternal lineages traced back to a single founder stallion. Methods and theory are described in detail in order to demonstrate the scope of this analysis for wider conservation strategies. We quantitatively demonstrate the critical nature of the genetic resources within the breed and offer a perspective on implementing this data in considered breed management strategies.

Introduction

In recent years there has been substantial interest in quantifying the genetic diversity of equine breeds using pedigree [1], molecular data [2] or a combination of both sources [3] in order to implement effective breed management strategies. The effectiveness of the use of both data types in the understanding and management of rare and native equine breeds have been investigated using both theoretical modelling, and studies of closed studbooks.

The Cleveland Bay horse is a heritage British breed which has its origins in the Cleveland Hills of Northern England [4]. The first studbook was published in 1885, and this contains

retrospective pedigrees of animals dating back to 1732 providing a closed non-Thoroughbred studbook dating back almost 300 years and for more than 38 generations. In addition, the breed Society now has a mandatory policy of microsatellite-based parentage testing at the time of registration. Unrestricted access to the microsatellite test data, as well as the stud book records provides a rare opportunity to evaluate both methods of assessing genetic diversity within the breed and, in turn, provides comprehensive guidance to breeders in terms of conservation practice for this endangered breed [5], whilst providing an important and potentially wide-ranging tool for wider conservation practices both *in situ* and *ex situ in vivo*.

The Cleveland Bay is a warm-blooded equine; a product of a cross of hot-blooded Oriental / Barb / Turkish or Mediterranean stock on the cold-blooded Northern European heavy draught horse [6]. It is reputed to have evolved in the matriline from the now extinct Chapman horse, which early records show were being bred on the monastic estates of the region well before the dissolution of the monasteries in the mid 16th century [4].

Although stated as being “free of blood” in the first three volumes of the studbook [7], early research into the founders of the breed recognised the contribution on the male side by some notable Thoroughbred stallions that were standing at stud or travelling in the region in the late 18th and early 19th Centuries [8].

Over the years the breed has been used extensively as both a work horse and a riding horse, and has been crossed with other breeds to produce carriage horses [9]. Indeed, at one time there was a separate breed society with its own studbook—The Yorkshire Coach Horse Society—for such animals [10]. Such has been the desirability of the pure Cleveland Bay for contributing weight carrying capacity when crossed with other equine breeds, that they have been exported globally [9]. In addition to North America, the breed has been exported to Australasia, Pakistan and Japan; a Cleveland Bay stallion stands at the Imperial stud [8].

The fashion for such effective cross-bred horses is one factor that brought the pure-bred Cleveland Bay horse to the edge of extinction. The substantial decrease in population size of the breed following the First World War when large numbers of Cleveland Bay horses were used to haul artillery on the battlefields of Northern Europe led to sustainability concerns regarding the remaining genetic resources of the breed [9]. The popularity of the breed continued to decline in the 1920s and 30s as the increasing use of motorised transport reduced the need for carriage horses. Moreover, following the technological developments of the Second World War, further mechanisation was implemented in farming practice and the purpose of the Cleveland Bay was further diminished [11].

In an attempt to improve the diversity of the home-based breeding population, the stallion Farnley Exchange was brought back from the United States of America (USA) in 1945 to stand at stud [9]. By the early 1960s there were only four stallions of breeding age left in existence and the breed is known to have gone through a genetic bottleneck at this time [8].

In the 1960s HM the Queen purchased the stallion Mulgrave Supreme, thus preventing his export, and stood him at public stud, both to promote, and help conserve the genetic diversity of the breed in the United Kingdom. Since that time the breed has seen a moderate recovery in numbers, partly because of patronage of the breed society by HM the Queen and the use of Cleveland Bay horses at the Royal Mews.

By the late 1990s, between 35 and 50 pure bred animals were being registered annually by the Cleveland Bay Horse Society (CBHS), whose studbook now includes animals being bred both in the United Kingdom, Europe, North America and Australasia [12].

The breed is one of only five equines listed as “Critical” by the UK Rare Breeds Survival Trust, indicating that the population has less than 300 breeding females. Earlier investigation of the CBHS Studbook records [7] indicated there were eight female ancestry lines existing within the breed.

A more recent study [13] restricted to animals entered in the CBHS studbook between 1934 and 1995, highlighted the limited genetic diversity in the breed and the increasing levels of inbreeding. It was recognised that further in-depth analysis of the status of the breed would be needed in order to aid in the development of breed management plans.

The aim of this study was to develop a comparative analysis of the genetic diversity in the Cleveland Bay Horse population using both genealogical and molecular methods and provide recommendations in order to support a global breed conservation strategy for the Cleveland Bay Horse, whilst sequentially detailing the theory and practice inherent in our approach leading to its applicability in the conservation of endangered breeds and species *in vivo*.

Materials and methods

Pedigree data

Summary data from the CBHS stud books volumes one to thirty eight was published in the Society's Centenary studbook [7]. Names and studbook numbers of all registered horses together with date of birth, sire and dam were listed and this information was digitised in Filemaker™ (Filemaker Inc.), to construct an electronic pedigree database for the breed, stored in Filemaker format. Registrations post-1985 have been added to the database on an annual basis up to and including for this study, Volume 38 of the studbook.

The Cleveland Bay Horse Society provided access to a total of 535 microsatellite parentage testing reports. These had been obtained by commercial analysis of hair follicle samples taken from individual animals for registration verification. Samples were tested for a panel of 16 microsatellite markers approved by the International Society for Animal Genetics (ISAG) equine genetics group, by the Animal Health Trust (Newmarket, UK.). Close examination of stud book records, recent Breed Society census records and the microsatellite dataset enabled the identification of a reference population of 402 animals, registered in the 10-year period 1997 to 2006 for which both microsatellite and pedigree data was available.

Pedigree completeness

Data correction routines within the programmes Genes [14] and Eva [15] were used to identify pedigree errors and correct infinite loops. Calculation of Pedigree Completeness was made using PopRep [16]. Using Eqs 1 and 2 to compute pedigree completeness index [17] (I_d):

$$I_d = \frac{4I_{d_{pat}}I_{d_{mat}}}{I_{d_{pat}} + I_{d_{mat}}} \quad \text{Eq 1}$$

$$I_{d_k} = \frac{1}{d} \sum_{i=1}^d a_i \quad k = pat, mat \quad \text{Eq 2}$$

Where k represents the paternal (pat) or maternal (mat) line of an individual, and a_i is the proportion of known ancestors in generation i ; d is the number of generations measured when calculating the pedigree completeness. Values for pedigree completeness will range from 0 to 1. Where all of the ancestors of an individual are known to some specified generation (d) then $I_d = 1$. However, where one of the parent animals is unknown, $I_d = 0$ [16].

Generation interval

Generation Interval is defined as the average age of the parent animals at the birth of selected offspring with offspring subsequently producing at least one progeny [18]. The generation interval

was calculated for each of the four possible lines of descent: sire to son; sire to daughter; dam to son and dam to daughter. The results were averaged for each year group using PopRep [16].

Founder and ancestor representation

Stallion and dam lines, defined respectively as: *unbroken descent through male or female animals only from an ancestor to a descendant* [3] were identified and detailed founder and ancestor analysis was performed using Endog 4.6 [19] to initially determine Number of Founders.

We make the assumption that all animals with two unknown parents are regarded as founders in this analysis [20]. In addition, if an animal has one known and one unknown parent, the unknown parent is regarded as a founder. The total number of founders contains limited information on the genetic basis for the population. Firstly, founders are assumed to be unrelated, as their parentage is unknown. However, this is most likely not the case in practice. Secondly, some founders have been used more intensely and therefore contribute more, in terms of genetic resource, to the current population than other founders.

The effective number of founders, f_e , has been designed to correct for this second shortcoming [21] and is defined as the number of equally contributing founders that would be expected to produce the same genetic diversity as in the population under study. This is computed as:

$$f_e = \left[\sum_{i=1}^{N_f} q_i^2 \right]^{-1} \quad \text{Eq 3}$$

Where q_k is the probability of gene origin of the k^{th} founder and N_f the real number of founders. In a scenario where every founder makes an equal contribution, the effective number of founders will equal the actual number of founders.

It is more common for founders to contribute unequally, leading to $f_e < N_f$. The genetic contributions will converge following 5 to 7 generations [22]. Once this convergence occurs, employing f_e as a measure of genetic contribution, will have limited usefulness as will remain constant irrespective of later changes in the population. Pedigrees of more than 7 generations can be characterized with a high effective number of founders even after a severe, recent bottleneck [23]. Whilst the effective number of founders is not an absolute measure of genetic diversity, it forms a basis for comparison of the effective population size (N_e) and the effective number of ancestors (f_a). In a population with minimum inbreeding, f_e would be expected to be approximately equal to $\frac{1}{2}N_e$ [22]. Where f_e diverges from this, there is compelling evidence that the breeding structure has been changed since the founder generation [24].

The Effective Number of Founder Genomes (f_g) was proposed by Lacy (1989) to account for unequal founder contributions, random loss of alleles caused by genetic drift and for bottleneck events. It is computed by the equation:

$$f_g = 1 / \sum_{i=1}^c (p_i^2 / r_i) \quad \text{Eq 4}$$

Where p_i is the expected proportional genetic contribution of a founder i ; r_i is the expected proportion of alleles from founder i which remain in the current population, and c is the total number of contributing founders [21]. This gives an indication of the number of equally contributing founders with no loss of founder alleles, that would produce the same degree of diversity as found in a reference population [25]. The f_g will be smaller than both f_e and the effective number of ancestors (f_a), even under minimum inbreeding pressure, and approximately equal to $\frac{1}{2}N_e$. The scale of these differences is indicative of the degree of random loss of alleles. Alleles will be lost with every generation of a pedigree and thus f_g will decrease as the depth of pedigree increases [24].

The Effective Number of Ancestors (f_a) supplements f_e and is calculated from the genetic contributions of ancestors with the largest marginal genetic contributions themselves [20]. Whilst genetic contributions of founders are independent and sum to unity, this is not the case for genetic contributions of ancestors. Indeed, the dam of a highly used sire has >50% contribution of her son, as the same genes are represented in both generations. Boichard et al. (1997) therefore introduced the marginal contribution to the pedigree genetic resource. The ancestors contributing most to the reference population are considered individually in a recursive process. For each round of the recursion, the ancestor with the highest contribution is chosen, and the contributions of all others are calculated conditionally on the contribution of the chosen ancestor. The marginal contribution is the genetic contribution from an individual after correcting for contributions of other ancestors already considered in the recursive process. The sum of marginal contributions of all ancestors will be equal to unity. Ancestors with a large marginal contribution to the reference population will correlate with individuals having genes passed through many descendants [24].

Assessment of the f_a helps to account for the losses of genetic variability produced by the unbalanced use of individuals in terms of reproduction within breeding programmes. This is conventional in domestic equines, whilst also accounting for bottlenecks in the pedigree.

The parameter f_a is computed as

$$f_a = 1 / \sum_{j=1}^a q_j^2 \quad \text{Eq 5}$$

where q_j is the marginal contribution of an ancestor j .

Inbreeding analysis

Inbreeding coefficients for each individual animal were calculated using ENDOG [19].

The Increase in Inbreeding (ΔF), is calculated for each generation using ENDOG 4.6 [19], by means of Eq 6.

$$\Delta F = \frac{F_t - F_{t-1}}{1 - F_{t-1}} \quad \text{Eq 6}$$

where F_t and F_{t-1} are the average inbreeding of offspring and their parents, respectively [18].

The Average Relatedness Coefficient (AR) [26] describes the probability that a randomly chosen allele from the whole population in the pedigree belongs to the animal under study. This parameter was calculated using ENDOG 4.6 [19]. The Additive Relationship Coefficient (R_{yz}), is estimated for two animals through calculating the hypothetical coefficient of inbreeding of an animal produced by mating the two individuals, irrespective of the sex of these assumed parents. The additive relationship between the two animals is then calculated as twice the coefficient of inbreeding of the hypothetical offspring. $R_{yz} = 2 F_x$, where F_x is the coefficient of inbreeding of the hypothetical offspring of individual Y and individual Z. This additive relationship has a minimum value of zero and a maximum value of two. The Additive Relationship is twice the value of the coefficient of kinship. The kinship of any two individuals is identical to the inbreeding coefficient of their progeny if they were mated. It is the probability that alleles drawn randomly from gametes of each of the two individuals are identical by descent.

Effective population size

The Effective Population Size from the rate of inbreeding is computed using the classic equation

$$N_e = \frac{1}{2\Delta F} \quad \text{Eq 7}$$

Where the rate of inbreeding per generation is calculated using Eq 6. The Effective Population Size from the number of parents is computed as

$$N_e = \frac{4N_m N_f}{N_m + N_f} \quad \text{Eq 8}$$

Where N_m and N_f are the number of male and female parents, respectively [18]. This method assumes that the ratio of breeding males to breeding females is 1:1, and that all individuals have an equal opportunity to contribute their genetic material to the next generation. This is seldom the case in managed livestock populations and there is a tendency for this method to overestimate N_e [16].

Microsatellites

Total DNA was isolated at the Animal Health Trust's laboratories, from hair follicle samples following standard commercial procedures and as previously described [27]. A set of 16 microsatellites (ASB17 VHL20 HTG10 HTG4 AHT5 AHT4 HMS3 HMS6 HMS7 ASB23 LEX3 LEX33 ASB2 HTG6 HTG7 HMS2) were analysed in all the sampled individuals. The GENETIX program was used to carry out factorial correspondence analyses and associated calculations on 15 of these markers [28]. Although microsatellite LEX3 appears in the panel of markers recommended for equine parentage verification by the International Society for Animal Genetics it was excluded from the analysis in this study because it is located on the X chromosome and as such is not appropriate for this type of analysis.

The Average Number of Alleles per Locus (A), corrected in order to account for sample size using Hurlbert's rarefaction method (1971) can be shown as:

$$A[g] = \sum_i \left[1 - \prod_{k=0}^{g-1} \frac{N-N_i-k}{N-k} \right] \quad \text{Eq 9}$$

where g is the specified sampled size for a collection containing N individuals, numbering N_i in the i^{th} species.

Nei's minimum distance (D_m) and Nei's standard distance (D_s [29]) are computed according to Eqs 10 and 11, respectively.

$$D_{mm} = \frac{f_{mm} + f_{kk}}{2} \quad \text{Eq 10}$$

$$D_s = -\ln \frac{f_{km}}{f_{kk} \times f_{mm}}^{1/2} \quad \text{Eq 11}$$

where f_{kk} and f_{mm} are the average coancestry between individuals belonging to population k or m , and f_{km} is the average coancestry between individuals belonging to populations k and m .

Population structure

F (fixation) statistics extend the study of inbreeding coefficients in the case of sub-divided populations [30]. The F_{IT} refers to the inbreeding of individuals in the total population. Conversely, F_{IS} describes the inbreeding of individuals within sub-populations. F_{ST} is not strictly a fixation index as it represents the correlation between two gametes taken at random in two sub-populations from the total population. It measures the degree of genetic differentiation of

the sub-populations. The three indices are computed as in Eqs 12, 13 and 14, respectively

$$F_{IS} = \frac{\bar{F} - \bar{f}}{1 - \bar{f}} \quad \text{Eq 12}$$

$$F_{ST} = \frac{\bar{f} - \bar{f}}{1 - \bar{f}} = \frac{\bar{D}}{1 - \bar{f}}, \quad \text{Eq 13}$$

and

$$F_{IT} = \frac{\bar{F} - \bar{f}}{1 - \bar{f}} \quad \text{Eq 14}$$

where \bar{f} and \bar{F} are, respectively, the mean coancestry and the inbreeding coefficient for the entire metapopulation, and, the average coancestry for the subpopulation, so that $(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$ [31].

ENDOG [19] was used to calculate F statistics and Nei's minimum distance [29]), D , the genetic distance between subpopulations i and j which is given by Eq 15

$$D_{ij} = D_{ij} - [(D_{ii} + D_{jj})/2] = [(f_{ii} + f_{jj})/2] - f_{ij} \quad \text{Eq 15}$$

The programme TREX [32, 33] was used to construct phylogenetic trees to illustrate the structure from the distance matrix data.

Bayesian model-based clustering was conducted using the programme STRUCTURE v2.1 [34], to assign individuals to homogeneous clusters or populations K , from a user defined range. An admixture model was adopted, with a burn in of 104 and 104 iterations of each value of K from 2 to 25.

Results

Pedigree completeness

The pedigree file included a total of 5422 animals, of which 2661 were male and 2761 were female. The reference population of 402 individual animals consisted of 193 male and 209 females for which microsatellite data as well as pedigree data was available.

The pedigree file was analysed to assess the number of fully traced generations for each individual, the maximum number of generations traced and the equivalent complete generations for each animal. The maximum number of traced generations was 36. Percentage average population completeness for each year of birth considering 1 through 6 generations are shown in Fig 1 with percentage population completeness for the reference population up to 6 generations being high (Table 1).

Average generation interval

Generation intervals for each of the four pathways (Table 2) ranged from 9.2 years to 10.0 years (sire-son and sire-daughter, respectively). The average generation interval for each breeding year (Fig 2) was found to range between 5.5 and 13 years, being at a minimum in the immediate post WW2 period 1946 to 1950, which coincides with the genetic bottleneck previously identified by Walling (1994).

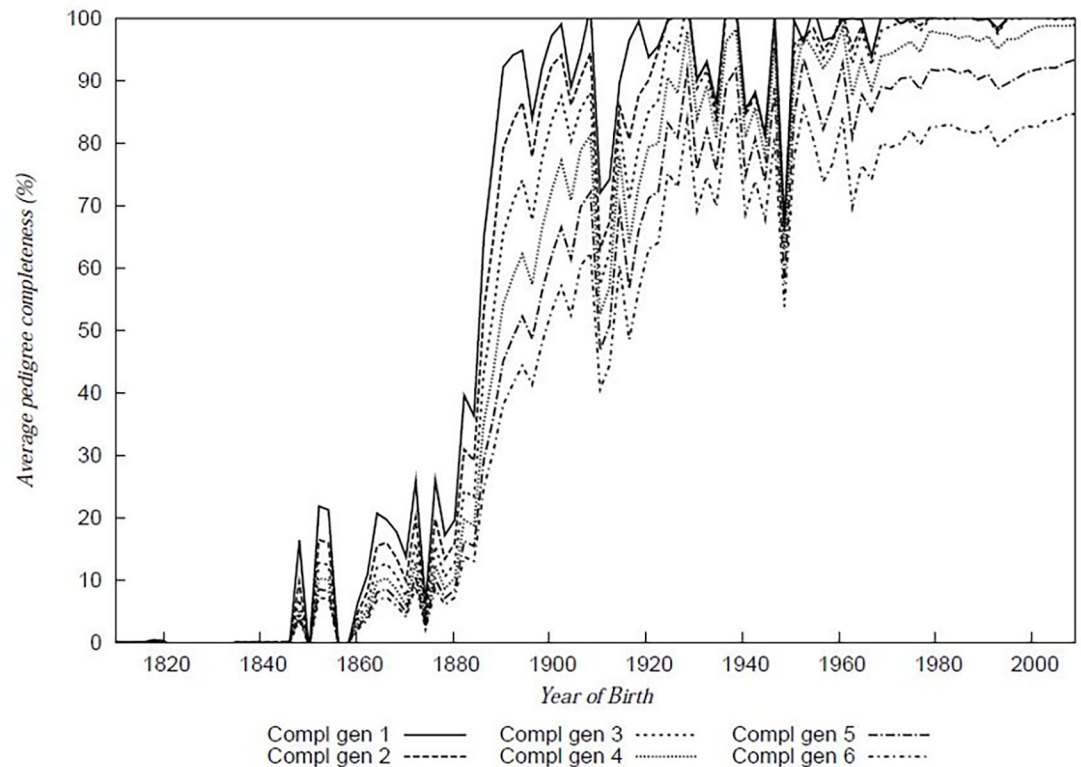


Fig 1. Percentage pedigree completeness over 6 generations. Average percentage completeness (%) is shown as a factor of individual birth year.

<https://doi.org/10.1371/journal.pone.0240410.g001>

Founder and ancestor representation

A total of 11 stallion lines were identified in the pedigree. A single paternal ancestry line is present in the reference (living) population.

Analysis of the female members of the studbook identified a total of 17 dam lines. Nine of these maternal ancestry lines are present in direct descent in the living population. Three of these lines (2,4 & 9) are only represented, in direct female descent, by either a single individual or two individual animals (Table 3). The three most common maternal lines constitute 70% of the present female population. However, analysis of the relative contributions of the most influential maternal ancestry lines to the genome of the reference population reveals that some of the lines least well represented in direct descent in fact continue to make a substantial genetic contribution as shown in Table 3.

Table 1. Pedigree completeness over 6 generations estimated from breed society records and pedigree recording data.

Generations	Completeness (%)
1	100
2	100
3	99.9
4	98.6
5	92.6
6	83.7

<https://doi.org/10.1371/journal.pone.0240410.t001>

Table 2. Average generation interval by pathway.

Pathway	Average generation interval (years)
Sire son	9.2
Sire daughter	10.0
Dam son	9.6
Dam daughter	9.3
Whole pop	9.6

<https://doi.org/10.1371/journal.pone.0240410.t002>

Analysis identified 194 founders in total of which 28 were represented in the reference population. The mean retention was 0.035. The number of founder genomes surviving was 6.285. Calculations on the same population show the founder genome equivalent to be 2.366 with the effective number of non-founders only 2.379. The proportion of ancestry known was 0.330 reflecting the fact that in early volumes of the studbook only a record of the sire of an individual animal was made. The Number of Ancestors contributing to the population was 424 and the number of ancestors describing 50% of the genome was 7 animals.

The number of Ancestors contributing to the Reference Population was calculated as 31 animals. The Effective Number of Founders/Ancestors [20] for the Reference Population were 40 and 9, respectively. The number of ancestors describing 50% of the genome of the living population was 3. Ancestors were selected following Boichard et al. (1997), while founders were selected by their individual Average Relatedness coefficient (AR).

Inbreeding analysis and effective population size

Across the whole analysed dataset, $F = 7.8\%$ with an associated mean average relatedness of 8.3%. Fig 3 shows Inbreeding and additive relationship coefficients by birth year between 1900 to 2006.

The average rate of change of the additive genetic relationships between 1901 and 2009 for the Cleveland Bay Horse breed was 0.00202 per year based on the slope regression. This results in a Δf per generation of 0.02629 (Table 4). The rate of change of the average inbreeding coefficients based on slope regression between 1901 and 2009 was 0.00214, which represents a ΔF per generation of 0.02709. The effective population sizes for the Cleveland Bay Horse breed, based on Δf and ΔF were 19 and 18, respectively (Fig 4). The pattern of inbreeding during which the reference population was foaled and Effective population size, calculated based on both the rate of inbreeding and the number of parents are tabulated in Table 5 for the period 1997 to 2006 with data calculated using POPREP [16].

Microsatellite variation

The total number of alleles found for 15 microsatellite loci within the reference population was 93. The mean number of alleles per locus was 2 ranging from 4 to 10. The mean Observed Heterozygosity (H_o) ranged between 0.052 (HTG7) and 0.716 (VHL20) the mean being 0.4486 whilst the mean Expected Heterozygosity (H_e) was 0.5341. The highest values for H_e were found for microsatellite LEX33 whilst the lowest were found for microsatellite HTG6 (Table 6).

Across the reference population there is complete heterozygosity. However, at subpopulation level 3 (Table 7), groups show homozygosity at multiple loci. Female Line 2 is 62.5% polymorphic with fixation at HMS3 and LEX3. Female Line 4 is 62.5% polymorphic with fixation of alleles at HMS3, ASB23, HTG4, HTG10 and LEX3. Female Line 8 is 93.75% polymorphic with fixation at LEX3.

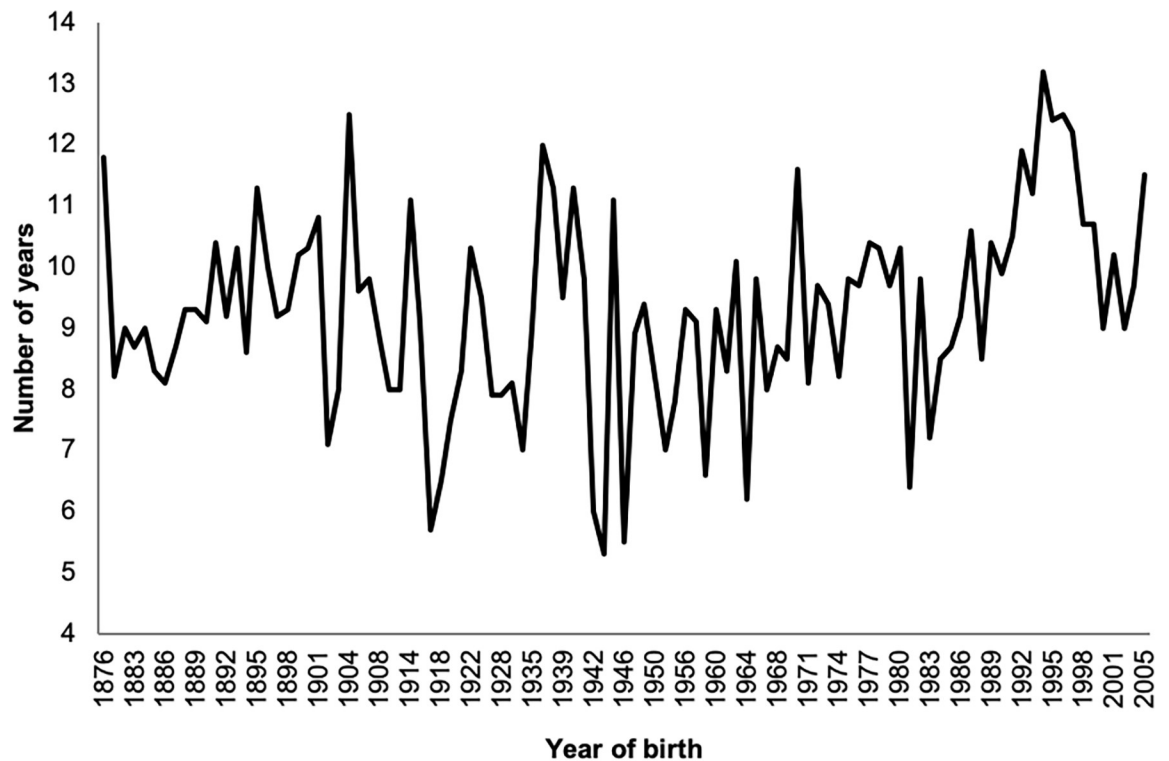


Fig 2. Average generation interval for whole population calculated as the average age of parents at the birth of offspring which in turn produce the next generation of breeding individuals.

<https://doi.org/10.1371/journal.pone.0240410.g002>

Allele frequencies are more restricted in populations 2, 4 and 9 (Fig 5), as is the expected heterozygosity. This will be influenced by the smaller membership and corresponding sample size for these subpopulations.

The analysis of allele frequencies identifies a significant number of gaps in the distribution of allele length or number of repeats. It has been reported that populations that have experienced genetic bottlenecks tend to exhibit such less cohesive distributions than stable populations [35].

Table 3. Relative contributions of maternal ancestry lines to the evolution of the whole and reference (1997–2006) populations.

Maternal Line	Whole Population		Evolution Rate Of Whole Population (%)	Reference Population		Evolution Rate Of Reference Population (%)
	N	%		N	%	
0	3103	57.23	-	0	0	-
1	424	7.82	11.38	77	19.15	23.94
2	231	4.26	11.59	3	0.75	27.45
3	269	4.96	0.52	56	13.93	2.14
4	26	0.48	0.08	1	0.25	1.29
5	500	9.22	0.09	82	20.40	15.89
6	611	11.27	-	115	28.06	21.54
7	177	3.26	-	6162	15.42	-
8	78	1.43	0.18	5	1.24	0.27
9	3	0.06	0.01	1	0.25	-
Total	5422	100.00		402	100.00	

<https://doi.org/10.1371/journal.pone.0240410.t003>

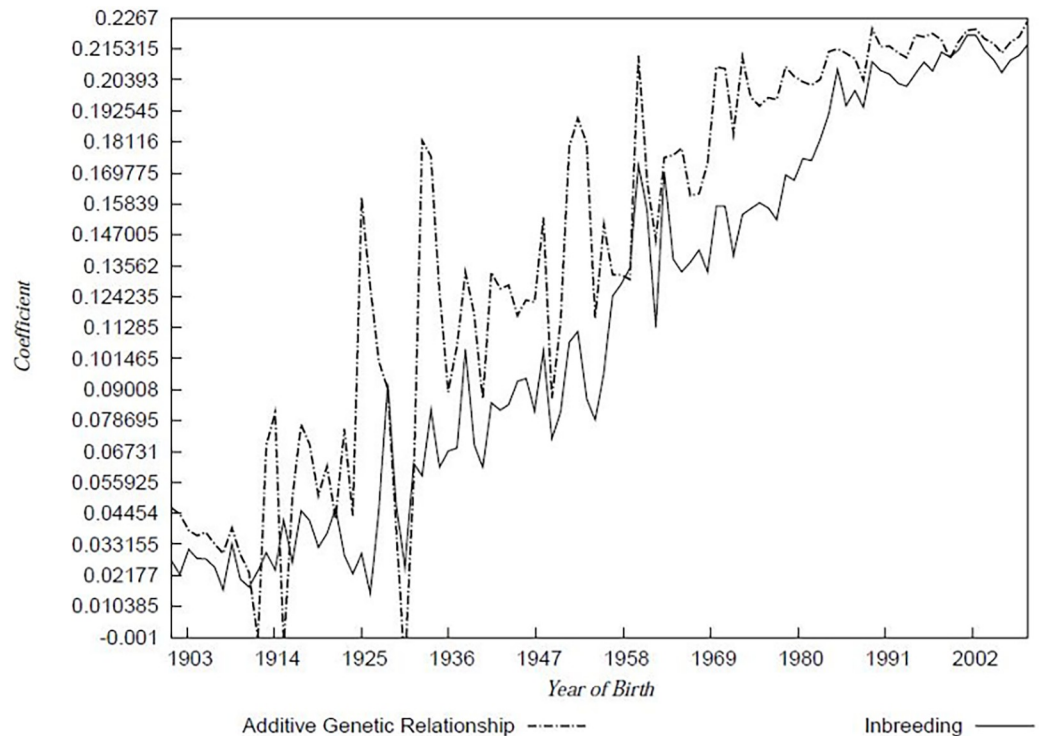


Fig 3. Inbreeding coefficient and additive genetic relationship 1900 to 2006 as a function of birth year of individuals.

<https://doi.org/10.1371/journal.pone.0240410.g003>

Bottleneck analysis

The microsatellite allele frequency data was tested for departure from mutation-drift equilibrium with the software BOTTLENECK 1.2 [23]. The results of the three tests of heterozygosity excess (Infinite Allele Model, IAM; Stepwise mutation Model, SMM; and Two-Phase Mutation Model, TPM) are shown in Table 8 and the results of the test for null hypothesis under Sign Test, Standard Difference Test and Wilcoxon Test in Table 9.

Under the Sign Test, the expected number of loci with heterozygosity excess were 8.93 ($p = 0.00120$) under IAM, 9.40 ($p = .029262$) under TPM, and 9.43 ($p = 0.06923$) under SMM. This suggests that the null hypothesis is rejected under IAM, but with $p > 0.05$ would appear to

Table 4. Change in inbreeding coefficient and average relatedness for 7 fully traced generations.

Complete Generations	N	F	AR
0	1562	0	0.01508601
1	1127	0.0109910	0.04204081
2	532	0.0409840	0.06222444
3	202	0.0733043	0.1084541
4	304	0.1299420	0.1471427
5	605	0.1761405	0.1711177
6	694	0.1931363	0.1795912
7	342	0.2047071	0.1813364
8	54	0.2167604	0.1838901

<https://doi.org/10.1371/journal.pone.0240410.t004>

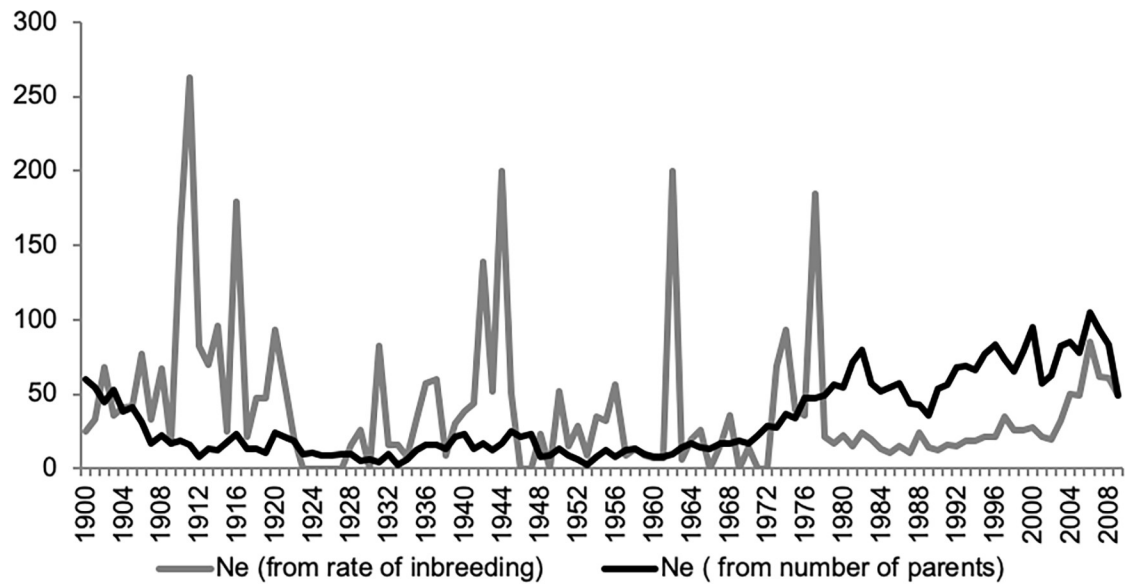


Fig 4. Effective population size from rate of change of inbreeding (grey series), and number of parents (black series) calculated with POPREP [16].

<https://doi.org/10.1371/journal.pone.0240410.g004>

be met under the other two tests. Therefore, only under the IAM is there clear evidence of a recent bottleneck event.

The standard difference test gives T2 probability statistics of 3.186 ($p = 0.00072$) under IAM; 0.902 ($p = 0.18357$) under TPM and -4.294 ($p = 0.00001$) under SMM. Probability values of less than 0.05 for both IAM and SMM under these two models suggest a recent bottleneck event.

Under the Wilcoxon rank test the probability values were 0.00042 (IAM); 0.11560 (TPM) and 0.97116 (SMM), thus rejecting the null hypothesis under IAM.

Mode shift indicator

The Bottleneck software [23] provides an alternative method for detecting potential genetic bottleneck events in the Mode Shift Indicator. Populations that have not experienced a bottleneck will be at or near mutation drift equilibrium and will be expected to have a large proportion of alleles with low frequency [36]. This pattern will show as a normal, L shaped

Table 5. Inbreeding coefficients, F , and effective population size (N_e) of animals by birth year 1997–2006.

Year	Number of Animals	Average F	Minimum Inbreeding	Maximum Inbreeding	ΔF	N_e (from rate of inbreeding)	Parents	N_e (from number of parents)
1997	57	0.2072	0.1327	0.2943	0.0141	35	84	73
1998	46	0.2139	0.1540	0.2943	0.0191	26	71	65
1999	54	0.2126	0.1448	0.3156	0.0193	26	85	79
2000	64	0.2139	0.1654	0.3079	0.0179	28	101	95
2001	37	0.2186	0.1783	0.3132	0.0236	21	60	57
2002	46	0.2218	0.1830	0.3084	0.0255	20	69	63
2003	52	0.2173	0.1830	0.3017	0.0156	32	85	82
2004	63	0.2133	0.1629	0.2852	0.0101	50	95	85
2005	54	0.2102	0.1100	0.2580	0.0103	49	84	78
2006	76	0.2065	0.0925	0.2616	0.0059	85	116	105

<https://doi.org/10.1371/journal.pone.0240410.t005>

Table 6. Summary statistics for the 15 microsatellite loci analysed. N_a represents the number of alleles; N , the sample size; H_o the Observed Heterozygosity; H_e the Expected Heterozygosity; HW the departure from Hardy-Weinberg equilibrium; F the Fixation Index; and N_m the Gene flow estimated from $F_{ST} = 0.25(1 - F_{ST})/F_{ST}$.

Locus	Chromosome Location	Size Range (nucleotides)	N_a	N	H_o	H_e	HW	F	F_{IS}	F_{IT}	F_{ST}	N_m
VHL20	30	83–102	5	402	0.716	0.697	NS	-0.015	-0.29	-0.144	0.113	1.968
HTG4	9	116–137	4	402	0.463	0.434	NS	-0.0356	-0.105	0.056	0.146	1.463
AHT4	24	140–166	7	402	0.53	0.529	NS	-0.0009	-0.268	-0.114	0.121	1.81
HMS7	1	167–186	5	402	0.697	0.706	ND	0.0048	-0.293	-0.15	0.11	2.017
HTG6	15	74–103	5	402	0.067	0.173	NS	0.5363	-0.24	0.392	0.51	0.24
AHT5	8	126–147	7	402	0.669	0.684	NS	0.0098	-0.345	-0.165	0.134	1.622
HMS6	4	154–170	5	402	0.59	0.572	NS	-0.0211	-0.271	0.041	0.245	0.769
ASB2	15	237–268	8	402	0.55	0.578	NS	0.0163	-0.219	-0.012	0.169	1.227
HTG10	21	83–105	4	402	0.687	0.675	NS	-0.0089	-0.07	0.116	0.174	1.189
HTG7	4	114–126	4	402	0.052	0.18	ND	0.606	-0.252	0.516	0.614	0.157
HMS3	9	146–170	7	402	0.187	0.203	ND	0.042	-0.197	0.041	0.199	1.007
HMS2	10	215–236	6	402	0.057	0.176	ND	0.5802	-0.243	0.452	0.559	0.197
ASB17	2	104–116	10	402	0.5	0.78	ND	0.2164	-0.02	0.316	0.329	0.509
ASB23	3	176–212	6	402	0.639	0.759	***	0.0777	-0.082	0.086	0.155	1.362
LEX33	4	203–217	10	402	0.575	0.805	ND	0.162	-0.196	0.03	0.189	1.073
MEAN			6.2	402	0.46527	0.53007		0.14467	-0.2061	0.0974	0.25113	1.10733

Significant deviations from HWE were observed for microsatellites AHT4, HTG1, and LEX33.

<https://doi.org/10.1371/journal.pone.0240410.t006>

distribution when displayed graphically. Fig 6 shows that the Cleveland Bay data displays a normal L-shaped distribution at low allele size class, but deviates from it in the latter quartiles. This would suggest a population not completely at mutation drift equilibrium, and showing evidence of having experienced a genetic bottleneck in the recent past.

As both the data plot and the trend show that at the higher size classes there is some departure from the normal L-shaped distribution; the absolute assumption of accepting the null hypothesis should be treated with caution. Indeed, on initial examination, the results of the analysis with Bottleneck [23] appear far from conclusive. Initial assessment suggests that under the IAM all of the tests provide evidence of a recent bottleneck event. However, under TPM and SMM, the evidence is somewhat contradictory indicates some reservation to assessment of the suggested recent bottleneck. The mutation drift model deviation from normal L-shaped distribution supports the above assumption, however, this conflicting evidence

Table 7. Summary of the microsatellite analysis results on a subpopulation by matriline basis and for the full dataset, where MNA represents the mean number of alleles per locus.

SUBPOPULATION	N	Observed Heterozygosity	Expected Heterozygosity	Unbiased Expected Heterozygosity	MNA	Polymorphic Loci
1	77	0.5590	0.5713	0.5769	4.9375	100.00%
2	3	0.4573	0.3196	0.3937	1.8125	62.50%
3	56	0.5893	0.6078	0.5978	4.5625	100.00%
4	1	0.5833	0.4208	0.5222	1.9333	62.50%
5	82	0.5919	0.6217	0.6009	4.5	100.00%
6	115	0.5659	0.6092	0.5749	4.875	100.00%
7	61	0.5818	0.6138	0.5857	4.6875	100.00%
8	4	0.6258	0.5483	0.5626	3.0625	93.75%
GRADING REGISTER	3	0.4722	0.3251	0.4021	1.75	62.50%
MEAN		0.5585	0.5153	0.5352	3.568978	86.81%

<https://doi.org/10.1371/journal.pone.0240410.t007>

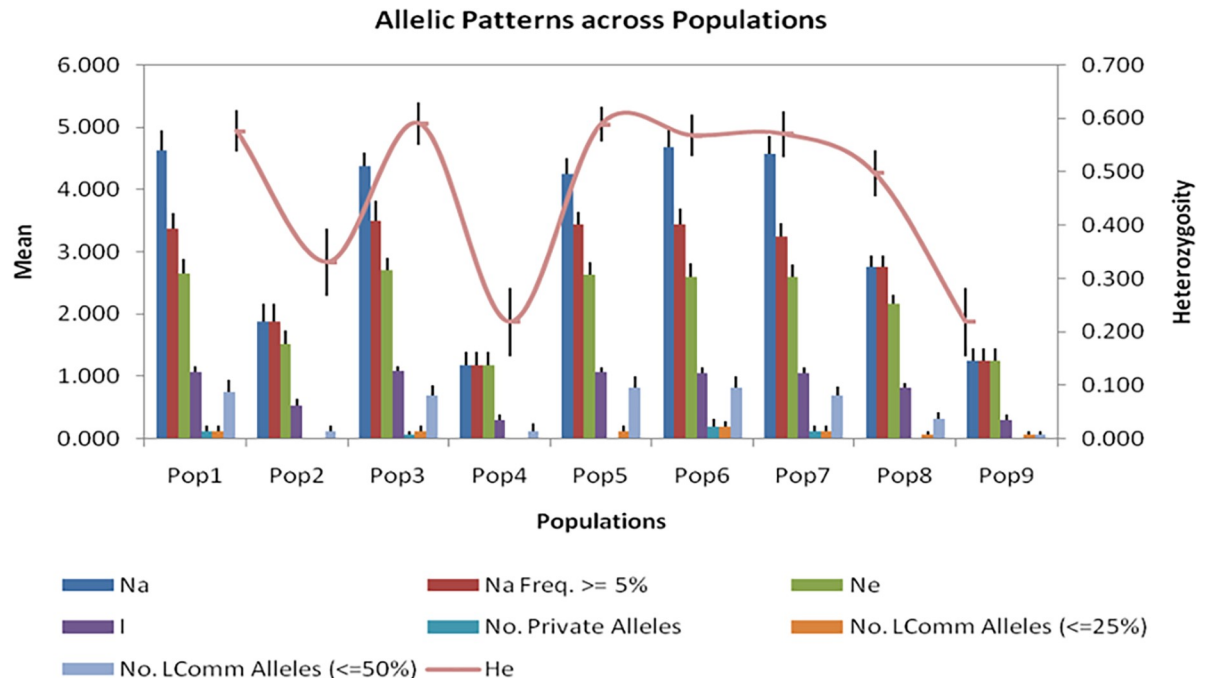


Fig 5. Summary statistics grouped by subpopulations, where N_a represents the number of different alleles; N_a ($Freq > = 5\%$) the number of different alleles with a frequency $\geq 5\%$, N_e the number of effective alleles, which is equal to $1 / (\sum p_i^2)$; I , the Shannon and Weaver information index, calculated as $-\sum (p_i \ln(p_i))$; $No. private alleles$, the number of alleles unique to a single population; $No. LComm alleles (< = 25\%)$, the number of locally common alleles ($Freq. > = 5\%$) found in 25% or fewer populations; $No. LComm alleles (< = 50\%)$, the number of locally common alleles ($Freq. > = 5\%$) found in 50% or fewer populations; H_e the Expected Heterozygosity; and UH_e the Unbiased Expected Heterozygosity, estimated as $H_e(2N / (2N-1))$.

<https://doi.org/10.1371/journal.pone.0240410.g005>

suggests the reduction in population size in the 1950s was perhaps not as significant a bottleneck event as previously reported [9]. When the theory behind the various models is re-examined [36] it becomes evident that gene diversity excess has only been demonstrated for loci evolving under the Infinite Allele Model. Given that there is very strong evidence to support a recent bottleneck event under this model, which is supported by testing of microsatellite allele frequency data herein, it is likely that the Cleveland Bay horse has indeed experienced a recent genetic bottleneck.

Population structure

Wright F Parameters [37] reflecting departure from Hardy–Weinberg equilibrium were calculated from the pedigree analysis for the reference population in terms of F_{IS} (0.006677), F_{ST} (0.040230) and F_{IT} (0.033821). Multilocus estimations of Wright's F statistics [38] from the microsatellite data showed an across population distribution of the following: F_{IS} (0.011362), F_{IT} (0.029308), and F_{ST} (0.018153).

Distance matrices [39] were constructed from both pedigree and molecular analysis, and phylogenetic trees were constructed using TReX [33] showing the relative positions of each female ancestry line (Figs 7 and 8).

Both the pedigree distance analysis (Fig 7) and the molecular analysis (Fig 8) are suggestive of a population structure rooted on three sub-divisions, or clades. However, neither analysis provides conclusive evidence of the causes or nature of this division. In addition to the pairwise distance matrices constructed assuming 9 subgroups within the population, GENALEX 6.4 [40] was also used to construct the much larger matrix of Nei distance between individuals

Table 8. Bottleneck heterozygosity excess test results based on 16 identified loci, where n represents the sample size; and k_o , the observed number of alleles under the assumption of mutation-drift equilibrium.

locus	Observed			IAM				TPM				SMM			
	n	k_o	H_e	H_{exp}	SD	D_H/SD	$Prob$	H_e	H_{exp}	SD	$Prob$	H_{exp}	SDs	D_H/SD	$Prob$
VHL20	804	5	0.7	0.4	0.19	1.56	0.02	0.51	0.15	1.21	0.06	0.64	0.09	0.67	0.28
HTG4	804	4	0.43	0.33	0.2	0.52	0.37	0.42	0.17	0.11	0.47	0.55	0.12	-0.96	0.16
AHT4	804	7	0.53	0.51	0.18	0.13	0.48	0.63	0.12	-0.89	0.16	0.75	0.06	-3.8	0.01
HMS7	804	6	0.75	0.45	0.19	1.62	0.01	0.57	0.14	1.34	0.03	0.71	0.07	0.7	0.25
HTG6	74	4	0.68	0.46	0.17	1.37	0.05	0.53	0.14	1.1	0.1	0.6	0.1	0.8	0.21
AHT5	804	7	0.67	0.5	0.18	0.94	0.19	0.63	0.11	0.31	0.45	0.74	0.06	-1.2	0.1
HMS6	804	5	0.57	0.39	0.2	0.9	0.22	0.51	0.16	0.42	0.4	0.63	0.1	-0.64	0.2
ASB2	800	7	0.57	0.5	0.18	0.43	0.41	0.63	0.12	-0.43	0.26	0.74	0.06	-2.73	0.02
HTG10	804	4	0.68	0.33	0.2	1.77	0.02	0.43	0.17	1.45	0.03	0.55	0.12	1.08	0.09
HTG7	78	3	0.5	0.33	0.18	0.9	0.26	0.41	0.16	0.57	0.37	0.46	0.14	0.28	0.48
HMS3	802	6	0.2	0.45	0.19	-1.36	0.14	0.58	0.13	-2.89	0.02	0.7	0.08	-6.29	0
HMS2	76	5	0.54	0.53	0.16	0.09	0.45	0.6	0.12	-0.52	0.25	0.68	0.08	-1.7	0.07
ASB17	634	9	0.72	0.59	0.15	0.84	0.21	0.71	0.09	0.08	0.45	0.8	0.05	-1.83	0.05
ASB23	742	5	0.72	0.41	0.19	1.66	0.02	0.51	0.15	1.4	0.03	0.63	0.1	0.93	0.14
LEX33	646	8	0.76	0.56	0.17	1.14	0.08	0.68	0.1	0.79	0.21	0.77	0.05	-0.34	0.29

The IAM, SMM and TPM mutation models simulate the coalescent processes of n genes. H_{exp} is the average heterozygosity and used to compare with the observed value in determining a heterozygosity excess or deficit at each locus. The standardised difference for each locus is estimated based on the inverse product of the Nei gene diversity and standard deviation (SD) of the mutation-drift equilibrium.

<https://doi.org/10.1371/journal.pone.0240410.t008>

[39]. This matrix in Phylip format was imported into the cluster drawing programme Splitstree4 [41] to construct a Neighbour-Net diagram. The Neighbour Net Diagram indicating a 2 clade subdivision of the population is shown at Fig 9 whilst a 3 clade subdivision is shown at Fig 10.

Examination of this net immediately suggests that the structure of the reference population could be explained by two broad groups or clades as shown in Fig 9. However, an alternative model with three clades, shown in Fig 10, is also possible.

Principal co-ordinate analysis *via* covariance matrix was conducted using Genalex 6.5 [40], with sub-populations assigned by both modern female and modern male ancestry lines, in order to examine alternative possible structuring of the reference population. Fig 11 presents the PCoA with subpopulations assigned by female ancestry and Fig 12 by male ancestry.

The PCoA analysis shows both male and female sub-populations distributed widely across principal axes, with little suggestion of structuring by sex group being the driving process of population sub-division in the microsatellite data. Variational Bayesian analysis of the microsatellite dataset, using the programme STRUCTURE [34] was carried out, in order to further investigate breed structure. 104 runs of the analysis were carried out for potential populations, K , numbering 2 to 25. The best fit of K appears at $K = 3$. Fig 13 provides a visual representation

Table 9. Tests for null hypothesis under three microsatellite evolution models.

TEST/MODEL	IAM	TPM	SMM
Sign Test: Number of loci with heterozygosity excess (probability)	8.93* (0.00120)	9.40 (0.29262)	9.43 (0.06923)
Standard differences test: Ti values (probability)	3.186* (0.00072)	0.902 (0.18357)	4.294* (0.00001)
Wilcoxon Rank Test (probability of heterozygosity excess)	0.00042*	0.11560	0.97116
*Rejection of null hypothesis (bottleneck) $P < 0.05$			

<https://doi.org/10.1371/journal.pone.0240410.t009>

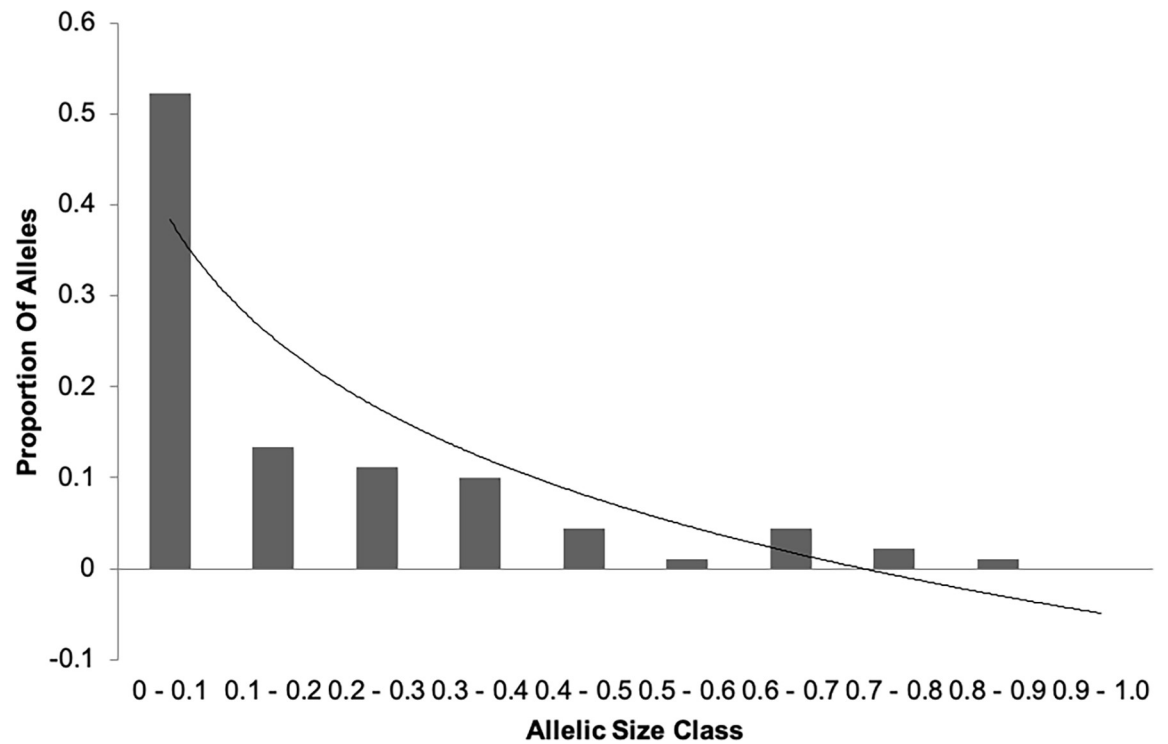


Fig 6. Allele distribution by size class. Trendline describes a natural logarithmic relationship according to $y = -0.188 \ln(x) + 0.3836$.

<https://doi.org/10.1371/journal.pone.0240410.g006>

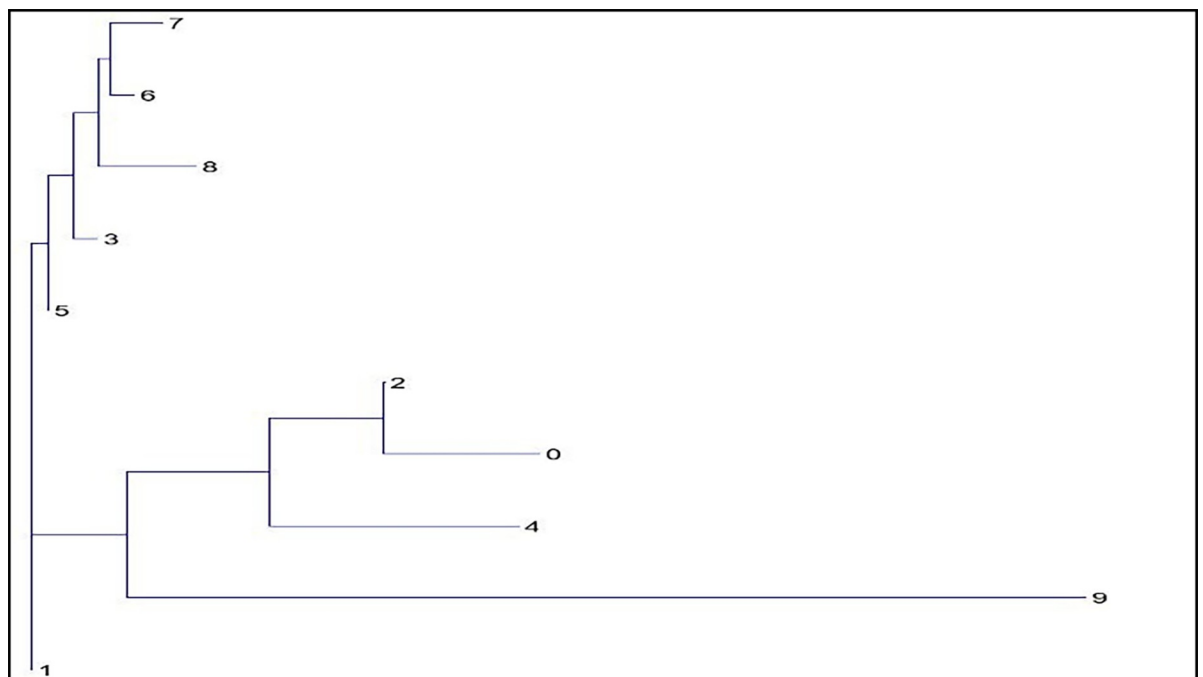


Fig 7. Neighbour joining tree showing relative genetic distance between subgroups from analysis of pedigree data assigned by female ancestry line.

<https://doi.org/10.1371/journal.pone.0240410.g007>

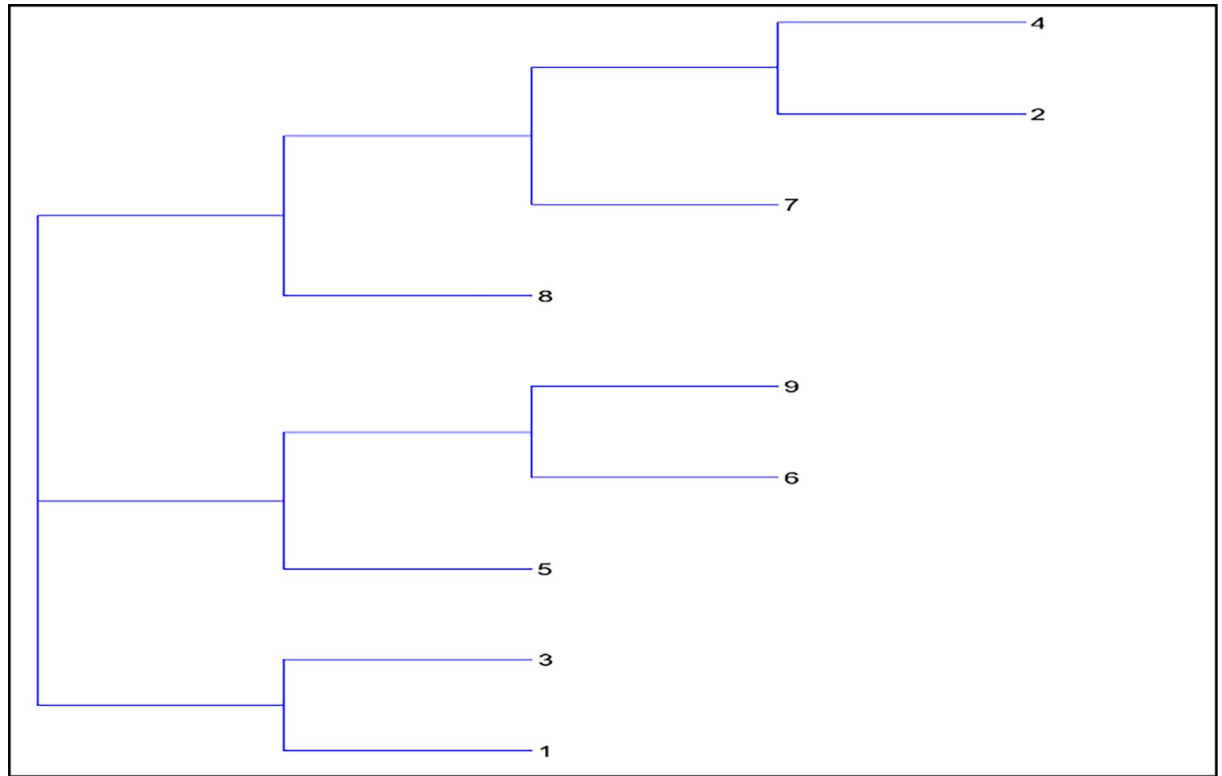


Fig 8. Neighbour joining tree from microsatellite analysis showing distances between subpopulations by maternal ancestry line.

<https://doi.org/10.1371/journal.pone.0240410.g008>

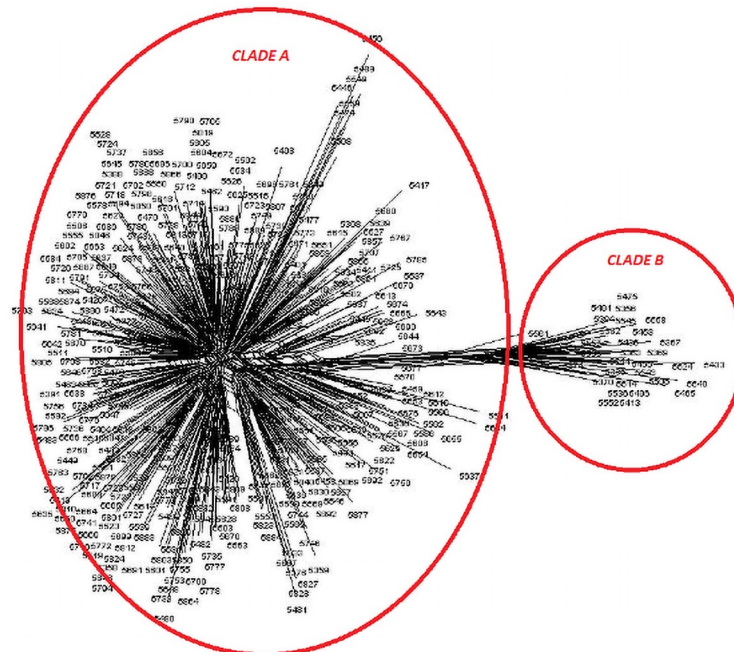


Fig 9. Neighbour-Net diagram of Nei genetic distance between individuals showing two clade model of structure.

<https://doi.org/10.1371/journal.pone.0240410.g009>

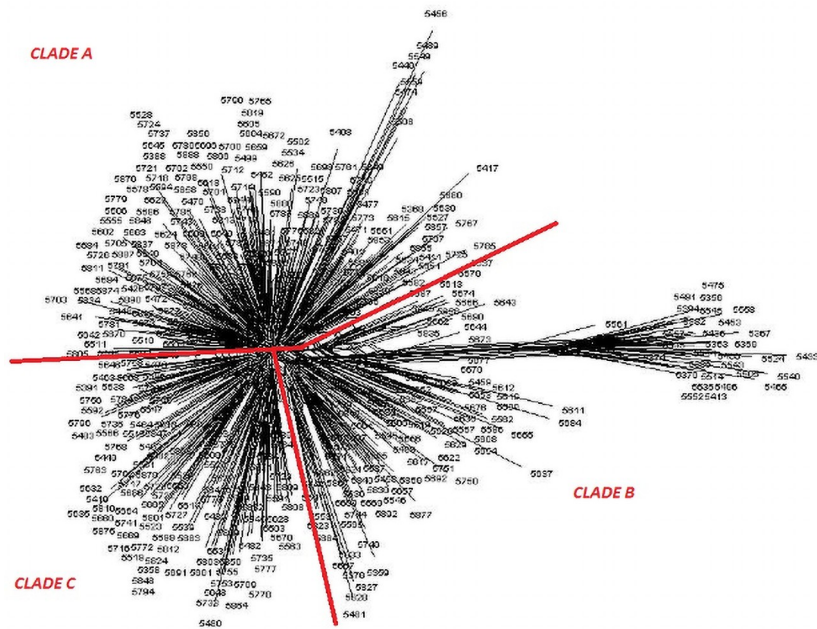


Fig 10. Neighbour-Net diagram of Nei genetic distance between individuals showing three clade model of structure.

<https://doi.org/10.1371/journal.pone.0240410.g010>

of this analysis for $K = 2$ to $K = 4$. There is a substantial increase in background noise in the display at $K = 4$, indicative that the number of clusters or sub-populations is below this level.

Further analysis of the population structure was conducted using the programme BAPS [42]. 17 clusters within the microsatellite dataset were identified, with a highly significant probability of 0.99998.

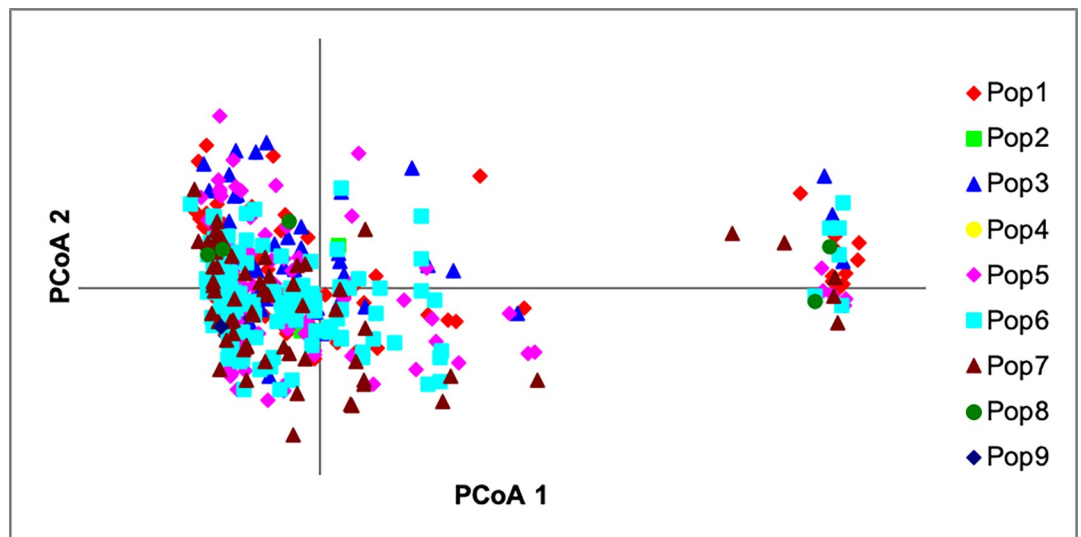


Fig 11. Principal coordinate analysis (PCoA) with subpopulations assigned by female ancestry across the two principal components (PCoA1, PCoA2).

<https://doi.org/10.1371/journal.pone.0240410.g011>

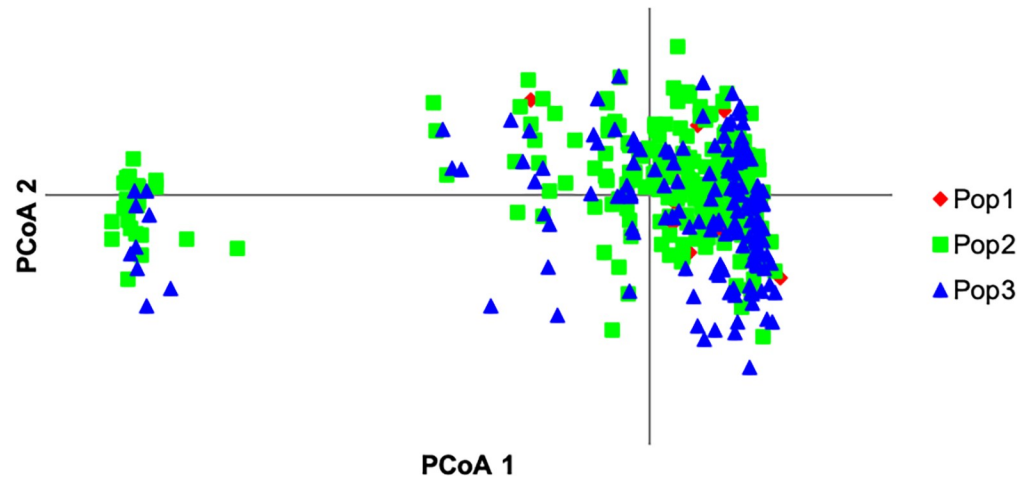


Fig 12. PCoA with subpopulations assigned by male ancestry.

<https://doi.org/10.1371/journal.pone.0240410.g012>

Discussion

The results presented herein highlight the significant losses of founder representation that have occurred in the Cleveland Bay Horse population across the past century. Approximately 91% of the stallion and 48% of the dam lines are lost in the reference population. The unbalanced representation of the founders is illustrated by the effective number of founder animals (f_e) and the effective number of ancestors (f_a). The parameter f_e constitutes over a third of the equivalent number of founder animals for the reference population, whilst the ratio f_a/f_e is 22.5%. This ratio is substantially lower than that reported in other horse breeds such as 41.7% in the Andalusian [43] or 54.4% in the Lipizzan [44]. Additionally, this is lower than the figure of 38.2% reported for the endangered Catalonian donkey [26].

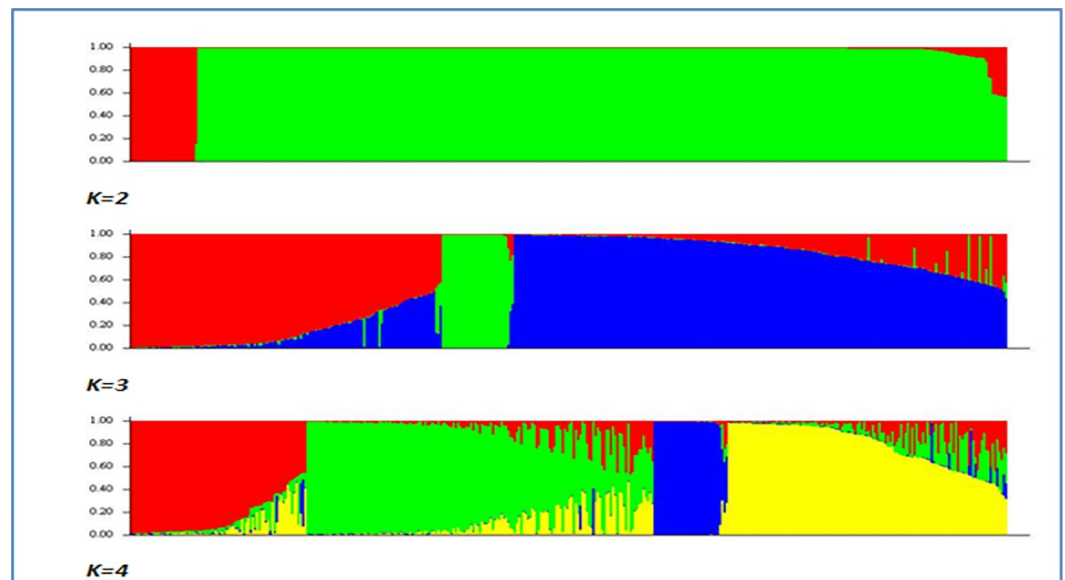


Fig 13. STRUCTURE analysis of population numbers $K = 2$ to $K = 4$. Each colour is a representation of a population, with individuals shown as vertical lines, which are split into coloured segments; the lengths of these describe the admixture proportions from K populations.

<https://doi.org/10.1371/journal.pone.0240410.g013>

The values of the generation interval presented herein (Table 2) are common in equines and identical to those observed in the literature [22, 43]. Suggesting some lack of regularised control measures or quantitative breeding strategies on the part of breeders and a decrease in genetic gain, which is directly linked to the generation interval. Breeders should start programmes at a younger age and decrease breeding extent over time.

The average inbreeding computed for the Cleveland Bay Horse at 20.64% in the reference population is substantially higher than most of the values reported in the literature [43], with typical values ranging from 6.5% to 12.5%. Although most of these inbreeding values have been computed in breeds with deep pedigrees such as Andalusian, Lipizzan or Thoroughbred there are significant differences in population sizes, and the accumulation of inbreeding in populations of restricted size will occur at a greater rate.

The smaller the number of individuals in a randomly mating breed the greater will be the accumulation of inbreeding due to the restricted choice of mates. Furthermore, we see a smaller N_e with increasing ΔF . The Cleveland Bay horse is therefore predisposed to inbreeding and associated loss of genetic variation. In the reference population of 402 individuals the Effective Population Size (N_e) computed *via* individual increase in inbreeding was 27.84. N_e computed *via* regression on equivalent generations was 26.29. Inbreeding and genetic loss under random mating will occur at $\frac{1}{2} N_e$ per generation. In the reference population, where Mean N_e is 32.32 under random mating, inbreeding can be expected to accumulate at 1.5% per generation.

This is reflected by the genealogical F_{IS} values. This parameter characterises the mating policy derived from the departure from random mating as a deviation from Hardy–Weinberg equilibrium. Positive F_{IS} values indicate that the average F value within a population exceeds the between-individuals coancestry, thus suggesting that matings between relatives have taken place [26]. Moreover, the average AR values computed for nine complete generations, (Table 3) are roughly equivalent the value of F . In an ideal scenario with random matings and no population subdivision, AR would be approximately twice the F value of the next generation [26].

Molecular information obtained in this study using microsatellite analysis suggests that genetic diversity within the breed is more restricted than has been reported in many other horse breeds and is based on an assessment of the tendency of genetic characteristics to vary accordingly (Table 10).

Populations that have experienced a recent reduction in their N_e exhibit a correlative reduction of the allele numbers (k) and gene diversity (H_e) at polymorphic loci. However, the allele numbers reduce faster than the genetic diversity. Thus, in a recently bottlenecked population, the observed gene diversity is higher than the expected equilibrium gene diversity (H_e) which is computed from the observed number of alleles, k , under the assumption of a constant-size or equilibrium population [36]. The existence of a population bottleneck in the mid twentieth century, when the number of breeding age Cleveland Bay stallions was reduced to four, has previously been reported [12]. There is clear genetic evidence of this event shown in the excess of observed heterozygosity across subpopulations, with the exception of ancestry line nine. The latter is of more recent origin having evolved from a grading up scheme in the latter half of the twentieth century. In all other subgroups, the excess is positive ranging from 2.12% in Line 5 to 19.6% in Line 4. However, this investigation has revealed that lines two, four, and eight are in fact not polymorphic. The observed heterozygosity excess amongst the five polymorphic lines peaks in line one at 6.1%.

Microsatellite multilocus estimations of Wright's F statistics [38] showed an across population F_{IS} , F_{IT} and F_{ST} of 0.01758, 0.02490, and 0.00745, respectively. This departure from random mating will have been influenced by a number of factors common to restricted

Table 10. Genetic variability from microsatellite DNA loci for Cleveland Bay and other domestic horse breeds.

Breed	He	Ho	MNA	Source
Cleveland Bay Horse	0.173–0.805	0.052–0.716	6.19	
Suffolk Punch	0.724	0.679	6.42	[45]
Dales Pony	0.654	0.715	5.58	[46]
Exmoor Pony	0.609	0.601	5.25	[47]
Fell Pony	0.731	0.782	6.42	[46]
Irish Draught	0.772	0.766	7.08	[48]
Shetland	0.661	0.642	5.3	[49]
Thoroughbred	0.646–0.732	0.628–0.671	4.7–7.5	[2]
Thoroughbred	0.695	0.674	6.25	[50]
Arabian	0.690	0.624	6.58	[27]
Lipizzan	0.675	0.663	7.1	[51]
Friesian	0.466	0.454	4.5	[52]
Spanish Celtic Horses	0.677–0.770	0.694–0.765	5.2–7.8	[53]
Portuguese Garrano	0.751	0.732	4.5	[52]
Lithuanian	0.442–0.770	0.452–0.785	2.0–4.7	[54]
Sorraia Horse	0.093–0.736	0.088–0.705	3.3	[55]

H_e denotes the expected heterozygosity, whilst H_o represents the observed heterozygosity, and MNA the mean number of alleles per locus.

<https://doi.org/10.1371/journal.pone.0240410.t010>

populations of domesticated equines. These include: selection by breeders for particular lines of descent; natural differences in fertility between individuals; a restricted number of male animals leaving significantly more offspring than females (disproportionate male founding) and geographic distribution of animals and breeders leading to logistical difficulties in some matings. The reduced number of alleles and fixation at certain loci in female ancestry lines is evidence of loss of founder representation from these lines. This lower heterozygosity is also indicative of the typical practice of the larger studs, where breeding tends to be carried out in pasture by free live cover, with the use of only one stallion per year, per herd and where the same stallion may be retained for several breeding years. This strategy is compounded by breeders with only a small number of breeding females sending their animals to these groups or to be covered in hand by the same stallion.

This strategy has different implications for the genetic diversity of the Cleveland Bay Horse compared that of mares travelling to stud to be covered in hand by a greater range of stallions that do not have their own herds of mares [55], as well as through trade or exchange, which will change geographic location albeit on an irregular basis. Although this latter practice has clear benefits in conservation programmes, there is the danger of inappropriate matings supplanting the more common and less frequent alleles. Whilst such matings increase the frequency of the rarer alleles, they simultaneously increase the frequency of those more common [56], highlighting the need for in-depth understanding of the genetic diversity of any rare breed, and for an effective management plan for conservation maintenance.

There has been considerable debate about the most effective methods of conserving and managing endangered populations [55]. Before the advent of mitochondrial and microsatellite DNA analysis, the accepted strategy involved minimizing inbreeding, whilst managing mean Kinship/average relatedness [57]. Moreover, the use of molecular methods has been proposed [58, 59]. Where pedigree data is robust and complete over a significant number of generations, it appears that genealogical data remains the preferred method by which to manage founder

contributions, inbreeding and kinship/relatedness. Indeed Lacy has highlighted the problems caused in conservation programmes based on private or rare alleles [56].

Variational Bayesian analysis of within-population structure using microsatellite data shows significant evidence for three main clades. Although this study has been based on the use of pedigree and microsatellite marker data for the Cleveland Bay horse there is now firm evidence of the value of mitochondrial DNA for such investigations and an increasing number of investigations consider the origins and relatedness of modern equines (Table 10).

The Cleveland Bay horse has been reported to belong to haplotype C [48] which is common amongst older northern European breeds such as the Exmoor, Icelandic, Fjord, Connemara and Scottish Highland. This correlates with the assertion that in the matriline the Cleveland Bay has evolved from the Chapman; an ancient Northern European breed. The comparative studies have been based on five Cleveland Bay mtDNA sequences deposited in GeneBank by Cothran and Frankham within which there are three haplotypes. There is scope for further sampling of all of the existing matriline to determine the number of haplotypes present in the reference population the level of correlation with the three Clades identified herein.

Conclusion

We have reported an in-depth genetic analysis of the Cleveland Bay Horse, using both pedigree and microsatellite data. It reveals substantial loss of genetic diversity and high levels of relatedness and inbreeding. The results of this study highlight the importance of the Cleveland Bay Horse community implementing an effective and sustainable breed management plan, such as management of Mean Kinship and Inbreeding Coefficients.

Acknowledgments

The authors thank the Breed Committee of the Cleveland Bay Horse Society for access to its microsatellite parentage testing records.

Author Contributions

Conceptualization: Andrew Dell, Mark Curry.

Data curation: Andrew Dell, Mark Curry, Philippe B. Wilson.

Formal analysis: Andrew Dell, Mark Curry, Kelly Yarnell, Philippe B. Wilson.

Funding acquisition: Mark Curry.

Investigation: Andrew Dell, Mark Curry.

Methodology: Andrew Dell, Mark Curry.

Project administration: Mark Curry, Gareth Starbuck, Philippe B. Wilson.

Resources: Gareth Starbuck, Philippe B. Wilson.

Software: Gareth Starbuck.

Supervision: Kelly Yarnell.

Validation: Philippe B. Wilson.

Visualization: Andrew Dell, Kelly Yarnell, Gareth Starbuck.

Writing – original draft: Andrew Dell.

Writing – review & editing: Kelly Yarnell, Gareth Starbuck, Philippe B. Wilson.

References

1. Petersen JL, Mickelson JR, Cleary KD, McCue ME. The american quarter horse: Population structure and relationship to the thoroughbred. *Journal of Heredity*. 2014. <https://doi.org/10.1093/jhered/est079> PMID: 24293614
2. Cunningham EP. Molecular methods and equine genetic diversity. *Conservation Genetics of Endangered Horse Breeds*. 2005.
3. Cunningham EP, Dooley JJ, Splan RK, Bradley DG. Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Anim Genet*. 2001. <https://doi.org/10.1046/j.1365-2052.2001.00785.x> PMID: 11736806
4. Boyd MM. A plea for a more extended use of the system of live-stock registration. *J Hered*. 1907. <https://doi.org/10.1093/jhered/os-3.1.255>
5. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000. PMID: 10835412
6. Khanshour AM, Hempsey EK, Juras R, Cothran EG. Genetic characterization of Cleveland Bay Horse Breed. *Diversity*. 2019. <https://doi.org/10.3390/d11100174>
7. EMMERSON S. Cleveland Bay Horse Society Centenary Studbook. Clevel Bay Horse Soc. 1984.
8. Dent AA. Cleveland Bay Horses. JA Allen & Company, Limited; 1978.
9. Fairfax-Blakeborough J. Cleveland bay horse, its history, evolution and importance today. 1950.
10. Reese HH. Breeds of light horses. US Department of Agriculture; 1918.
11. Johnson D. Horse breeds. Voyageur Press; 2008.
12. WALLING G. Cleveland Bay Horse Society Studbook Vol XXXIII. Clevel Bay Horse Soc. 1994.
13. Khanshour AM, Juras R, Cothran EG. Microsatellite analysis of genetic variability in Waler horses from Australia. *Aust J Zool*. 2014; 61: 357–365.
14. Lacy RC. Management of limited animal populations. Bottleneck dolphin reproduction workshop. 2000.
15. Ansari-Mahyari S, Berg P. Power of QTL mapping using both phenotype and genotype information in selective genotyping.
16. Groeneveld E, Westhuizen BD, Maiwashe A, Voordewind F, Ferraz JB. POPREP: a generic report for population management. *Genet Mol Res*. 2009. <https://doi.org/10.4238/vol8-3gmr648> PMID: 19866435
17. Maccluer JW, Boyce AJ, Dyke B, Weitkamp LR, Pfenning DW, Parsons CJ. Inbreeding and pedigree structure in standardbred horses. *J Hered*. 1983. <https://doi.org/10.1093/oxfordjournals.jhered.a109824>
18. J-M G., Falconer DS. Introduction to Quantitative Genetics. Popul (French Ed. 1962. <https://doi.org/10.2307/1525780>
19. Gutiérrez JP, Goyache F. A note on ENDOG: A computer program for analysing pedigree information. *J Anim Breed Genet*. 2005. <https://doi.org/10.1111/j.1439-0388.2005.00512.x> PMID: 16130468
20. Boichard D, Maignel L, Verrier É. The value of using probabilities of gene origin to measure genetic variability in a population. *Genet Sel Evol*. 1997. <https://doi.org/10.1186/1297-9686-29-1-5>
21. Lacy RC. Analysis of founder representation in pedigrees: Founder equivalents and founder genome equivalents. *Zoo Biol*. 1989. <https://doi.org/10.1002/zoo.1430080203>
22. Bijma P, Woolliams JA. Prediction of genetic contributions and generation intervals in populations with overlapping generations under selection. *Genetics*. 1999. PMID: 10049935
23. Cornuet JM, Luikart G. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*. 1996. <https://doi.org/10.1093/oxfordjournals.jhered.a111627> PMID: 8978083
24. Sørensen AC, Sørensen MK, Berg P. Inbreeding in danish dairy cattle breeds. *J Dairy Sci*. 2005. [https://doi.org/10.3168/jds.S0022-0302\(05\)72861-7\(05\)72861-7](https://doi.org/10.3168/jds.S0022-0302(05)72861-7(05)72861-7) PMID: 15829680
25. Lacy RC. Clarification of genetic terms and their use in the management of captive populations. *Zoo Biol*. 1995. <https://doi.org/10.1002/zoo.1430140609>
26. Gutiérrez JP, Marmi J, Goyache F, Jordana J. Pedigree information reveals moderate to high levels of inbreeding and a weak population structure in the endangered Catalanian donkey breed. *J Anim Breed Genet*. 2005. <https://doi.org/10.1111/j.1439-0388.2005.00546.x> PMID: 16274421
27. Khanshour A, Conant E, Juras R, Cothran EG. Microsatellite analysis of genetic diversity and population structure of Arabian horse populations. *J Hered*. 2013; 104: 386–398. <https://doi.org/10.1093/jhered/est003> PMID: 23450090
28. Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F. GENETIX 4.05, Windows TM software for population genetics. *Lab génome, Popul Interact CNRS Umr*. 2004;5000.

29. Hill WG. *Molecular Evolutionary Genetics*. By Masatoshi Nei. New York: Columbia University Press. 1987. 512 pages. U.S. \$50.00. ISBN 0 231 06320 2. *Genet Res*. 1988. <https://doi.org/10.1017/s001667230002735x>
30. Wright S. Variability within and among natural populations. *Evolution and the genetics of populations*. 1978.
31. Caballero A, Toro MA. Analysis of genetic diversity for the management of conserved subdivided populations. *Conserv Genet*. 2002. <https://doi.org/10.1023/A:1019956205473>
32. Makarenkov V. T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*. 2001. <https://doi.org/10.1093/bioinformatics/17.7.664> PMID: 11448889
33. Alix B, Boubacar DA, Vladimir M. T-REX: A web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res*. 2012. <https://doi.org/10.1093/nar/gks485> PMID: 22675075
34. Pritchard JK. Documentation for structure software: Version 2. 2. *Statistics (Ber)*. 2007.
35. Chikhi L, Bruford M. Mammalian population genetics and genomics. *Mammalian Genomics*. 2004. <https://doi.org/10.1079/9780851999104.0539>
36. Luikart G, Cornuet JM. Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conserv Biol*. 1998. <https://doi.org/10.1046/j.1523-1739.1998.96388.x>
37. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. *Evolution (N Y)*. 1984. <https://doi.org/10.1111/j.1558-5646.1984.tb05657.x> PMID: 28563791
38. Weir BS, Hill WG. Estimating F-Statistics. *Annu Rev Genet*. 2002. <https://doi.org/10.1146/annurev.genet.36.050802.093940> PMID: 12359738
39. Nei M. *Molecular evolutionary genetics*. Columbia university press; 1987.
40. Peakall R, Smouse PE. GenAEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*. 2012; 28: 2537–2539. <https://doi.org/10.1093/bioinformatics/bts460> PMID: 22820204
41. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*. 2006. <https://doi.org/10.1093/molbev/msj030> PMID: 16221896
42. Corander J, Waldmann P, Marttinen P, Sillanpää MJ. BAPS 2: Enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*. 2004. <https://doi.org/10.1093/bioinformatics/bth250> PMID: 15073024
43. Valera M, Molina A, Gutiérrez JP, Gómez J, Goyache F. Pedigree analysis in the Andalusian horse: Population structure, genetic variability and influence of the Carthusian strain. *Livest Prod Sci*. 2005. <https://doi.org/10.1016/j.livprodsci.2004.12.004>
44. Zechner P, Sölkner J, Bodo I, Druml T, Baumung R, Achmann R, et al. Analysis of diversity and population structure in the Lipizzan horse breed based on pedigree information. *Livest Prod Sci*. 2002. [https://doi.org/10.1016/S0301-6226\(02\)00079-9](https://doi.org/10.1016/S0301-6226(02)00079-9)
45. Aberle K, Wrede J, Distl O. Analyse der Populationsstruktur des Süddeutschen Kaltbluts in Bayern. *Berl Munch Tierarztl Wochenschr*. 2004.
46. Fox-Clipsham LY, Brown EE, Carter SD, Swinburne JE. Population screening of endangered horse breeds for the foal immunodeficiency syndrome mutation. *Vet Rec*. 2011. <https://doi.org/10.1136/vr.100235> PMID: 22016514
47. Prystupa JM, Hind P, Cothran EG, Plante Y. Maternal lineages in native Canadian equine populations and their relationship to the Nordic and Mountain and Moorland pony breeds. *J Hered*. 2012. <https://doi.org/10.1093/jhered/ess003> PMID: 22504109
48. McGahern AM, Edwards CJ, Bower MA, Heffernan A, Park SDE, Brophy PO, et al. Mitochondrial DNA sequence diversity in extant Irish horse populations and in ancient horses. *Anim Genet*. 2006. <https://doi.org/10.1111/j.1365-2052.2006.01506.x> PMID: 16978181
49. Brinkmann L, Gerken M, Riek A. Adaptation strategies to seasonal changes in environmental conditions of a domesticated horse breed, the Shetland pony (*Equus ferus caballus*). *J Exp Biol*. 2012. <https://doi.org/10.1242/jeb.064832> PMID: 22399650
50. Gu J, Orr N, Park SD, Katz LM, Sulimova G, MacHugh DE, et al. A genome scan for positive selection in thoroughbred horses. *PLoS One*. 2009. <https://doi.org/10.1371/journal.pone.0005767> PMID: 19503617
51. Achmann R, Curik I, Dovc P, Kavar T, Bodo I, Habe F, et al. Microsatellite diversity, population subdivision and gene flow in the Lipizzan horse. *Anim Genet*. 2004. <https://doi.org/10.1111/j.1365-2052.2004.01157.x> PMID: 15265067

52. Luís C, Juras R, Oom MM, Cothran EG. Genetic diversity and relationships of Portuguese and other horse breeds based on protein and microsatellite loci variation. *Anim Genet*. 2007. <https://doi.org/10.1111/j.1365-2052.2006.01545.x> PMID: [17257184](https://pubmed.ncbi.nlm.nih.gov/17257184/)
53. Cañon J, Checa ML, Carleos C, Vega-Pla JL, Vallejo M, Dunner S. The genetic structure of Spanish Celtic horse breeds inferred from microsatellite data. *Anim Genet*. 2000. <https://doi.org/10.1046/j.1365-2052.2000.00591.x> PMID: [10690360](https://pubmed.ncbi.nlm.nih.gov/10690360/)
54. Juras R, Cothran EG, Klimas R. Genetic Analysis of Three Lithuanian Native Horse Breeds. *Acta Agric Scand—Sect A Anim Sci*. 2003. <https://doi.org/10.1080/09064700310012971>
55. Luís C, Cothran EG, Oom MDM. Inbreeding and genetic structure in the endangered Sorraia horse breed: Implications for its conservation and management. *J Hered*. 2007. <https://doi.org/10.1093/jhered/esm009> PMID: [17404326](https://pubmed.ncbi.nlm.nih.gov/17404326/)
56. Lacy RC. Should we select genetic alleles in our conservation breeding programs? *Zoo Biology*. 2000. [https://doi.org/10.1002/1098-2361\(2000\)19:4<279::AID-ZOO5>3.0.CO;2-V](https://doi.org/10.1002/1098-2361(2000)19:4<279::AID-ZOO5>3.0.CO;2-V)
57. Mills LS, Ballou JD, Gilpin M, Foose TJ. Population Management for Survival and Recovery: Analytical Methods and Strategies in Small Population Conservation. *J Wildl Manage*. 1997. <https://doi.org/10.2307/3802439>
58. Pearl MC. Research Techniques in Animal Ecology Methods and Cases in Conservation Science. *J Wildl Manage*. 2000. <https://doi.org/10.2307/3803113>
59. Fraser DJ, Bernatchez L. Adaptive evolutionary conservation: Towards a unified concept for defining conservation units. *Molecular Ecology*. 2001. <https://doi.org/10.1046/j.1365-294X.2001.t01-1-01411.x> PMID: [11903888](https://pubmed.ncbi.nlm.nih.gov/11903888/)