SCIENTIFIC REPORTS
natureresearch

OPEN

Genomic inbreeding trends, influential sire lines and selection in the global Thoroughbred horse population

Beatrice A. McGivney[1], Haige Han[1,2], Leanne R. Corduff[1], Lisa M. Katz[3], Teruaki Tozaki[4], David E. MacHugh[2,5] & Emmeline W. Hill[1,2*]

The Thoroughbred horse is a highly valued domestic animal population under strong selection for athletic phenotypes. Here we present a high resolution genomics-based analysis of inbreeding in the population that may form the basis for evidence-based discussion amid concerns in the breeding industry over the increasing use of small numbers of popular sire lines, which may accelerate a loss of genetic diversity. In the most comprehensive globally representative sample of Thoroughbreds to-date ($n = 10{,}118$), including prominent stallions ($n = 305$) from the major bloodstock regions of the world, we show using pan-genomic SNP genotypes that there has been a highly significant decline in global genetic diversity during the last five decades (F_{IS} $R^2 = 0.942$, $P = 2.19 \times 10^{-13}$; F_{ROH} $R^2 = 0.88$, $P = 1.81 \times 10^{-10}$) that has likely been influenced by the use of popular sire lines. Estimates of effective population size in the global and regional populations indicate that there is some level of regional variation that may be exploited to improve global genetic diversity. Inbreeding is often a consequence of selection, which in managed animal populations tends to be driven by preferences for cultural, aesthetic or economically advantageous phenotypes. Using a composite selection signals approach, we show that centuries of selection for favourable athletic traits among Thoroughbreds acts on genes with functions in behaviour, musculoskeletal conformation and metabolism. As well as classical selective sweeps at core loci, polygenic adaptation for functional modalities in cardiovascular signalling, organismal growth and development, cellular stress and injury, metabolic pathways and neurotransmitters and other nervous system signalling has shaped the Thoroughbred athletic phenotype. Our results demonstrate that genomics-based approaches to identify genetic outcrosses will add valuable objectivity to augment traditional methods of stallion selection and that genomics-based methods will be beneficial to actively monitor the population to address the marked inbreeding trend.

The Thoroughbred is among the fastest animals selected by humans for sport, originating from *"the commingled blood of Arabs, Turks and Barbs"*[1] crossed with local British and Irish mares[2] *"but selection and training have together made him a very different animal from his parent-stocks"*[1]. The Thoroughbred is now a large (N ~ 500,000) global breed but, in the context of modern horse breeds, it has very low genetic diversity[3,4] due to the limited foundation alleles at the establishment of the stud book and restriction of external gene flow subsequent to the closing of the population[5–7]. In Thoroughbred horse breeding selection of potential champion racehorses is a global multi-billion-dollar business, but there is no systematic industry-mediated genomic selection or genetic population management. We hypothesised that the market-driven emphasis on highly valuable pedigrees and the common practice of inbreeding to successful ancestors in attempts to reinforce favourable variants in offspring has resulted in a global reduction in genetic diversity. Here, we apply population genetics approaches to

[1]Plusvital Ltd, The Highline, Dun Laoghaire Business Park, Dublin, Ireland. [2]UCD School of Agriculture and Food Science, University College Dublin, Dublin, Ireland. [3]UCD School of Veterinary Medicine, University College Dublin, Dublin, Ireland. [4]Genetic Analysis Department, Laboratory of Racing Chemistry, Utsunomiya, Tochigi, Japan. [5]UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland. *email: Emmeline.Hill@ucd.ie

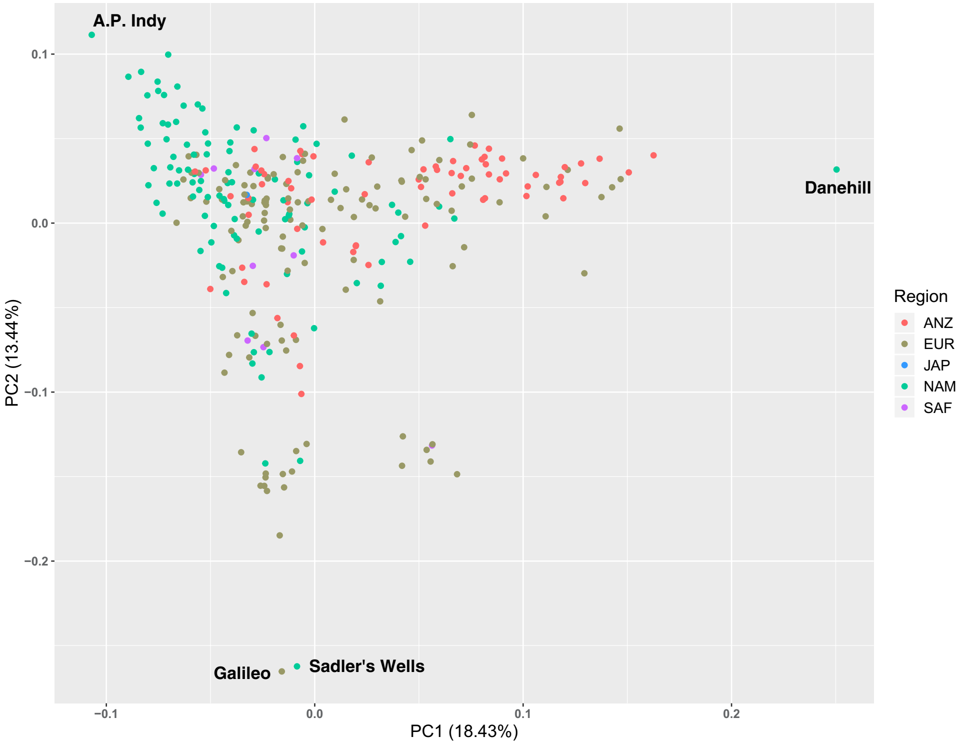

Figure 1. Principal component analysis plot of the genetic relatedness matrix based on genotype data for prominent global Thoroughbred stallions ($n = 305$). Individuals are colour coded based on region of birth: Australia/New Zealand (ANZ), red; Europe (EUR), green; Japan (JAP), blue; North America (NAM), light green; South Africa (SAF), purple.

assess temporal population-wide variation across the longest time span to-date for this population, and for the first time perform a comparative analysis of genetic diversity among the major bloodstock regions of the world. Additionally, we identify genes of interest for the Thoroughbred phenotype in signatures of positive selection that are likely to be most impacted by inbreeding.

Results and Discussion

Genetic diversity driven by 'breed-shaping' stallions within a highly homogeneous global population.

We evaluated genetic diversity in a principal component analysis (PCA) of the genetic relatedness matrix between Thoroughbreds and representatives of putative founding populations and within the global Thoroughbred population. We show that the Thoroughbred breed is divergent from the founding populations (S1-S2 Figures and S1 Text). However, although the population is geographically dispersed, with the majority of horses located in Australasia (ANZ), Europe (EUR), Japan (JAP), North America (NAM) and South Africa (SAF), the Thoroughbred ($n = 10{,}118$, 1970–2017) is largely genetically homogeneous maintained in a single cluster, albeit with some level of geographic population structure particularly towards EUR samples. The outliers driving diversity from the main cluster are prominent stallions (S3-S4 Figures) and visualisation of relatedness among the stallion population ($n = 305$) reveals partitioning in PC1 and PC2 apparently driven by the 'breed-shaping' sire lines in each region; *Sadler's Wells* (1981, NAM) in EUR; *Danehill* (1986, NAM) in ANZ and *A.P. Indy* (1989, NAM) in NAM (Fig. 1, S5 Figure). PCA plots illustrating the genetic diversity in each of the major regions EUR, ANZ and NAM are provided in S6–S11 Figures.

Northern Dancer was arguably the most successful stallion of the 20th century and his descendants have been the dominant sire lines in Australia and Europe for the last quarter century. *Sadler's Wells*, a son of *Northern Dancer*, is the single most successful stallion of the modern era siring, among many world-class horses, *Galileo*. *Galileo* has been the leading stallion in Great Britain and Ireland for the last decade (except *Danehill Dancer* in 2009). *Danehill*, also a grandson of *Northern Dancer*, was champion sire in Australia (1995–1997, 2000–2005), Great Britain and Ireland (2005, 2006, 2007) and France (2001, 2007) and has dominated Australian pedigrees since the 1990s. *A.P. Indy,* in contrast to 97% of the Thoroughbred population (S1 Text), does not trace his ancestry back to the influential *Northern Dancer*. *A. P. Indy* was the leading sire in North America in 2003 and 2006 and was among the top ten sires for ten consecutive years.

The influence of these sire lines in the PCA is reflected in the economic demand for popular bloodlines in the yearling sales market in which horses are valued largely on the basis of pedigree. In 2018, the top 24% (by average sale price) of stallions ($n = 82$) with progeny sold in the Australian Magic Millions Gold Coast sale (Book 1), were responsible for siring 67% of horses sold ($n = 695$) and received 79% of the total sale income (AUD 156.9 M). In Europe (Ireland) the 2018 Goffs Orby yearling sale grossed €43.5 M from horses sired by 78 stallions and the top

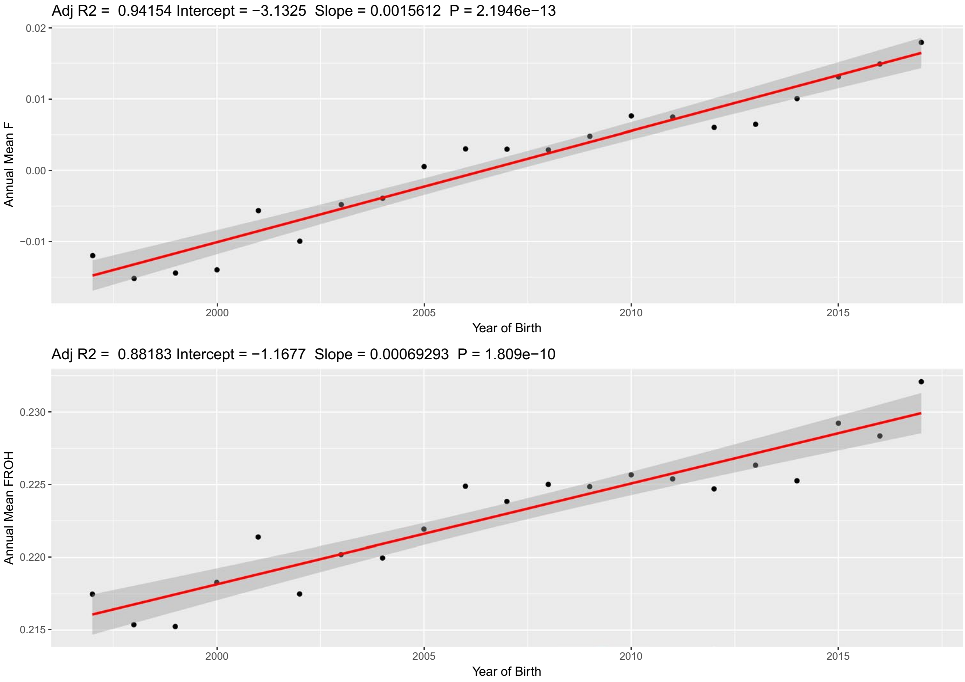

Figure 2. The regression of the mean annual F_{IS} [top] and F_{ROH} [bottom] on year of birth for $n = 10{,}118$ Thoroughbred horses born between 1996 and 2017.

25% of stallions (by gross sale income) were responsible for 78% of the total sale income (€34.1 M). In both markets, 16 and 19 of the top stallions, respectively, traced directly to *Northern Dancer* by paternal descent.

A marked increase in inbreeding in the Thoroughbred population over five decades. Here, in the most comprehensive genetic analysis performed in this population ($n = 10{,}118$) we show a striking temporal increase in inbreeding and regional variance across the global Thoroughbred population during the last five decades. Individual inbreeding coefficients (F_{IS}) estimated using 9,212 pruned SNPs and runs of homozygosity (F_{ROH}) (minimum length 1 Mb) characterised using an unpruned set of 46,478 SNPs revealed a highly significant increase in inbreeding over time (F_{IS}: $R^2 = 0.942$, $P = 2.19 \times 10^{-13}$; F_{ROH}: $R^2 = 0.88$, $P = 1.81 \times 10^{-10}$) (Fig. 2) with the greatest rate of change observed since the 2000s (S12 Figure). A linear regression model was used to test for significance and directionality of change in annual mean inbreeding with respect to year of birth (S12 Figure). Inbreeding estimates were determined using two methods: individual inbreeding coefficients (F_{IS}) measure observed versus expected genetic diversity in an individual in a population. F_{ROH} estimates the proportion of the genome covered by runs of homozygosity. The values are correlated but the unit of measurement is different so the absolute values cannot be directly compared. Similar trends for both measures were observed within each geographic region (S13-S16 Figures). Results from Student's t-tests indicated that inbreeding estimates were higher in EUR compared to ANZ (F_{IS}, $P = 1.9 \times 10^{-21}$; F_{ROH}, $P = 0.043$) and NAM (F_{IS}, $P < 3.1 \times 10^{-19}$; F_{ROH}, $P = 0.00059$). A temporal reduction of Thoroughbred genetic diversity has previously been reported among much smaller sample sets from single geographic regions[8,9]. Expansion of the timeline to include the most recent decade using a sample size more than 20-fold larger indicates that despite industry cognizance of inbreeding and previous cautions[8], there has been no arrest in the rate of increase in inbreeding and it is a global, population-wide phenomenon.

Breeding practices that promote inbreeding have not resulted in a population of faster horses[4,10] and our results, generated for the first time using a large cohort of globally representative genotypes, corroborate this. When evaluating Timeform handicap ratings for EUR horses, no association was found between F_{IS} and racing performance for horses born between 1996–2013 ($n = 1{,}886$, $r^2 = 0.034$, $P = 0.780$) and 2009–2013 ($n = 1{,}065$, $r^2 = -0.014$, $P = 0.65$). Purposeful inbreeding attempts to duplicate favourable gene variants that are selected at each successive generation, but homozygosity arising from inbreeding tends to be associated with decreased trait values since the proportionate trait gain arising from beneficial mutations is generally limited by pleiotropy[11]. Also, among domesticates, there is an increased mutational load and a higher proportion of deleterious alleles in regions under selection[12–15]. Therefore, there may be a genetic cost of inbreeding, which in dogs has negatively influenced intra-breed genetic diversity and increased the frequency of disease-causing alleles[15,16] and in cattle, is associated with decreases in milk, fat and protein yields and negatively impacts reproductive traits[17].

While inbreeding depression generally negatively impacts health and fertility through the accumulation of deleterious alleles, the 'tipping point' at which there is an irreversible accumulation of unfavourable mutations is not currently known. In horses mutational load and purging of deleterious alleles has been assessed in whole genome sequence data and revealed a significant relationship between inbreeding and mutational load[18].

Surprisingly, according to a recent study[18], Thoroughbreds appear to have a lower than expected mutational load, suggested to be due to effective purging through negative selection on phenotypes. This may be facilitated in the Thoroughbred in practice by the unusually large census population size relative to the effective population size, with a high proportion of horses that do not ever race[19]. However, the results from that study are likely not representative of the genetics of the current breeding population in the major bloodstock regions of the world; the sample cohort of Thoroughbreds examined was small ($n = 19$) and all the horses were registered Thoroughbreds in Korea. For example, ROH identified in that study were up to 11 Mb, whereas we have detected here ROH up to 40 Mb (S1 Table); increasing ROH increases the likelihood of deleterious alleles being exposed.

It is interesting to consider also that many of the performance limiting genetic diseases in the Thoroughbred do not generally negatively impact on suitability for breeding; some diseases, with known heritable components, are successfully managed by surgery (osteochondrosis desicans, recurrent laryngeal neuropathy), nutritional and exercise management (recurrent exertional rhabdomyolysis) and medication (exercise induced pulmonary haemorrhage). This facilitates retention of risk alleles in the population and enhances the potential for rapid proliferation of risk alleles if they are carried by successful stallions. Furthermore, since single gene variants have not yet been identified for these diseases, it is likely that they result from polygenic additive genetic variation that may not be easily exposed and purged by negative selection on individual ROH. In order to fully understand the consequences of inbreeding in the Thoroughbred, inbreeding measures should be regressed on large population-scale phenotypes and the extent of mutational load should be determined in a large cohort of representative samples.

There are currently concerns about the impact of the increasing use of small numbers of stallions in the breeding population on inbreeding and population viability. While the census population size of the Thoroughbred is large, a more appropriate method to assess population health is to estimate the effective population size (N_e), an estimate of the size of an idealised population that is representative of the genetic diversity in the actual population[20–22]. To better understand the impact of inbreeding on diversity in the population we estimated N_e for the global population and regional populations using the subset of horses born 2013–2017 ($n = 3{,}341$) to represent the current breeding population. The global population had $N_e = 330$ and among regional cohorts N_e was highest in NAM ($N_e = 226$) and lowest in SAF ($N_e = 93$); also ANZ ($N_e = 197$), EUR ($N_e = 198$). The observation that N_e was higher when all regions were considered together indicates that there is opportunity to exploit regional variation to improve genetic diversity in the global population. This may be achieved in practice, for example, by selecting stallions that are genetically distant from mares, or more broadly by the permanent movement of stallions to regions that have a genetically diverse population of mares. A comparison of N_e with census population sizes (N_c, included here as $N_c = 100{,}000$ for NAM, EUR, ANZ and SAF and $N_c = 500{,}000$ global) indicates that the genetic variation in the population deviates substantially from expectations. N_e/N_c was $<0.2\%$ in all regional cohorts and was $<0.1\%$ when the population was considered as a whole. In most wild species N_e is 10–50% of the census population size and among domesticates N_e/N_c has a median of 3%[23]. A limitation to interpretation of the results from this study is the potential bias in the samples used, many of which came from the same breeding farms. Also, in some cases, the sample sizes are relatively small (e.g. SAF) and may not be representative of the entire population. Nonetheless, these results highlight the need for a systematic evaluation of genetic diversity that may be applied for longitudinal monitoring of the Thoroughbred population.

Selection for the thoroughbred phenotype. As well as the effect of close relationships among breeding individuals, inbreeding is a consequence of selection for favourable traits. Here, to systematically identify genes that have been targets of selection for the Thoroughbred athletic phenotype, and may be most impacted by inbreeding, we used the composite selection signal (CSS) approach, a weighted measure that combines the signals from identification of highly differentiated loci (F_{ST}), increase in selected allele frequency (ΔSAF) and cross-population extended haplotype homozygosity (*XP-EHH*) tests[24,25], to analyse 48,896 pan-genomic SNPs genotyped in elite Thoroughbreds ($n = 110$) and non-Thoroughbreds ($n = 84$) representing putative founder populations. We identified 15 significant candidate selected genomic regions (S2 Table, S17 Figure), defined as clusters of ≥5 SNPs among the top 1% of the smoothed CSS statistic result ($-\log_{10}P$), seven of which overlapped with previously reported genomic regions with evidence for selection in Thoroughbreds[26,27] (Table 1). As a high prevalence of runs of homozygosity (ROH) can inform on selection, the top 1,000 SNPs ranked by the percentage of individuals with a certain SNP located within a ROH were extracted for ROH >1 Mb and >5 Mb (S3-S4 Tables, S18 Figure). Among the CSS regions, there was substantial overlap with regions with a high prevalence of ROH.

We interrogated the CSS peaks as well as flanking regions (±0.5 Mb) for candidate genes that may contribute to the Thoroughbred phenotype and identified genes with functions in behaviour, musculoskeletal, cardiac and respiratory function, conformation and metabolism (Table 1). Given that positive selection at a specific genomic locus tends to reduce ('sweep') variation across a larger region, it can be difficult to identify the gene under selection. Here, we identified plausible candidate genes driving selection based on the location of the highest-ranking SNPs in the selected region and by reviewing known biological functions of genes that we hypothesise may be selected for the Thoroughbred phenotype. Seven of the top 10 ranked SNPs were located in the top ranked region on ECA1 and the top three SNPs spanned 95 kb closest to the ZW10 interacting kinetochore protein gene (*ZWINT*). ZWINT (also known as SIP30) is abundantly expressed in the brain, modulates neurotransmitter release and functions in the mediation of peripheral nerve injury-induced neuropathic pain[28–30]. In rodents ZWINT influences pain-evoked emotional responses[31] and since human athletes have been reported to have a greater ability to tolerate pain than normally active controls[32], it is intriguing to speculate ZWINT may be involved in the equine response to exercise-induced pain. The region containing *ZWINT* was also identified as having reduced genetic diversity in Japanese Thoroughbreds compared to other breeds[27]. There were particularly long ROH in this region in the current study; 31 of the 305 stallions had ROH >16 Mb between ECA1: 37.8–76.3 Mb (S1 Table); the longest ROH spanned almost 40 Mb.

Chr	Region (Mb)	Top 1% SNPs (n)	Top CSS value	Cluster genes (n)	Candidate genes	Gene function	ROH overlap
1	38.24–45.47*	87	5.39	29	*ZWINT*	neuropathic pain response[28–30]	ROH > 1 Mb ROH > 5 Mb
17	20.69–23.86*	31	5.11	22	*DLEU7*	growth traits[51]; overgrowth[52]	ROH > 1 Mb
					KCNRG	aortic root diameter[92]	
1	121.51–122.22*	15	3.99	2	*THSD4*	lung function[93]	—
4	18.95–19.85	14	3.79	3	*VWC2*	bone formation[53]	ROH > 1 Mb ROH > 5 Mb
30	24.6–25.26	18	3.49	2	*ASPM*	cerebral cortex size[36]	—
18	47.23–47.73	6	3.27	1	*XIRP2*	cardiac hypertrophy[94]	ROH > 1 Mb ROH > 5 Mb
1	21.94–22.78	10	3.15	0	*XPNPEP1*	response to stress[38]; behavioural hyperactivity[39]	—
					SORCS1	insulin metabolism[95]	—
1	178.8–179.86	16	3.088555	1	*FKBP25*	protective response to ischemic injury[96]	ROH > 1 Mb
21	54.66–55.46	15	2.94	1	*IRX1*	hip geometry[49]; bone mineral density[46]	ROH > 1 Mb
					ADAMTS16	high altitude adaptation[47]	—
6	21.92–23.58	26	2.96	7	*COL6A3*	collagen/tendon[57,58]	ROH > 1 Mb
14	41.78–42.44*	11	2.601678	1	*FSTL4*	extinction of inhibitory avoidance memory; regulates BDNF[97–99]	—
14	46.78–47.42*	14	2.541632	3	*MEGF10*	muscle cell proliferation; regulation of myogenesis[55,56]	—
7	68.3–68.6*	7	2.54	3	*DGAT2*	lipid metabolism in adipocyte browning[63–67]; Charcot-Marie Tooth disease[62]	—
					WNT11	neural crest development; embryonic cardiac development[100]	—
2	104.13–104.32	8	2.46	1	eca-mir-147b	vascular smooth muscle cell proliferation and migration[101]	—
22	29.61–29.86*	7	2.43	0			—

Table 1. Selected genomic regions in Thoroughbreds containing ≥ 5 SNPs among the top 1% (480) SNPs ranked by the composite selection signal (CSS) value. Chr = ECA; Region = EqCab2.0. *Regions previously identified as being under selection in Thoroughbreds[26,27].

We identified other neurological/behaviour associated genes in the selected regions including the follistatin like protein 4 gene (*FSTL4*), the abnormal spindle microtubule assembly gene (*ASPM*) and the X-prolyl aminopeptidase 1 gene (*XPNPEP1*). Knockdown of *FSTL4* in mice results in the extinction of inhibitory avoidance memory indicating its involvement in synaptic plasticity and memory formation. Interestingly it functions by directly interacting with the exercise-induced brain derived neurotrophic factor (BDNF)[33,34] which decreases in response to chronic stress[35]. *ASPM* is a major determinant of the size of the cerebral cortex in mammals[36,37] that plays a key role in memory, attention, perception and awareness. Knockdown of *XPNPEP1*, which encodes an important downstream regulator of the stress response[38], in mice results in enhanced locomotor activity and impaired contextual fear memory[39]. An emerging theme in equine transcriptomics and genomics research suggests a link between the exercise phenotype and behavioural plasticity. For example, in the skeletal muscle transcriptome response to exercise training, neurological processes were the most significantly over-represented gene ontology (GO) terms, with the top three ranked GO terms being *Neurological system process* ($P = 4.85 \times 10^{-27}$), *Cognition* ($P = 1.92 \times 10^{-22}$) and *Sensory perception* ($P = 4.21 \times 10^{-21}$)[40]. Furthermore, in genome-wide association (GWA) studies genes involved in behavioural plasticity are the most strongly associated with economically important traits in racing Thoroughbreds: precocity (early adaptation to racing)[41] and the likelihood of racing[19]. For Thoroughbred horses, behavioural plasticity enables adaptation to the rigours of an intense exercise training programme in an unnatural environment, with considerable variation in the abilities of horses raised in the same environment to adapt to stress. The presence of these genes in genomic regions under selection in the Thoroughbred population supports human-mediated adaptation of the Thoroughbred towards heightened awareness and the ability to learn and adapt to stress.

Genes for aesthetic physical phenotypes including height, stature, coat and plumage colour are commonly encountered in genomic regions under selection in domestic animal populations since they are easily identifiable[42–45]. Conformation characteristics are among the most discernible traits in horses and are the principal observable traits on which Thoroughbreds are selected. Here, the selection signal on ECA21 centred on the iroquois homeobox 1 gene (*IRX1*) and the ADAM metallopeptidase with thrombospondin type 1 motif 16 gene (*ADAMTS16*), a locus that has been shown to be associated with hip geometry in humans[46]. While *ADAMTS16* has been identified among selection signals for high altitude adaptation in pigs[47], *IRX1* is associated with bone mineral density in humans[46] and influences chondrocyte differentiation and may therefore contribute to joint flexibility[48]. In horses, genes with functions in bone mineral density have been associated with measurements of joint angles[49]. Since the angle of the pelvis is a major determinant of physical conformation in Thoroughbreds and is associated with injury and performance[50], we hypothesise that the selection signature on ECA21 reflects the evolution of the Thoroughbred conformation phenotype.

Other genes relating to musculoskeletal form and function were identified among selected regions. For example, the deleted in lymphocytic leukemia 7 gene (*DLEU7*) is associated with growth traits in chickens[51] and overgrowth in humans[52] and may therefore contribute to stature in Thoroughbreds. In bone physiology, the von Willebrand factor C domain containing 2 gene (*VWC2*), a bone morphogenic protein, promotes bone formation, regeneration and healing[53,54]. In the context of Thoroughbred musculature, the multiple EGF like domains 10 gene (*MEGF10*) controls muscle cell proliferation and is involved in the regulation of myogenesis[55,56] and the collagen type VI alpha 3 chain gene (*COL6A3*) plays a major role in the maintenance of strength of muscle and connective tissue[57]. *COL6A3* is one of three genes encoding components of collagen VI, which in the horse is expressed in developing cartilage[58]. Also, collagen VI disruption in horses has been associated with osteochondrosis, a common developmental orthopaedic disease with a major economic impact in the Thoroughbred industry[59]. *COL6A3* may also be relevant to the muscle metabolism phenotype since its expression in adipocytes is associated with insulin resistance and obesity[60,61]. Mutations in the diacylglycerol O-acyltransferase 2 gene (*DGAT2*) mutations cause Charcot-Marie Tooth disease in humans[62] and the DGAT2 protein also functions in lipid metabolism[63–67].

Polygenic adaptation in the Thoroughbred. There are likely numerous endophenotypes on which selection acts to generate the athletic phenotype. Adaptation driven by selective sweeps at a number of key genomic loci is likely to be important; however, modest allele frequency changes at multiple loci are also expected to occur as a consequence of polygenic adaptation[68,69]. Therefore, in addition to 'core' genes that are critical to the phenotypic outcome, highly granular additive genetic variation—essentially encompassing the entire genome—combined with epistatic interactions differing across cell types, may largely shape the phenotype[70]. Considering this, we performed an enrichment analysis to identify functional processes over-represented among the set of 387 of the 462 genes contained within the putative selection regions that mapped to the Ingenuity Pathway Analysis (IPA) database. The top canonical pathway was *Airway inflammation in asthma* (S5 Table). Manual curation of the top 50 canonical pathways points to a process of polygenic adaptation among functional modules in cardiovascular signalling, cellular growth, proliferation and development, cellular immune response, cellular stress and injury, metabolic pathways (fatty acid and lipid degradation/biosynthesis), neurotransmitters and other nervous system signalling and organismal growth and development (S6 Table). Our results support the development of the Thoroughbred phenotype via a contribution from major allele frequency shifts at 'core' genes contributing to behavioural, metabolic and conformation traits and genome-wide changes in functional modules that shape a range of exercise-relevant physiological adaptations.

Concluding remarks. We report here a highly significant increase in inbreeding in the global Thoroughbred population during the last five decades, which is unlikely to be halted due to current breeding practices. Inbreeding results in mutational load in populations that may negatively impact on population viability. 'Genetic rescue' of highly inbred populations may be possible by the introduction of genetically diverse individuals[71]; however, rescuing genetic diversity in the Thoroughbred will be challenging due to the limitations of a closed stud book. Furthermore, the population has a small effective population size (N_e) and a limited numbers of stallions have had a disproportionate influence on the genetic composition of the Thoroughbred; 97% of pedigrees of the horses included here feature the ancestral sire, *Northern Dancer* (1961) and 35% and 55% of pedigrees in EUR and ANZ contain *Sadler's Wells* (1981) and *Danehill* (1986), respectively (S1 Text).

Pedigree data can be useful to illustrate broad trends in breeding practices, but since pedigree-based estimates of inbreeding and relatedness have poor correlations with estimates using genomic methods[27,72–75] relying on pedigree alone for outcrossing is likely to be inefficient in reversing the trends observed here (S1 Text, S19 Figure). Directives to prevent over-production from popular sire lines and the global movement of stallions that are distinct from the local population of mares may act to maintain and increase genetic diversity in the population. However, given the limited diversity in current Thoroughbred pedigrees, genomics-based measures using high-density genome-wide SNP information and a large reference population are likely to offer the best opportunity to slow and reverse the potential effects of inbreeding. The introduction of an industry-mediated longitudinal programme of genomics-based monitoring of inbreeding and the implementation of guidelines and strategies for genome-enabled breeding that are comparable to methods used in other domestic species, will contribute to promoting economic gain and safeguarding the future of the breed.

Methods

Ethics statement. Samples from animals used in this study were collected by owners and submitted to Plusvtial Ltd. for commercial genetic testing. Consent for use of samples in research was obtained during the sample submission process and methods were carried out in accordance with the agreement. No experimental procedure was performed on live animals.

Sampling and population assignment. The following data was collected for $n = 10{,}118$ Thoroughbred (TB) horses: horse name, sire, dam, sex, year of birth and country of birth (S7 Table). Based on country of birth, horses were assigned to the following geographic regions: Europe/Middle East (EUR), Australasia (ANZ), North America (NAM), South Africa (SA) and Japan (JAP). An additional set of $n = 84$ horses from four breeds that were chosen to represent putative TB founder populations included $n = 20$ Akhal Teke (AKTK), $n = 24$ Arabian

Dataset	Analysis	n samples SNP50	n samples SNP670	n samples SNP70	Reconstructed Genotypes	n SNPs post QC and imputation	n SNPs post pruning
Thoroughbred and Founders	Founder PCA and CSS analysis	73	79	42	0	31,722	Not Applicable
Unpruned Thoroughbred Set	Inbreeding ROH	356	6109	3569	84	48,896	Not Applicable
Pruned Thoroughbred Set	PCA and Inbreeding (FIS)	356	6109	3569	84	48,896	9,212
Pruned Sire Set (Subset of pruned Thoroughbred Set)	PCA and population structure	53	75	93	84	48,896	9,212

Table 2. Summary of data sets used for each analysis. Horses were genotyped using the Illumina EquineSNP50 BeadChip (SNP50), the Illumina EquineSNP70 BeadChip (SNP70) or the Affymetrix Axiom™ Equine 670 K SNP genotyping array (SNP670).

(ARR), $n = 23$ Moroccan Barb (MOR) and $n = 17$ Connemara pony (CMP). The ARR and AKTK sample data used were obtained from publicly available equine Illumina SNP50 Beadchip genotype data[3].

Assembly of comparative SNP data set, quality control and filtering of SNPs. DNA was isolated from blood or hair samples and genotyped using the Illumina EquineSNP50 BeadChip (SNP50), the Illumina EquineSNP70 BeadChip (SNP70) or the Affymetrix Axiom™ Equine 670 K SNP genotyping array (SNP670). Only animals and SNPs with a genotyping rate >95% were included with a minor allele frequency (MAF) threshold >0.05 applied. A set of 48,896 autosomal SNPs derived originally from the SNP50 and SNP70 arrays was used for the analysis. This SNP set was extracted from the genotype data from each of the three platforms. 1,821 SNPs were not present on the SNP670 array. SNPs that failed quality control in <5% of samples or were not present on one of the array platforms were imputed using the software program BEAGLE (version 3.3.2)[76,77]. For 10 horses genotyped using both the SNP50 and SNP70 arrays and 10 different horses separately genotyped using the SNP70 and SNP670 array post-imputation concordance was greater than 99%.The TB dataset ($n = 10{,}118$) comprising of the set of 48,896 SNPs was pruned using the PLINK software suite V1.9 (http://pngu.mgh.harvard.edu/purcell/plink/)[78] -indep function with the following parameters: a five-step sliding window size of 50 with a VIF threshold of 50 where the VIF is $1/(1 - R^2)$. This pruned set of 9,212 SNPs was used for principal component analysis (PCA) within the Thoroughbred population and for the calculation of individual inbreeding coefficients (F) as outlined below.

The Thoroughbred population was compared to founder populations using PCA and composite selection signature (CSS) analysis. We randomly selected $n = 229$ elite horses (CPI > 5, *i.e.* earned more than five times the average; CPI is a class performance index defined by the American Jockey Club) from the TB dataset ($n = 10{,}118$). Genomic relationships among all horses, Thoroughbreds and non-Thoroughbreds, were estimated using autosomal identity by descent (pi-hat) values in PLINK v1.9[79]. After removing individual horses with pi-hat > 0.25, 150 horses ($n = 23$ MOR, $n = 17$ CMP, and $n = 110$ TBE) remained. Then, the newly genotyped data was merged with publicly available data for AKTK ($n = 20$) and ARR ($n = 24$). Each population was pre-processed separately to include only individuals and SNPs with a genotyping rate > 95% and with MAF > 0.05. Following breed specific quality control a final round of quality control (MAF > 0.01 and genotyping rate >0.95) was applied to the combined data set of 194 horses. This resulted in 31,722 SNPs remaining for the mixed breed PCA and CSS analyses. The datasets created are summarised in Table 2. The sex, year of birth and region of origin of the Thoroughbred samples are provided in S7 Table.

Stallion genotype reconstruction. To increase the representation of prominent stallions genotypes were reconstructed for 127 horses. Of these 43 had previously been genotyped and were used to assess the accuracy of the genotype reconstruction (S1 Text). The genotypes (~46,000 autosomal SNPs) were reconstructed for horses where genotypes were available for $n \geq 20$ progeny of an individual stallion (S8 Table). An adaptation of the method described by Gomez-Raya[80] was used to infer genotypes. The main difference here was that population genotype frequencies were included allowing accurate reconstruction with just 20 offspring. If all offspring do not share one allele at a locus the sire must be heterozygous at the locus. If all offspring do share one allele at a locus the sire may be either heterozygous at the locus or homozygous for the shared allele. However, the probability of the sire being homozygous or heterozygous can be calculated based on the proportion of the different genotypes in the offspring and the observed proportion of each genotype in the population. The chi-square statistic was calculated using observed and expected offspring allele frequencies for each of the three possible sire genotypes. In order to compare the three sire genotype possibilities and assign a relative probability to each, the chi-square test statistics were first converted to likelihood scores. To do this the density of a chi-square distribution was taken at each of the three points (the density at each point represents the likelihood of the sire having that genotype, given the observed offspring genotypes). The density values were then divided by the sum of the three densities to normalize. Normalizing these three likelihoods by their sum gave the relative likelihood of each sire genotype and the genotype with the maximum likelihood was assigned as the sire genotype.

Principal component analysis (PCA). PCA was conducted using smartPCA from the EIGENSOFT package (version 4.2)[81].

The option outliersigmathresh is used to identify samples which exceed a defined number of standard deviations along one of the top principal components and classify as outliers. The threshold for this was increased from the default threshold of six to a threshold of 10 to ensure that samples from the Founder populations were not mis-classified as outliers. All other parameters were set to default values. Principal components (PCs) were plotted with data points colour-coded based on each horse's geographic region of origin. A series of analyses were performed to investigate Thoroughbred population sub-structure using the Thoroughbred and Founders, the Pruned Thoroughbred and the Pruned Sire set as described above and summarised in Table 2.

Inbreeding estimates. Using the Pruned Thoroughbred Set individual inbreeding coefficients (F_{IS}) for each horse were calculated in PLINK based upon reduction in heterozygosity relative to Hardy-Weinberg expectation (thehet option in PLINK). Genomic inbreeding was also evaluated in the Unpruned Thoroughbred Set by identifying runs of homozygosity (ROH) with the–homozyg command. ROH were defined as tracts of homozygous genotypes that were >1 Mb in length identified for one SNP per 1000 kb on average and two consecutive SNPs <1000 kb apart. No more than two missing genotypes and one heterozygous genotype were allowed. The following parameters were set:–homozyg;–homozyg-kb 1000;–homozyg-snp 30;–homozyg-gap 1000;–homozyg-window-het 1;–homozyg-window-snp 30;–homozyg-density 1000;–homozyg-window-missing 2 to identify runs of homozygosity spanning at least 1 Mb. The threshold for the number of SNP per RoH was set at 30 in order to identify shorter RoH of 1–2 Mb. As over 80% of 50 SNP windows span a region of greater than 2 mb these shorter RoH could not be detected using a minimum number of 50 SNPs per RoH. The analysis was also run with the parameter homozyg-kb increased to 5000 to identify runs of homozygosity of 5 Mb or greater reflective of more recent inbreeding[82] (Keller 2011). First, the individual sum of total ROH per animal was calculated. The F_{ROH} statistic proposed by McQuillan *et al.*[83] was then calculated, whereby the total length of ROH covering an individual animal's genome (L_{ROH}) is divided by the length of the autosomal genome (L_{AUTO}); $F_{ROH} = L_{ROH}/L_{AUTO}$. Here, consistent with other equine studies[84,85] we used the length of the equine autosomal genome (assembly EquCab 2)[86] as 2,242,960 kb (www.ncbi.nlm.nih.gov/genome/145?genome_assembly_id=22878).

The annual mean, SE, SD and CI were calculated based on the horse's year of birth for both measures of inbreeding (S16-S17 Table). Following a preliminary analysis, the pre-1996 samples were excluded from further analysis as sample distribution pre-1996 was low. The post-1995 data is of the most interest as this coincides with the introduction of "big books" for stallions; i.e. large numbers of mares bred. The shuttling of stallions to Australia from Europe for dual hemisphere breeding seasons peaked in 2001. A summary of the year of birth, region of birth and sex of the horses used in these analyses is provided in S7 Table. A pair-wise Student's t-Test was used to compare measures of inbreeding across the main geographic regions represented in the data set; Europe/Middle East (EUR), Australasia (ANZ) and North America (NAM).

A linear model was used to assess the relationship between year of birth and inbreeding for the global population and for each of the geographic regions. Mean annual inbreeding values were used in the regression model (S9 Table, S10 Table). Pearson correlation was run to assess the relationship between inbreeding and racing performance defined by handicap (Timeform) rating. Significance was calculated by testing the null hypothesis of no linear correlation (r = 0).

Effective population size (N_e) estimates. The pruned set of 9,212 SNPs was used for the calculation of effective population size (N_e). To estimate N_e, plink-formatted data was first converted to GENEPOP format using PGDSpider 2.1.1.5[87]. Then the LD method in NeEstimator2x[88] was used to calculate N_e using the converted GENEPOP data as input. N_e was calculated for global thoroughbreds and individual regions (ANZ, EUR, NAM and SAF). Year of birth for the horses used for N_e calculation were restricted to 2013–2017.

Composite selection signals (CSS). The composite selection signals (CSS) method following the procedure of Randhawa *et al.*[24,25] was used to identify signatures of selection in elite Thoroughbred horses ($n = 110$) using the Thoroughbred and Founders data set summarised in table two. A mixed set of non-Thoroughbred horses representing putative founder populations (MOR, $n = 23$; CMP, $n = 17$; ARR, $n = 24$; AKTK, $n = 20$) as the comparator population. The Elite Thoroughbred population was assigned as the population under selection and the non-Thoroughbreds were assigned as the reference population.

The CSS approach was developed to investigate genomic signatures of selection and has been successful at localizing genes for monogenic and polygenic traits under selection in livestock[24,25,89]. The CSS uses fractional ranks of constituent tests and does not incorporate the statistics with *P* values, allowing a combination of the evidence of historical selection from different selection tests. For the present study, the CSS combined the fixation index (F_{ST}), the change in selected allele frequency (ΔSAF) and the cross-population extended haplotype homozygosity (*XP-EHH*) tests into one composite statistic for each SNP. F_{ST} statistics were computed as the differentiation index between the population/s of interest (*i.e.* selected) and the contrasting/reference population/s (*i.e.* non-selected). *XP-EHH* and ΔSAF statistics were computed for the selected population(s) against the reference population. The CSS were computed as follows:

(1) For each constituent method, test statistics were ranked (1,…, n) genome-wide on n SNPs.
(2) Ranks were converted to fractional ranks (r′) (between 0 and 1) by $1/(n + 1)$ through $n/(n + 1)$.
(3) Fractional ranks were converted to z-values as $z = \Phi-1(r')$ where $\Phi-1(\cdot)$ is the inverse normal cumulative distribution function (CDF).
(4) Mean z scores were calculated by averaging z-values across all constituent tests at each SNP position and *P*-values were directly obtained from the distribution of means from a normal N $(0, m^{-1})$ distribution where m is the number of constituent test statistics.
(5) Logarithmic ($-\log_{10}$ of *P*-values) of the mean z-values were declared as CSS and were plotted against the genomic positions to identify the significant selection signals.
(6) To reduce spurious signals, the individual test statistics were averaged (smoothed) over SNPs across chromosomes within 1 Mb sliding windows.

According to the approach proposed by Randhawa *et al.*[24], significant genomic regions were defined as those that harbour at least one significant SNP (top 0.1%) surrounded by at least five SNPs among the top 1%. Here, we relaxed the stringency to define significance as regions harbouring at least five SNPs among the top 1% since the numbers of regions would otherwise be small (i.e. ~48 SNPs). Also, since linkage disequilibrium extends up to 0.4 Mb in the Thoroughbred[90], we considered 1 Mb sliding windows reasonable in this population. Therefore, SNPs among the top 1% smoothed CSS values within the sliding windows were considered significant. Genes underlying the selection peaks (with at least one top 1%) as well as flanking regions (±0.5 Mb) were identified by mapping to an annotated protein coding gene list downloaded from NCBI (accessed: 2018-05-17). These genes were then examined for evidence of functional significance.

Gene enrichment. Ingenuity® Pathway Analysis (IPA®: Qiagen, Redwood City, CA, USA; release date June 2017) was used to perform an enrichment analysis to identify functional processes over-represented among the gene lists. Settings were such that the reference set was Ingenuity Knowledge Base (genes only); relationships to include were direct and indirect; interaction networks did not include endogenous chemicals. All other settings were default. *P* is reported as -$\log_{10}$ of the adjusted *P* value obtained with the Benjamini-Hochberg procedure[91]. Ratio denotes the ratio of genes specific to the pathway identified among selected regions in this study divided by the total number of the genes in this pathway designated by the Ingenuity Knowledge Base. Manual curation to group biological functions/pathways was performed using pathway information from https://targetexplorer.ingenuity.com.

Received: 9 October 2019; Accepted: 30 December 2019;
Published online: 16 January 2020

References

1. Darwin, C. The Variation of Animals and Plants under Domestication. (John Murray, 1868).
2. Bower, M. A. *et al.* The cosmopolitan maternal heritage of the Thoroughbred racehorse breed shows a significant contribution from British and Irish native mares. *Biol Lett* **7**, 316–320, https://doi.org/10.1098/rsbl.2010.0800 (2011).
3. Petersen, J. L. *et al.* Genetic diversity in the modern horse illustrated from genome-wide SNP data. *PLoS One* **8**, e54997, https://doi.org/10.1371/journal.pone.0054997 (2013).
4. Todd, E. T. *et al.* Founder-specific inbreeding depression affects racing performance in Thoroughbred horses. *Sci Rep* **8**, 6167, https://doi.org/10.1038/s41598-018-24663-x (2018).
5. Willett, P. *Makers of the modern thoroughbred*. (University Press of Kentucky, 1986).
6. Willett, P. *An introduction to the thoroughbred*. Revised edn, (Paul, 1975).
7. Cunningham, E. P., Dooley, J. J., Splan, R. K. & Bradley, D. G. Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Anim Genet* **32**, 360–364, https://doi.org/10.1046/j.1365-2052.2001.00785.x (2001).
8. Binns, M. M. *et al.* Inbreeding in the Thoroughbred horse. *Anim Genet* **43**, 340–342, https://doi.org/10.1111/j.1365-2052.2011.02259.x (2012).
9. Kakoi, H., Kikuchi, M., Tozaki, T., Hirota, K. I. & Nagata, S. I. Evaluation of recent changes in genetic variability in Japanese thoroughbred population based on a short tandem repeat parentage panel. *Anim Sci J* **90**, 151–157, https://doi.org/10.1111/asj.13143 (2019).
10. Gaffney, B. & Cunningham, E. P. Estimation of genetic trend in racing performance of thoroughbred horses. *Nature* **332**, 722–724, https://doi.org/10.1038/332722a0 (1988).
11. Joshi, P. K. *et al.* Directional dominance on stature and cognition in diverse human populations. *Nature* **523**, 459–462, https://doi.org/10.1038/nature14618 (2015).
12. Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* **2**, e64, https://doi.org/10.1371/journal.pgen.0020064 (2006).
13. Corbin, L. J. *et al.* The utility of low-density genotyping for imputation in the Thoroughbred horse. *Genet Sel Evol* **46**, 9, https://doi.org/10.1186/1297-9686-46-9 (2014).
14. Zhang, Q., Guldbrandtsen, B., Bosse, M., Lund, M. S. & Sahana, G. Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC Genomics* **16**, 542, https://doi.org/10.1186/s12864-015-1715-x (2015).
15. Marsden, C. D. *et al.* Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences* **113**, 152–157 (2016).
16. Lewis, T. W., Abhayaratne, B. M. & Blott, S. C. Trends in genetic diversity for all Kennel Club registered pedigree dog breeds. *Canine Genetics and Epidemiology* **2**, 13, https://doi.org/10.1186/s40575-015-0027-4 (2015).
17. Pryce, J. E., Haile-Mariam, M., Goddard, M. E. & Hayes, B. J. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. *Genet Sel Evol* **46**, 71, https://doi.org/10.1186/s12711-014-0071-7 (2014).
18. Orlando, L. & Librado, P. Origin and Evolution of Deleterious Mutations in Horses. *Genes (Basel)* **10**, https://doi.org/10.3390/genes10090649 (2019).
19. McGivney, B. A. *et al.* A genomic prediction model for racecourse starts in the Thoroughbred horse. *Anim Genet* **50**, 347–357, https://doi.org/10.1111/age.12798 (2019).
20. Wang, J., Santiago, E. & Caballero, A. Prediction and estimation of effective population size. *Heredity (Edinb)* **117**, 193–206, https://doi.org/10.1038/hdy.2016.43 (2016).
21. Wang, J. Estimation of effective population sizes from data on genetic markers. *Philos Trans R Soc Lond B Biol Sci* **360**, 1395–1409, https://doi.org/10.1098/rstb.2005.1682 (2005).
22. Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**, 195–205, https://doi.org/10.1038/nrg2526 (2009).
23. Hall, S. J. Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. *Animal* **10**, 1778–1785, https://doi.org/10.1017/S1751731116000914 (2016).
24. Randhawa, I. A., Khatkar, M. S., Thomson, P. C. & Raadsma, H. W. Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet* **15**, 34, https://doi.org/10.1186/1471-2156-15-34 (2014).
25. Randhawa, I. A., Khatkar, M. S., Thomson, P. C. & Raadsma, H. W. Composite Selection Signals for Complex Traits Exemplified Through Bovine Stature Using Multibreed Cohorts of European and African Bos taurus. *G3 (Bethesda)* **5**, 1391–1401, https://doi.org/10.1534/g3.115.017772 (2015).
26. Petersen, J. L. *et al.* Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet* **9**, e1003211, https://doi.org/10.1371/journal.pgen.1003211 (2013).

27. Fawcett, J. A. *et al*. Genome-wide SNP analysis of Japanese Thoroughbred racehorses. *PLoS One* **14**, e0218407, https://doi.org/10.1371/journal.pone.0218407 (2019).
28. Peng, G. *et al*. SIP30 Is Regulated by ERK in Peripheral Nerve Injury-induced Neuropathic Pain. *Journal of Biological Chemistry* **284**, 30138–30147, https://doi.org/10.1074/jbc.M109.036756 (2009).
29. Zhang, Y. Q. *et al*. Role of SIP30 in the development and maintenance of peripheral nerve injury-induced neuropathic pain. *Pain* **146**, 130–140, https://doi.org/10.1016/j.pain.2009.07.011 (2009).
30. Lee, H. K., Safieddine, S., Petralia, R. S. & Wenthold, R. J. Identification of a novel SNAP25 interacting protein (SIP30). *J Neurochem* **81**, 1338–1347, https://doi.org/10.1046/j.1471-4159.2002.00937.x (2002).
31. Han, M. *et al*. SIP30 Is Required for Neuropathic Pain-Evoked Aversion in Rats. *The Journal of Neuroscience* **34**, 346–355 (2014).
32. Tesarz, J., Schuster, A. K., Hartmann, M., Gerhardt, A. & Eich, W. Pain perception in athletes compared to normally active controls: a systematic review with meta-analysis. *Pain* **153**, 1253–1262, https://doi.org/10.1016/j.pain.2012.03.005 (2012).
33. Marlatt, M. W., Potter, M. C., Lucassen, P. J. & van Praag, H. Running throughout middle-age improves memory function, hippocampal neurogenesis, and BDNF levels in female C57BL/6J mice. *Dev Neurobiol* **72**, 943–952, https://doi.org/10.1002/dneu.22009 (2012).
34. Neeper, S. A., Gomez-Pinilla, F., Choi, J. & Cotman, C. W. Physical activity increases mRNA for brain-derived neurotrophic factor and nerve growth factor in rat brain. *Brain Res* **726**, 49–56 (1996).
35. Krishnan, V. & Nestler, E. J. The molecular neurobiology of depression. *Nature* **455**, 894–902, https://doi.org/10.1038/nature07455 (2008).
36. Bond, J. *et al*. ASPM is a major determinant of cerebral cortical size. *Nature Genetics* **32**, 316–320, https://doi.org/10.1038/ng995 (2002).
37. Johnson, M. B. *et al*. Aspm knockout ferret reveals an evolutionary mechanism governing cerebral cortical size. *Nature* **556**, 370–375, https://doi.org/10.1038/s41586-018-0035-0 (2018).
38. Xu, J. *et al*. Genetic regulatory network analysis reveals that low density lipoprotein receptor-related protein 11 is involved in stress responses in mice. *Psychiatry Res* **220**, 1131–1137, https://doi.org/10.1016/j.psychres.2014.09.002 (2014).
39. Bae, Y. S. *et al*. Deficiency of aminopeptidase P1 causes behavioral hyperactivity, cognitive deficits, and hippocampal neurodegeneration. *Genes Brain Behav* **17**, 126–138, https://doi.org/10.1111/gbb.12419 (2018).
40. Bryan, K. *et al*. Equine skeletal muscle adaptations to exercise and training: evidence of differential regulation of autophagosomal and mitochondrial components. *BMC Genomics* **18**, 595, https://doi.org/10.1186/s12864-017-4007-9 (2017).
41. Farries, G. *et al*. Genetic contributions to precocity traits in racing Thoroughbreds. *Anim Genet* **49**, 193–204, https://doi.org/10.1111/age.12622 (2018).
42. Li, D. *et al*. Population genomics identifies patterns of genetic diversity and selection in chicken. *BMC Genomics* **20**, 263, https://doi.org/10.1186/s12864-019-5622-4 (2019).
43. Randhawa, I. A., Khatkar, M. S., Thomson, P. C. & Raadsma, H. W. A Meta-Assembly of Selection Signatures in Cattle. *PLoS One* **11**, e0153013, https://doi.org/10.1371/journal.pone.0153013 (2016).
44. Rochus, C. M. *et al*. Revealing the selection history of adaptive loci using genome-wide scans for selection: an example from domestic sheep. *BMC Genomics* **19**, 71, https://doi.org/10.1186/s12864-018-4447-x (2018).
45. Rubin, C. J. *et al*. Strong signatures of selection in the domestic pig genome. *Proc Natl Acad Sci USA* **109**, 19529–19536, https://doi.org/10.1073/pnas.1217149109 (2012).
46. Hsu, Y. H. *et al*. Meta-Analysis of Genomewide Association Studies Reveals Genetic Variants for Hip Bone Geometry. *J Bone Miner Res*, e3698, https://doi.org/10.1002/jbmr.3698 (2019).
47. Dong, K. *et al*. Genomic scan reveals loci under altitude adaptation in Tibetan and Dahe pigs. *PLoS One* **9**, e110520, https://doi.org/10.1371/journal.pone.0110520 (2014).
48. Askary, A. *et al*. Iroquois Proteins Promote Skeletal Joint Formation by Maintaining Chondrocytes in an Immature State. *Dev Cell* **35**, 358–365, https://doi.org/10.1016/j.devcel.2015.10.004 (2015).
49. Gmel, A. I., Druml, T., von Niederhausern, R., Leeb, T. & Neuditschko, M. Genome-Wide Association Studies Based on Equine Joint Angle Measurements Reveal New QTL Affecting the Conformation of Horses. *Genes (Basel)* **10**, https://doi.org/10.3390/genes10050370 (2019).
50. Weller, R., Pfau, T., Verheyen, K., May, S. A. & Wilson, A. M. The effect of conformation on orthopaedic health and performance in a cohort of National Hunt racehorses: preliminary results. *Equine Vet J* **38**, 622–627 (2006).
51. Abdalhag, M. A. *et al*. Single nucleotide polymorphisms associated with growth traits in Jinghai yellow chickens. *Genet Mol Res* **14**, 16169–16177, https://doi.org/10.4238/2015.December.8.6 (2015).
52. Kamien, B. *et al*. Narrowing the critical region for overgrowth within 13q14.2-q14.3 microdeletions. *European Journal of Medical Genetics* **58**, 629–633, https://doi.org/10.1016/j.ejmg.2015.10.006 (2015).
53. Almehmadi, A. *et al*. VWC2 Increases Bone Formation Through Inhibiting Activin Signaling. *Calcif Tissue Int* **103**, 663–674, https://doi.org/10.1007/s00223-018-0462-9 (2018).
54. Ohyama, Y. *et al*. Modulation of matrix mineralization by Vwc2-like protein and its novel splicing isoforms. *Biochem Biophys Res Commun* **418**, 12–16, https://doi.org/10.1016/j.bbrc.2011.12.075 (2012).
55. Saha, M. *et al*. Consequences of MEGF10 deficiency on myoblast function and Notch1 interactions. *Hum Mol Genet* **26**, 2984–3000, https://doi.org/10.1093/hmg/ddx189 (2017).
56. Holterman, C. E., Le Grand, F., Kuang, S., Seale, P. & Rudnicki, M. A. Megf10 regulates the progression of the satellite cell myogenic program. *J Cell Biol* **179**, 911–922, https://doi.org/10.1083/jcb.200709083 (2007).
57. Pan, T. C. *et al*. COL6A3 protein deficiency in mice leads to muscle and tendon defects similar to human collagen VI congenital muscular dystrophy. *J Biol Chem* **288**, 14320–14331, https://doi.org/10.1074/jbc.M112.433078 (2013).
58. Henson, F. M. D., Davies, M. E., Schofield, P. N. & Jeffcott, L. B. Expression of types II, VI and X collagen in equine growth cartilage during development. *Equine Vet J* **28**, 189–198, https://doi.org/10.1111/j.2042-3306.1996.tb03772.x (1996).
59. Henson, F. M. D., Davies, M. E. & Jeffcott, L. B. Equine dyschondroplasia (osteochondrosis)—Histological findings and type VI collagen localization. *The Veterinary Journal* **154**, 53–62, https://doi.org/10.1016/S1090-0233(05)80008-5 (1997).
60. McCulloch, L. J. *et al*. COL6A3 Is Regulated by Leptin in Human Adipose Tissue and Reduced in Obesity. *Endocrinology* **156**, 134–146, https://doi.org/10.1210/en.2014-1042 (2015).
61. Dankel, S. N. *et al*. COL6A3 expression in adipocytes associates with insulin resistance and depends on PPARγ and adipocyte size. *Obesity* **22**, 1807–1813, https://doi.org/10.1002/oby.20758 (2014).
62. Hong, Y. B. *et al*. DGAT2 Mutation in a Family with Autosomal-Dominant Early-Onset Axonal Charcot-Marie-Tooth Disease. *Hum Mutat* **37**, 473–480, https://doi.org/10.1002/humu.22959 (2016).
63. Hung, Y. H., Carreiro, A. L. & Buhman, K. K. Dgat1 and Dgat2 regulate enterocyte triacylglycerol distribution and alter proteins associated with cytoplasmic lipid droplets in response to dietary fat. *Biochim Biophys Acta Mol Cell Biol Lipids* **1862**, 600–614, https://doi.org/10.1016/j.bbalip.2017.02.014 (2017).
64. Irshad, Z., Dimitri, F., Christian, M. & Zammit, V. A. Diacylglycerol acyltransferase 2 links glucose utilization to fatty acid oxidation in the brown adipocytes. *J Lipid Res* **58**, 15–30, https://doi.org/10.1194/jlr.M068197 (2017).
65. McFie, P. J., Banman, S. L. & Stone, S. J. Diacylglycerol acyltransferase-2 contains a c-terminal sequence that interacts with lipid droplets. *Biochim Biophys Acta Mol Cell Biol Lipids* **1863**, 1068–1081, https://doi.org/10.1016/j.bbalip.2018.06.008 (2018).

66. Ning, T. *et al*. Genetic interaction of DGAT2 and FAAH in the development of human obesity. *Endocrine* **56**, 366–378, https://doi.org/10.1007/s12020-017-1261-1 (2017).
67. Zang, L. *et al*. Identification of a 13 bp indel polymorphism in the 3'-UTR of DGAT2 gene associated with backfat thickness and lean percentage in pigs. *Gene* **576**, 729–733, https://doi.org/10.1016/j.gene.2015.09.047 (2016).
68. Pritchard, J. K. & Di Rienzo, A. Adaptation - not by sweeps alone. *Nat Rev Genet* **11**, 665–667, https://doi.org/10.1038/nrg2880 (2010).
69. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* **20**, R208–215, https://doi.org/10.1016/j.cub.2009.11.055 (2010).
70. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186, https://doi.org/10.1016/j.cell.2017.05.038 (2017).
71. Stronen, A. V. *et al*. Genetic rescue of an endangered domestic animal through outcrossing with closely related breeds: A case study of the Norwegian Lundehund. *PLoS One* **12**, e0177429, https://doi.org/10.1371/journal.pone.0177429 (2017).
72. Al Abri, M. A., Konig von Borstel, U., Strecker, V. & Brooks, S. A. Application of Genomic Estimation Methods of Inbreeding and Population Structure in an Arabian Horse Herd. *J Hered* **108**, 361–368, https://doi.org/10.1093/jhered/esx025 (2017).
73. Wang, J. Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology* **107**, 4–13, https://doi.org/10.1016/j.tpb.2015.08.006 (2016).
74. Kardos, M., Luikart, G. & Allendorf, F. W. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity* **115**, 63–72, https://doi.org/10.1038/hdy.2015.17 (2015).
75. Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* **16**, 33, https://doi.org/10.1038/nrg3821, https://www.nature.com/articles/nrg3821#supplementary-information (2014).
76. Ayres, D. L. *et al*. BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Syst Biol*, https://doi.org/10.1093/sysbio/syz020 (2019).
77. Ayres, D. L. *et al*. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol* **61**, 170–173, https://doi.org/10.1093/sysbio/syr100 (2012).
78. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575, https://doi.org/10.1086/519795 (2007).
79. Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, https://doi.org/10.1186/s13742-015-0047-8 (2015).
80. Gomez-Raya, L. Inferring unknown genotypes of sires at codominant deoxyribonucleic acid markers in half-sib families1. *Journal of Animal Science* **87**, 1872–1882, https://doi.org/10.2527/jas.2008-1425 (2009).
81. Price, A. L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909, https://doi.org/10.1038/ng1847 (2006).
82. Keller, M. C., Visscher, P. M. & Goddard, M. E. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* **189**, 237–249, https://doi.org/10.1534/genetics.111.130922 (2011).
83. McQuillan, R. *et al*. Runs of homozygosity in European populations. *American Journal of Human Genetics* **83**, 359–372, https://doi.org/10.1016/j.ajhg.2008.08.007 (2008).
84. Grilz-Seger, G. *et al*. Analysis of ROH patterns in the Noriker horse breed reveals signatures of selection for coat color and body size. *Anim Genet* **50**, 334–346, https://doi.org/10.1111/age.12797 (2019).
85. Han, H. *et al*. Refinement of Global Domestic Horse Biogeography Using Historic Landrace Chinese Mongolian Populations. *J Hered* **110**, 769–781, https://doi.org/10.1093/jhered/esz032 (2019).
86. Wade, C. M. *et al*. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865–867, https://doi.org/10.1126/science.1178158 (2009).
87. Lischer, H. E. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299, https://doi.org/10.1093/bioinformatics/btr642 (2012).
88. Do, C. *et al*. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Mol Ecol Resour* **14**, 209–214, https://doi.org/10.1111/1755-0998.12157 (2014).
89. Browett, S. *et al*. Genomic Characterisation of the Indigenous Irish Kerry Cattle Breed. *Front Genet* **9**, 51, https://doi.org/10.3389/fgene.2018.00051 (2018).
90. McCue, M. E. *et al*. A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet* **8**, e1002451, https://doi.org/10.1371/journal.pgen.1002451 (2012).
91. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Method.* **57**, 289–300 (1995).
92. Wild, P. S. *et al*. Large-scale genome-wide analysis identifies genetic variants associated with cardiac structure and function. *J Clin Invest* **127**, 1798–1812, https://doi.org/10.1172/JCI84840 (2017).
93. Soler Artigas, M. *et al*. Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function. *Am J Respir Crit Care Med* **184**, 786–795, https://doi.org/10.1164/rccm.201102-0192OC (2011).
94. McCalmon, S. A. *et al*. Modulation of angiotensin II-mediated cardiac remodeling by the MEF2A target gene Xirp2. *Circ Res* **106**, 952–960, https://doi.org/10.1161/CIRCRESAHA.109.209007 (2010).
95. Kebede, M. A. *et al*. SORCS1 is necessary for normal insulin secretory granule biogenesis in metabolically stressed β cells. *The Journal of clinical investigation* **124**, 4240–4256, https://doi.org/10.1172/JCI74072 (2014).
96. Jiang, Q. *et al*. Elucidation of the FKBP25-60S Ribosomal Protein L7a Stress Response Signaling During Ischemic Injury. *Cell Physiol Biochem* **47**, 2018–2030, https://doi.org/10.1159/000491470 (2018).
97. Ross, R. E., Saladin, M. E., George, M. S. & Gregory, C. M. High-Intensity Aerobic Exercise Acutely Increases Brain-derived Neurotrophic Factor. *Med Sci Sports Exerc*, https://doi.org/10.1249/MSS.0000000000001969 (2019).
98. Etnier, J. L. *et al*. The Effects of Acute Exercise on Memory and Brain-Derived Neurotrophic Factor (BDNF). *J Sport Exerc Psychol* **38**, 331–340, https://doi.org/10.1123/jsep.2015-0335 (2016).
99. Neeper, S. A., Gomez-Pinilla, F., Choi, J. & Cotman, C. Exercise and brain neurotrophins. *Nature* **373**, 109, https://doi.org/10.1038/373109a0 (1995).
100. Ossipova, O., Kerney, R., Saint-Jeannet, J. P. & Sokol, S. Y. Regulation of neural crest development by the formin family protein Daam1. *Genesis* **56**, e23108, https://doi.org/10.1002/dvg.23108 (2018).
101. Yue, Y., Lv, W., Zhang, L. & Kang, W. MiR-147b influences vascular smooth muscle cell proliferation and migration via targeting YY1 and modulating Wnt/beta-catenin activities. *Acta Biochim Biophys Sin (Shanghai)* **50**, 905–913, https://doi.org/10.1093/abbs/gmy086 (2018).

Author contributions

B.A.M. and E.W.H. designed the research; B.A.M., H.H. and L.R.C. performed research and analysed data; T.T. provided samples; H.H., T.T., L.M.K. and D.E.M. contributed to the manuscript; B.A.M. and E.W.H. wrote the manuscript.

Competing interests

This research was conducted with the financial support of Plusvital Ltd. (www.plusvital.com) and a grant to EWH from Science Foundation Ireland (www.sfi.ie) (SFI/11/PI/1166). EWH is Chief Science Officer for Plusvital Ltd. BAM, HH and LRC are employees of Plusvital Ltd. EWH and DEM are shareholders in Plusvital Ltd. LMK and DEM are consultants of Plusvital Ltd. Other than the authors, the funders played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Plusvital Ltd. has been engaged by the American Jockey Club (The Jockey Club) to provide expert advice and services, which may relate to data/results included in the manuscript.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-57389-5.

Correspondence and requests for materials should be addressed to E.W.H.