nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-017-0453-7

A southern African origin and cryptic structure in the highly mobile plains zebra

Casper-Emil T. Pedersen[1*], Anders Albrechtsen[1], Paul D. Etter[2], Eric A. Johnson[2], Ludovic Orlando[3], Lounes Chikhi[4,5], Hans R. Siegismund[1] and Rasmus Heller[1*]

The plains zebra (*Equus quagga*) is an ecologically important species of the African savannah. It is also one of the most numerous and widely distributed ungulates, and six subspecies have been described based on morphological variation. However, the within-species evolutionary processes have been difficult to resolve due to its high mobility and a lack of consensus regarding the population structure. We obtained genome-wide DNA polymorphism data from more than 167,000 loci for 59 plains zebras from across the species range, encompassing all recognized extant subspecies, as well as three mountain zebras (*Equus zebra*) and three Grevy's zebras (*Equus grevyi*). Surprisingly, the population genetic structure does not mirror the morphology-based subspecies delineation, underlining the dangers of basing management units exclusively on morphological variation. We use demographic modelling to provide insights into the past phylogeography of the species. The results identify a southern African location as the most likely source region from which all extant populations expanded around 370,000 years ago. We show evidence for inclusion of the extinct and phenotypically divergent quagga (*Equus quagga quagga*) in the plains zebra variation and reveal that it was less divergent from the other subspecies than the northernmost (Ugandan) extant population.

The African savannah harbours an unsurpassed diversity of large mammals, with ungulates being particularly well represented[1]. Their current distribution is the likely result of past expansions from—and contractions to—refugia where populations survived during adverse climatic conditions (reviewed in ref. [2]). Inferring the geographical origin of a cornerstone species that performs important ecosystem services can provide an indirect route to locating past climatic refugia. The plains zebra is distributed throughout southern and eastern Africa (Fig. 1). Extremely mobile and able to feed on relatively low-quality forage, it represents a pioneer species in the succession of grazers on the African savannah[3,4], thus providing a good model for understanding the phylogeography of savannah-adapted grazers in general.

Our knowledge concerning the phylogeography of the species is, however, limited. A study using mitochondrial DNA and seven microsatellites found less genetic structure in this species than any other savannah ungulate and identified a trend of isolation by distance[5]. This contrasts with the historical narrative of discrete morphological variation across the geographical range, with six recognized subspecies[6]: *Equus quagga borensis*, *Equus quagga boehmi*, *Equus quagga crawshayi*, *Equus quagga chapmani*, *Equus quagga burchelli* and *Equus quagga quagga* (the 'quagga', which became extinct in the late 1800s[7]). Different sources have provided different subspecies classifications based on striping patterns and cranial morphometrics[4–9], underlining the lack of consensus regarding the underlying population structure. The lack of resolvable population structure has precluded phylogeographic and demographic inference; hence, these are largely unknown despite the importance of the species in the savannah ecosystem. In this study, we test whether the genetic structure in the plains zebra coincides with the recognized subspecies. If there are distinct populations that evolved as genetically distinct lineages, management efforts should take this into account[10,11].

It has been estimated that at least 600,000 plains zebras inhabit the African savannahs[4], yet recent extinction of populations suggests the species was even more prolific in the past. The now extinct quagga zebra (*E. q. quagga*) is one notable example. The quagga was distributed south of the plains zebra range, with a possible overlap zone north of the Orange River[12]. Its relation to the plains zebra was initially controversial. Some observers considered it a distinct species due to its distinct morphology[13]—notably, a strongly reduced striping pattern and tan base colour on the hind half of the body. The quagga's status has since been addressed in several studies based on mitochondrial DNA sequences from museum specimens[5–7,14,15]. These studies found that quagga mitochondrial DNA haplotypes were diffusely distributed within the plains zebra haplotypes and predominantly interspersed within the southern plains zebra.

This study applies genome-wide single nucleotide polymorphism (SNP) data from restriction-site-associated (RAD) DNA sequencing to identify genetic structure in the plains zebra and unravel the within-species evolutionary processes; that is, phylogeographic and demographic history. Additionally, we include the quagga in the analyses to refine our understanding of the processes leading to its divergence from other plains zebras.

Results

Genetic structure and phylogeography. A total of 65 individuals were RAD sequenced (Supplementary Table 1). The Analysis of Next Generation Sequencing Data (ANGSD) pipeline yielded genotype likelihood information from a total of 7,051,457 sites sequenced in all zebras. Polymorphism was high with 3.4% variable positions. The median depth per sample was 15, with means per individual varying between 7 and 38. The number of SNPs in the called genotype datasets ranged between 67,993 and 167,963 (Supplementary Table 2).

[1]Department of Biology, Section for Computational and RNA Biology, University of Copenhagen, Copenhagen, Denmark. [2]Institute of Molecular Biology, University of Oregon, Eugene, OR, USA. [3]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. [4]Instituto Gulbenkian de Ciência, Oeiras, Portugal. [5]Centre National de la Recherche Scientifique, Université Paul Sabatier, École Nationale de Formation Agronomique, UMR 5174 Laboratoire Évolution et Diversité Biologique, Toulouse, France. *e-mail: capedersen@bio.ku.dk; rheller@bio.ku.dk

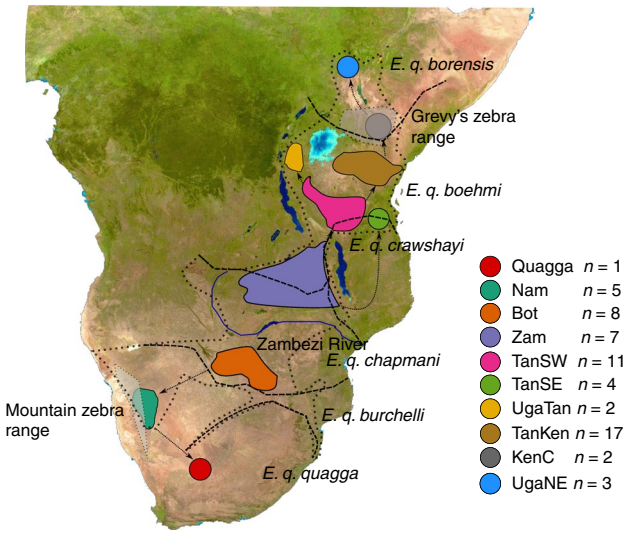

Fig. 1 | Sampling areas for the identified plains zebra populations. In the Namibian (Nam) and central Kenyan (KenC) populations, sampling was overlapping with the reported ranges of the mountain zebra and Grevy's zebra[9,65]. The exact sampling location for the Quagga museum sample is not known[20]. The current geographical spread of plains zebras (dotted outline) is adopted from ref. [4], while the morphologically determined subspecies are indicated by dashed lines[5]. Other populations are abbreviated as follows: Bot, Botswanan; TanKen, Tanzanian-Kenyan; TanSE, southeastern Tanzanian; TanSW, southwestern Tanzanian; UgaNE, northeastern Ugandan; UgaTan, Ugandan–Tanzanian; Zam, Zambian.

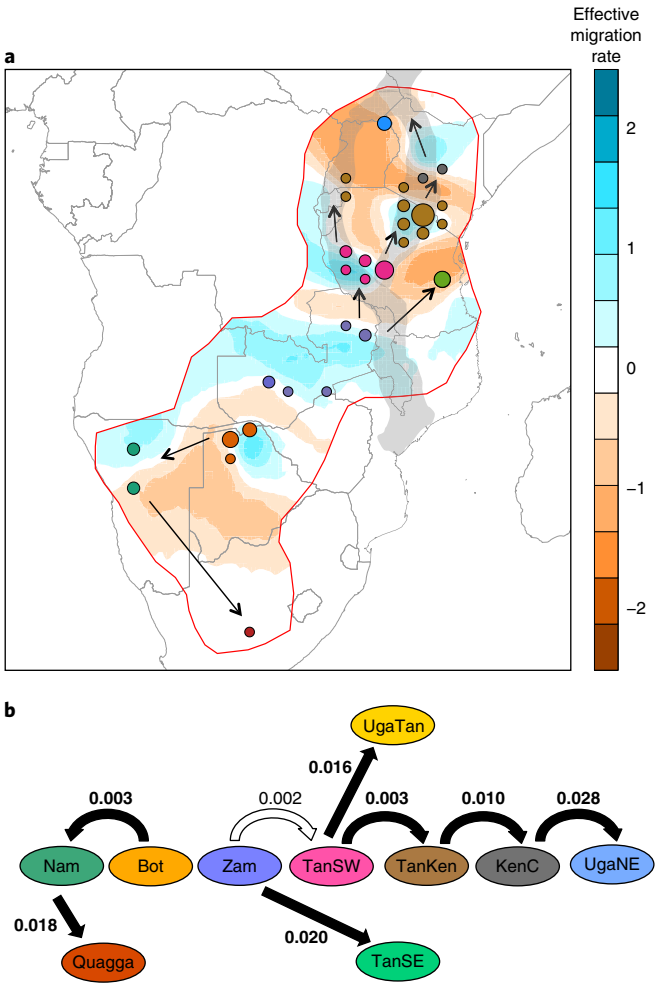

Fig. 2 | Estimating effective migration surfaces and directionality index. **a**, Analysis using dataset 2. Here, the estimated effective migration rates are shown using a gradient of colours from dark orange to dark blue. The numbers on the right denote the effective migration rate on the $\log_{10}$ scale relative to the overall migration rate across the habitat[48]. The shaded grey area roughly outlines the Rift Valley in East Africa. The colour scheme used for the samples is the same as in Fig. 1. Arrows indicate the inferred source–sink connection between populations (see **b**). **b**, ψ results from Supplementary Table 7. Here, the populations are lined along the south-west–north-east axis seen in Fig. 1. The filled black arrows indicate significant directionality between two connected populations (see bold values in Supplementary Table 7). The unfilled arrow indicates a non-significant directionality value. Significance was assessed using bootstrapped SFSs. The number above each arrow shows the ψ value for each population pair (in bold for significant values).

We identified nine extant populations based on the geographical and genetic context (Figs. 2–4; see also Supplementary Note 1 'Population identification criteria' and Supplementary Table 3). We labelled these nine populations according to their geographical location: Namibia, Botswana, Zambia, southwestern Tanzania, southeastern Tanzania, Uganda/Tanzania, Tanzania/Kenya, central Kenya and northeastern Uganda (see Fig. 1). An individual-based plot shows isolation by distance between the populations, but not within them, except for the sub-structured Tanzania/Kenya population (Supplementary Fig. 1; see discussion below). This is an indication of demic structure rather than the clinal variation indicated from less comprehensive genetic markers[5]. We found limited overlap between the inferred genetic structure and the current morphology-based subspecies designations (except in the case of the Namibian and Botswanan populations; Supplementary Figs. 2 and 3 and Supplementary Tables 1 and 4). A single individual, 3783 from Rungwa, showed aberrant clustering (for example, Figs. 3 and 4). We believe this is a likely migrant individual, as Rungwa is connected to the Tanzania/Kenya population by a known migration corridor[16]. The population structure is more pronounced in the northern part of the species range (Fig. 2), with six of the nine inferred extant populations found in Tanzania and northwards. We also found an increasing slope in genetic distance as a function of geographic distance in the northern region (Fig. 2b, Supplementary Table 5 and Supplementary Fig. 1).

The inferred root of the species tree as estimated by TreeMix was between the populations in Botswana and Zambia (Fig. 5), which suggests a cradle of all extant populations in a location close to these populations. However, low internal branch lengths made the inference of tree topology challenging. This is most likely caused by the high mobility of the plains zebra, resulting in persistent gene flow[17]. Low pairwise fixation index values (F_{ST}; Supplementary Table 5) between the geographical spine of populations formed by Botswana, Zambia, southeastern Tanzania and Tanzania/Kenya corroborated the limited genetic drift separating these populations. However, a permutation test where individuals from southeastern Tanzania and Tanzania/Kenya were randomly shuffled 100 times confirmed that even the lowest pairwise F_{ST} was significant. For the 100 permutation datasets, we found F_{ST} values ranging between 4.4×10^{-6} and 7.9×10^{-4}, with a mean value of 2.5×10^{-4}. In the marginal populations, pairwise F_{ST} values were higher and TreeMix branches were longer, supporting more genetic drift in these populations. This finding was robust to sample size and we report the pairwise F_{ST} values from random downsampling to three samples per population (Supplementary Table 5).

The directionality index values were congruent with a range expansion scenario where Botswana and Zambia are the populations

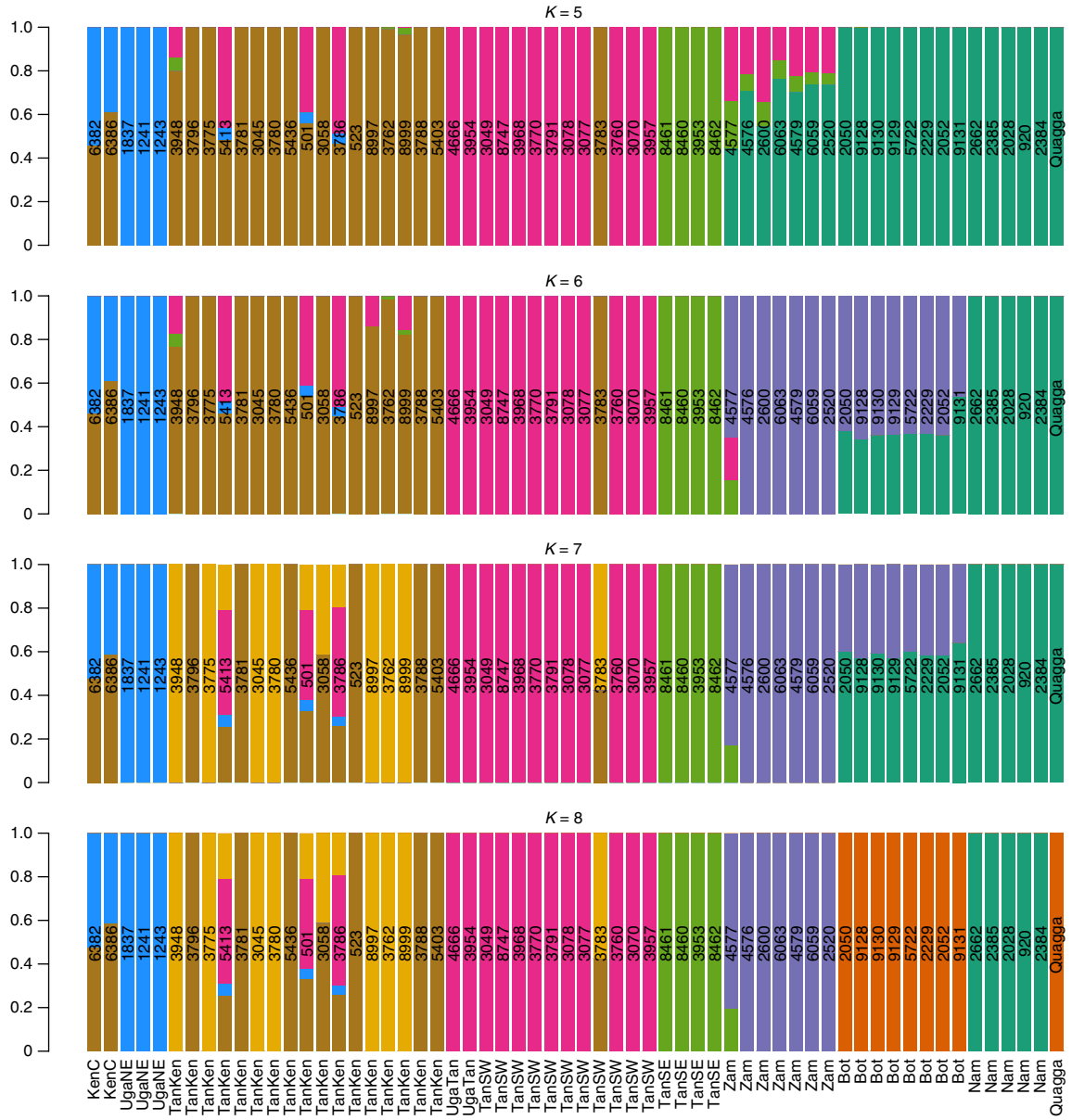

Fig. 3 | Admixture proportions using NGSadmix. Analyses were carried out under the assumption of various numbers of ancestral populations ($K = 5–8$). This analysis uses the 60 individuals in dataset 2. Columns represent individuals and are grouped into coloured clusters according to the proportion of their ancestry components. See Supplementary Fig. 19 for the same figure using subspecies labels for individuals.

closest to the origin of the expansion. The directionality index (ψ) values were consistently negative when Botswana or Zambia was population 1 (Fig. 2b). ψ was low for many population pairs (particularly between Botswana and Zambia), but bootstrapped two-dimensional site frequency spectra (SFSs) with ancestry correction confirmed that the ψ values were significantly different from zero in all pairs with a source–sink relationship according to Fig. 2b, except between Botswana and Zambia and Zambia and southeastern Tanzania (see Fig. 2b). Large negative values in the direction of northeastern Uganda were found for all pairwise comparisons, confirming high levels of drift in this population. Note that for all pairs, we used a downsampled SFS from each population to avoid possible bias due to sample size variability.

The population in southeastern Tanzania was clearly differentiated from the adjacent populations with no indications of admixture (Figs. 2–4 and Supplementary Fig. 1), suggesting a strong barrier to gene flow in eastern–central Tanzania. The more basal position of southeastern Tanzania in the population tree (Fig. 5) shows that this population derives from a dispersal event originating in Zambia that is different from that leading to the other East African populations. Similarly, the Uganda/Tanzania population formed a distinct cluster that diverged from southeastern Tanzania independent from the Tanzania/Kenya divergence (Fig. 5).

Inbreeding was low overall (Supplementary Figs. 4–9) with the majority of individuals being outbred ($F < 0.05$). The three samples that were moderately inbred ($F = 0.05–0.15$) were from localities with a high human impact: one from Kidepo National Park (northeastern Uganda), which has recently been severely impacted by human conflict (and equine disease), one from the fenced Nakuru National Park (central Kenya) and one from the Burigi game reserve, which has been heavily exploited by refugees from the Rwandan conflict. The two most inbred individuals had inbreeding coefficients of $F \sim 0.25$, which is at the level of sibship mating. They came from the Ovita Farm in Namibia and the Maun Educational

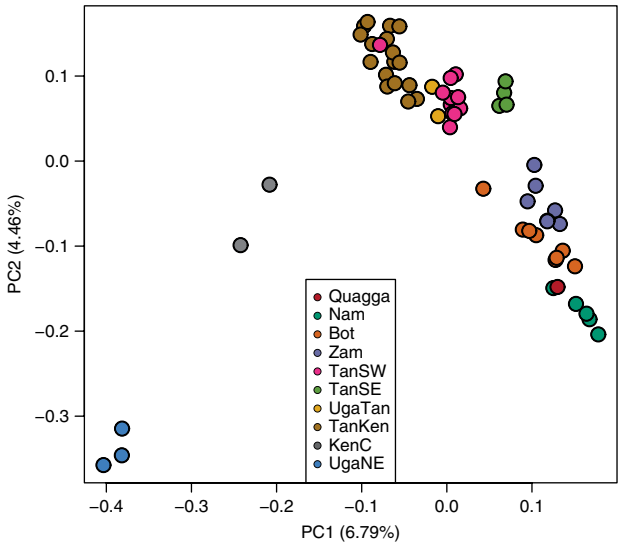

Fig. 4 | Principal component analysis plot showing clustering in the plains zebras using the 60 individuals in dataset 2. The axis labels show the variation explained by the two principal components (PC1 and PC2). The colour scheme is the same as in Fig. 1. Clusters were identified and matched with genetic clusters found in other analyses (see Supplementary Information section 'Population identification criteria'). See Supplementary Fig. 20 for a principal component analysis using subspecies labels for individuals.

Park in Botswana, respectively. Both are semi-natural populations recently established by humans.

Interspecific admixture. The inclusion of the extinct quagga in these analyses confirmed its status as the southernmost variant of plains zebra, rather than a distinct species[7,15]. The quagga individual was by most measures less divergent from its neighbouring plains zebra population than northeastern Uganda—the population at the opposite extreme of the range (Figs. 4 and 5). Analyses showed that the quagga was genetically closest to Namibia (Figs. 2 and 4). Using a method based on the number of derived mutations[18,19] between the quagga and Namibia, we estimated that the two populations had a coalescence time (T_{MRCA}) of 655–718 thousand years ago (ka). This estimate was close to that of Tanzania/Kenya and Botswana, as well as Namibia and Botswana (639–769 ka and 600–722 ka, respectively), suggesting that the T_{MRCA} of the quagga and plains zebra is the same as that between plains individuals from different populations.

In accordance with a whole-genome study using one sample per species, we found a signal of gene flow between the mountain zebra and plains zebra[20]. The *D*-statistic ((plains; Grevy's); mountain); horse) was significantly negative for the four southern plains zebra populations (Supplementary Fig. 10) and nearly significant for the three northern ones, suggesting gene flow from mountain to plains zebras. Results from TreeMix were inconclusive regarding gene flow (see above), but the highest likelihood trees with one and two migration edges supported interspecies admixture (Supplementary Fig. 11).

Demographic history. A stairway plot of demographic history showed concordant patterns of population size changes among most of the populations (Fig. 6). For all populations except northeastern Uganda, there was an inferred population expansion starting 800–900 ka and a more recent reduction (notably between 10 and 50 ka) to a lower modern effective size. It is well known that population structure confounds the shape of 'instantaneous' population size estimates[21,22]. However, simulation-based analyses suggested that even when structure and gene flow were included in a simplified three-population model, a signal of recent population decline remained (see below). We tested whether the results were sensitive to variable sample sizes by downsampling the larger population samples, and found no such effect (beyond a loss of demographic history resolution with small samples sizes; Supplementary Fig. 12). We also tested the impact of pooling samples randomly across the species range, which in some cases had an effect on the most recent part of the inferred history (Supplementary Fig. 13).

The fastsimcoal2 analysis suggested an initial divergence of plains zebra populations around 367 ka (confidence interval: 326–656 ka; Supplementary Fig. 14 and Supplementary Table 6). This estimate is close to the 275–350 ka divergence time between the plains zebra and quagga estimated in ref. [20], although the inclusion of migration in our model results in an older divergence time as expected. Fastsimcoal2 also confirmed the trends in the stairway plot, identifying a pre-divergence expansion around 600–700 ka. The triangulation method for determining T_{MRCA} yielded an estimate of 600–750 ka, which is also in agreement with previous estimates[15,20,23]. Note the difference between the population divergence time and coalescence time (T_{MRCA}), which was found in the present study as well as in ref. [20]. Inferred migration rates were high at 1.8–8.0 migrants per generation, and the rates were asynchronous with more migration northwards than southwards, supporting a geographic expansion northwards from Botswana. Using a model without migration, the fastsimcoal2 divergence time decreased to 250 ka (in line with ref. [20]).

Discussion

Uncovering cryptic structure in the plains zebra. The operational population scheme presented here contrasts with the prevailing morphology-based[6] and genetic-based[5] conclusions about plains zebra population structure. We identified up to nine extant populations with distinct evolutionary properties and show that the genetic structure is discrete rather than clinal[5,24]. Although isolation by distance between geographic localities has been reported before[5], this does not distinguish between true clinal variation and a stepping stone pattern with discrete demes. Most likely, this SNP marker set allowed us to discover fine-scale structure in the highly mobile plains zebra that was not detectable with mitochondrial DNA and microsatellite markers. Furthermore, the inferred genetic structure does not coincide with the morphology-based subspecies classification used by conservation authorities (refs [4,6]; International Union for Conservation of Nature Red List). For example, *E. q. boehmi* is not a meaningful evolutionary unit. It is subdivided into at least two to four genetic clusters with distinct phylogeographic histories (Supplementary Table 4). The same is true for *E. q. crawshayi*, which lumps the phylogeographically distinct southwestern Tanzania and southeastern Tanzania populations into one. These observations underline the necessity of supplementing morphology-based species assessment with genetic analyses to identify population structure of relevance for understanding phylogeography, as well as for prioritizing the conservation of genetic resources. There are several known cases of cryptic population structure in African mammals, including the kob[25], Grant's gazelle[5,26] and African buffalo[27].

We acknowledge that denser and more comprehensive sampling could reveal further substructure and/or additional clusters in regions that were not represented here. For example, the sampling in the border region between Tanzania and Kenya did not allow us to tease apart the fine-scale structure in this region, where structure appeared to be more pronounced than in other regions of similar size. The Rift Valley system in East Africa is a known driver of biodiversity and within-species structuring, probably caused by the complex topology and climatic history of the region (reviewed in ref. [2]). The results for Tanzania/Kenya, which included admixture from three distinct clusters at variable proportions at $K=8$

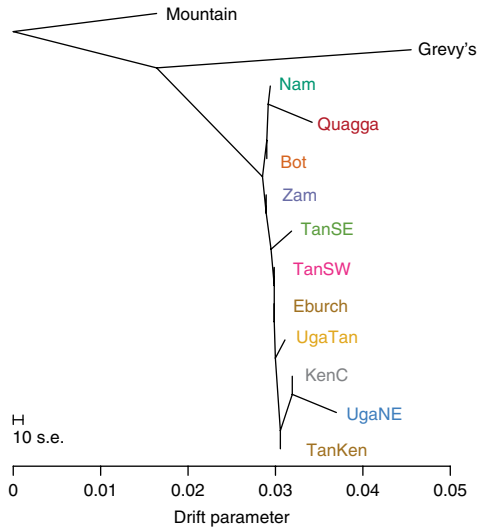

Fig. 5 | Maximum likelihood tree obtained using TreeMix without migration edges for the 67 individuals in dataset 1. The Grevy's zebra and mountain zebra are included and the mountain zebra is specified as the outgroup. *Eburch* and *Quagga* represent the 'burchelli' and 'quagga' samples from ref. [20]. The scale bar indicates ten times the average s.e. The colour scheme is the same as in Fig. 1.

(Fig. 3), support the presence of multiple differentiated populations connected by gene flow in this area, in agreement with physical observations[16]. The phylogeographic inference further corroborates a prominent role of the Rift Valley as a continental migration corridor for the plains zebra. The estimated effective migration surface and directionality analyses showed a bipartition in the northwards expansion overlapping strikingly with the two main branches of the Great Rift Valley system (Fig. 2). Hence, we show that while the Rift Valley constitutes a barrier to gene flow in some species[28], it can also play an important phylogeographic role by facilitating longitudinal gene flow.

A possible savannah refugium in the Zambezi–Okavango wetland areas. Here, we provide insights into the phylogeographic history of the plains zebra and identify the region encompassing northern Botswana and Zambia as the most likely location of the ancestral population, which is consistent with a southern African refugium for savannah ungulates[2]. Among the most notable geomorphological features within this region are two major wetland areas—the Zambezi River basin and the Okavango delta. The Zambezi River basin has been climatically stable over a long time scale[29]. The Okavango delta is also listed as a hydro-refugium, being part of the larger Makgadikgadi paleo-megalake hydrological system[30]. In accordance with these findings, an unspecified southern origin of the plains zebra was proposed[31], positing more temporally stable morphology in southern than eastern African fossils. Furthermore, colonization of eastern Africa from a southern refugium is hypothesized for several co-distributed bovids[2], including the impala[32], eland[33] and wildebeest[34].

A range expansion along the main south-west–north-east axis was indicated by a number of analyses (Figs. 2b, 5 and 6, Supplementary Figs. 10, 11 and 15 and Supplementary Table 7). It is unlikely that recent human influences could be responsible for this pattern of gradual fixation of alleles moving away from the Zambian and Botswanan populations in both directions. The species expansion did not result in strong founder effects, as the genetic diversity is only marginally reduced while moving away from the origin (Supplementary Fig. 16). However, the presence of discrete structure suggests that there must have been barriers to gene flow one or more times during the species history. These could be caused by cycles of climate change related to glacial periods (four such events are known within the past ~360 thousand years[35]).

The simulation analyses favoured relatively ancient population divergence with high gene flow and large population sizes among the central spine of plains zebra populations, rather than recent divergences with little or no gene flow and lower population sizes. In contrast, the marginal populations—particularly the northernmost northeastern Uganda—have markedly reduced genetic diversity and show more genetic isolation and a clear signal of an expansion founder effect[36]. These observations suggest a recent founding of the populations close to the expansion front—in other words, the plains zebra was still expanding its range until recently.

Neutral processes explain most of the genetic differentiation in the plains zebra (Supplementary Fig. 17). However, there were some indications of positive genic and exonic enrichment, always involving the Namibia population (Supplementary Table 8), which, incidentally, is the only plains zebra population residing outside the tropical region, plausibly necessitating a suite of physiological adaptations. A non-significant trend of negative enrichment in exonic sites when Namibia was excluded indicates a stronger role of negative than positive selection acting among populations of the plains zebra (Supplementary Table 8). Hence, we tentatively conclude that local adaptation is not prominent among populations of plains zebra, which is concordant with the high mobility of the species.

We found that the mountain zebra has most likely admixed with the plains zebra. A similar signal was found in ref. [20], although the finding was cautioned as a possible artefact of complex interspecies admixture events involving other equine species. Since all plains zebra populations showed the signal of mountain zebra admixture (Supplementary Fig. 10), the admixture most likely predates the divergence of the plains zebras. This is consistent with a southern origin of the plains zebra and a longstanding presence in this region. However, we found a positive correlation between distance from the mountain zebra and the value of the *D*-statistic (Supplementary Fig. 10), which could suggest recurrent mountain zebra introgression into the southern populations after the plains zebra had diverged and expanded northwards.

The demographic analyses showed that effective population sizes were large for the majority of the species history. Although we saw a boom–bust dynamic of population size, we warn that structure can bias these types of analyses[21,22]. The fastsimcoal2 analysis favoured a model with recent population decline even when gene flow was included, but this apparent increase in genetic drift could possibly be explained by more complex structure than we could model here. Nonetheless, plains zebra populations will not be able to maintain their high levels of genetic diversity if migration corridors are severed. The similarities in the stairway plots suggest that they pertain to the combined meta-population of plains zebras and consequently highlight the importance of gene flow in maintaining high genetic diversity and low differentiation overall. The non-standardized geographic spread of samples within each inferred population (Fig. 2, Supplementary Note 1 and Supplementary Table 1) could potentially affect the SFS patterns as demonstrated in ref. [37] (see also Supplementary Fig. 13). However, the similarity of the inferred population histories (Fig. 6) under these non-standardized sampling configurations reassures us that there is little or no significant sampling configuration bias.

Conclusion

The extant populations of plains zebra derive from a range expansion originating near the largest wetlands in southern Africa—the Zambezi and the Okavango—about 367 ka, providing a likely location of the presumed southern African savannah refugium. We found nine extant evolutionary units in which genetic variability is more discrete than a continuous isolation-by-distance model suggests.

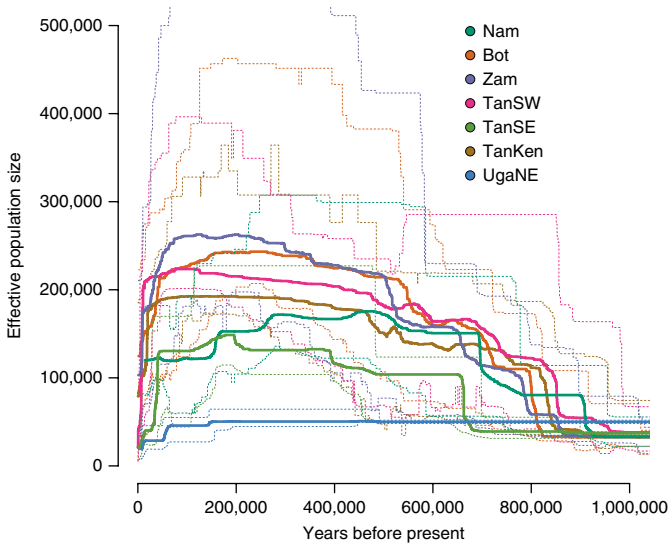

Fig. 6 | Stairway plots for the seven plains zebra populations with at least three samples. Estimates were obtained using the stairway plot method, which is based on the SFS. The estimates are based on the assumption of a mutation rate of 9.05×10^{-10} per site per year and a generation time of 8 years[20]. Dotted lines represent 95% confidence intervals. This analysis was based on 6,912,036–9,246,742 sites from the ancestry-corrected SFS (see Methods).

These are not congruent with the recognized plains zebra subspecies. The plains zebra has high effective population sizes and high genetic diversity, with notable exceptions such as the populations in northeastern Uganda, southwestern Uganda and southeastern Tanzania. We conclude that gene flow—and consequently habitat connectivity—is the main prerequisite for maintaining genetically diverse plains zebra populations.

Methods

Samples and data generation. We generated RAD sequencing data from a sample set of 59 plains zebra (*Equus quagga*), three Grevy's zebra (*Equus grevyi*) and three mountain zebra (*Equus hartmannae*) (see Supplementary Table 2). *Sbf*I RAD libraries were prepared using 250 ng of DNA per individual[38,39]. After P1 (barcoded adapter) ligation, samples with similar quality characteristics, based on agarose gel analyses, were pooled into sub-libraries that were sheared separately through sonication. More fragmented libraries were sheared for a shorter amount of time than less fragmented libraries to retain larger RAD fragments in those individuals. Sub-libraries were mixed at equal molar quantities and sequenced on the Illumina HiSeq 2000 instrument at the University of Oregon Genomics Core Facility.

In addition, we included data from a whole-genome study of the equids[20]. These were a single, captive plains zebra and a specimen of the extinct quagga *E. q. quagga* (Supplementary Tables 1 and 2).

RAD sequencing library construction and sequencing data processing. Before population genetic analyses, we filtered and demultiplexed the raw data (removing reads with excessive low-quality bases and trimming to 80 base pair read length due to declining base qualities near the end of the reads[40,41]). The resulting FastQ files were mapped to the horse reference genome EquCab2.0 using bwa-mem with default settings[42]. GATK[43] was used to make local realignments using the default settings of IndelRealigner[44]. The samples yielded 857,746–7,101,332 (mean 2,465,795) reads mapping with quality ≥ 30. We used the resulting BAM files as input to analyses in ANGSD version 0.91128[45]. Among the 7,051,457 sites present in all RAD sequencing samples, the per-sample mean depth of coverage was 7–38 (median 15).

While some analyses were carried out on raw genotype likelihoods estimated with ANGSD[45], we supplemented these with genotype calls from a high-depth subset of the sites to allow complementary analyses. The genotype calling was done collectively on a total of 67 samples comprising our 65 zebra samples, one quagga and one plains zebra sample from ref. [20]. This was done in ANGSD using the following filters: a minimum base quality of 20, a minimum mapping quality of 30, a polymorphism *P* value threshold of 10^{-6}, a minimum depth of 10 per sample to call a genotype and at least 65 individuals passing the filters. Genotype calls were processed to PLINK format files[46] and used in subsequent analyses. Some analyses were carried out on subsets of the complete dataset, comprising either the quagga and remaining plains zebra or only the plains zebra (Supplementary Table 2). For all datasets using genotype calls, we serially filtered the genotype file using a missing data filter of 20% and removed sites that were non-variable in the subset.

Several of the population genetic analyses involve estimation of the SFS, either per population or per population pair (one- and two-dimensional SFS). This was done using the -doSaf and -realSFS methods in ANGSD[45]. We excluded sites that were not present in all individuals within populations. Since some of the analyses can be sensitive to the correct identification of the ancestral allele, we inferred the SFS in a two-stage process. For the one-dimensional SFS, first, we estimated unfolded site allele frequencies for each of the populations, including the two outgroup species—the Grevy's and Mountain zebra—polarized with respect to the horse reference genome. Then, we estimated the three-dimensional SFS between the mountain, Grevy's and each plains zebra population and extracted the target population SFS conditional on the site being fixed ancestral in the two outgroups. This process minimizes misidentification of the derived allele in the plains zebra populations by removing sites that experienced a mutation on the branch ancestral to Grevy's and plains zebra. Preliminary analyses confirmed that failure to do so would overestimate the number of high-frequency derived alleles considerably, which would lead to biased demographic inference. We estimated pairwise two-dimensional SFS for each of the 21 pairs of populations in a similar way, by first estimating three-dimensional SFS for the same pair with the inclusion of Grevy's zebra and using only the sites that were fixed for the horse reference allele in the Grevy's zebra.

Population identification. Due to limited and non-uniform sample sizes from each geographical location, we took an ad-hoc approach to the identification of populations (see Supplementary Note 1 'Population identification criteria'). Here, populations are loosely defined as geographically coherent groups of samples that show evidence of being genetically distinct entities. The purpose of this operational definition of populations is to provide a null model for population structure within the plains zebra that is useful for understanding the phylogeography of the species and identifying prospective evolutionarily significant units.

As some inferred populations had low sample sizes and some analyses may be sensitive to varying sample size, we repeated these analyses using a randomly selected standardized (downsampled) number of samples (see details below).

Genetic clusters and phylogeography. As an initial representation of the genetic differentiation, we performed a principal component analysis on the dataset including our plains zebras and the quagga (dataset 2) (Supplementary Table 2). This was implemented using the R package snpRelate[47]. The programme EEMS[48] was used to estimate and visualize barriers to gene flow in a geographical context among the non-captive plains zebra in dataset 2. First, we calculated the pairwise genetic dissimilarities using the co-distributed tool bed2diffs, without imputation of missing SNP genotypes. This dissimilarity measure is simply defined as the average number of allelic differences between each pair of individuals across all genotyped loci. We set up a deme grid of 300 points and ran the Markov chain Monte Carlo programme for 5 million iterations, using 1 million iterations as burn-in and a default thinning interval of 10%. Visual inspection of the likelihood trace, as well as tolerable acceptance rates for the parameters (between 12 and 42%), confirmed good mixing and convergence of the chain. This result was corroborated by a second independent run, which showed almost identical results (Supplementary Fig. 18). Hyperprior values were selected based on a preliminary run, modifying the individual priors as recommended in the programme documentation[48].

We ran NGSadmix[49] to infer admixture patterns and identify genetic clusters within the plains zebra samples. To achieve this, we produced a Beagle genotype likelihood file using ANGSD version 0.911, restricting to sites that were found in all 60 individuals within dataset 2 (-minInd 60 option in ANGSD). We ran NGSadmix with 1–10 ancestral clusters and each successive *K* was repeated 100 times or until 3 replicate runs were within 2 likelihood units of the highest value. Here, we only report the admixture runs that converged using these criteria.

We characterized the relationship between the inferred plains zebra individuals and populations using different metrics as proxies for the genetic structure. First, we constructed a neighbour-joining tree using the NJ function in the R package APE[50] on dataset 1, using the mountain zebra samples as the outgroup and the genetic dissimilarity matrix calculated in estimated effective migration surface as a distance measure. Due to short internal branch lengths, we plotted the topology of the tree without using branch lengths. Having confirmed that the phylogenetic grouping of individuals agreed with the populations outlined in Fig. 1, we then inferred the population tree of the plains zebras and the quagga (dataset 1) using TreeMix[51]. We initialized the analysis using blocks of 100 SNPs as we assume that the large population size of plains zebra has broken down any extended linkage disequilibrium (LD) tracts. We ran from 0 to 5 migration events each with 100 different starting points. The tree topology and placement of the migration edges were not consistent across replicate runs. We therefore show the highest likelihood tree without migration edges. We repeated the TreeMix runs without the two samples from ref. [20].

To test the hypothesis of a range expansion in contrast with a simple equilibrium isolation-by-distance model, we calculated the directionality index ψ[36]. The rationale for this statistic is that the further a population sample is from the origin of a range expansion, the higher the probability that a SNP increases in allele frequency or becomes fixed[52]. We randomly downsampled each population to two samples using dataset 2. We estimated the pairwise two-directional SFS as described above between all 36 such population pairs and applied equation (1b) in ref. [36] to calculate the pairwise ψ values. The SFSs were bootstrapped 100 times to obtain confidence intervals around the directionality index.

For each population pair, a two-dimensional SFS was used to estimate a genome-wide value of Hudsons F_{ST} (as described in refs [53,54]) between the populations. Hudson's F_{ST} was preferred because it has been shown to be less biased by sample size than other F_{ST} estimators[54,55], which we also confirmed with this dataset.

To further test and corroborate cluster composition and admixture signals, we carried out *D*-statistic[56] tests of various groupings of samples. This was done using custom R scripts on the called genotypes in dataset 1. The *D*-statistic was used to test interspecies admixture, which was previously found to be common among equids[20] and to further validate the identification of populations by testing the genetic groupings in the species. Preliminary *D*-statistic analyses found that the Grevy's and mountain zebra were not admixed with the plains zebra and were therefore eligible to be used as true outgroups in the analyses (Supplementary Note 2 'Preliminary test of admixture in Grevy's and mountain zebra' and Supplementary Figs. 19–24).

Heterozygosity and runs of homozygosity. We calculated the proportion of heterozygous genotypes from the total number of genotypes for each individual plains zebra using dataset 3. To do so, we used realSFS[45] to estimate the unfolded SFS for each individual. Specifically, this meant estimating the counts of three classes of genotype within each individual—two homozygous genotypes and one heterozygous genotype. The genome-wide heterozygosity was calculated as the proportion of heterozygous genotypes of the total number of genotypes per individual. We also determined the extent of inbreeding using runs of homozygosity (ROH) of each individual mapped to the horse reference genome. For this analysis, we used PLINK on the called genotypes in dataset 3. We excluded ROH segments if they spanned fewer than 5,500 kilobases and required at least 200 SNPs to be present within each ROH segment to make sure that these were indeed identity-by-descent segments. These requirements were sensible, as there have been observations that ROH segments in lengths of more than 4 megabases were found in outbred individuals[57]. Additionally, using these settings, we excluded determining segments as ROH with a limited SNP density. Furthermore, we wanted conservative filters given the relatively sparse marker density in our RAD sequencing data. Here, we use the term inbreeding coverage, $F_{coverage}$, as has been done previously[19]. We plotted the chromosomal distribution of ROH for the most inbred individuals using a cutoff of $F_{coverage} = 0.0625$—corresponding to an offspring of a first cousin mating. $F_{coverage}$ was approximated as the fraction of the summed ROH length out of the total genome size.

Stairway plots, divergence times and coalescent simulations. To investigate the demographic history of the plains zebra, we used two complementary SFS-based analyses—the stairway plot[58] and fastsimcoal2[59,60]. For both analyses, we used a mutation rate of 9.05×10^{-10} per site per year and a generation time of 8 years[20]. In the stairway plot analyses, we used dataset 2 and included only the populations with ≥3 samples, hence excluding Uganda/Tanzania, central Kenya and the quagga.

In the fastsimcoal2 analysis, we used only the three largest population samples—Botswana, southeastern Tanzania and Tanzania/Kenya—to keep the demographic model tractable and allow the use of the multiSFS option in fastsimcoal2, which leads to better estimation of the composite likelihoods than using pairwise two-dimensional SFS[60]. We estimated the three-dimensional SFS between the three populations using the strategy outlined above. We devised a historical model for the populations based on the results from the other analyses (mainly the order of population splits). We used the expectation conditional maximization procedure described in ref. [60] to estimate 15 model parameters running 50 replicates of the optimization procedure each with 40 expectation conditional maximization cycles and 100,000 simulations per cycle. The replicate with the highest estimated likelihood was used as the maximum likelihood parameter estimate. Two alternative and simpler demographic models were tested: a model without gene flow and a model without the recent change in population size. Both yielded a significantly poorer fit to the observed three-dimensional SFS by Akaike information criterion model comparison[60]. We performed parametric bootstrapping under the optimal scenario by simulating 100 three-dimensional SFSs under these maximum likelihood parameter values, then using these to estimate a confidence interval of the model parameters[60]. Finally, we tested the fit of the best model by computing a bootstrapped distribution of the composite likelihood ratio using the same 100 simulations as above[60]. For all analyses, we restricted the optimization to the SFS cells with more than 20 entries.

As a complementary approach to gauge the divergence time of plains zebras, we also inferred the time since the most recent common ancestor of plains zebras, or T_{MRCA}, using a triangulation method used for ancient human genomes[18,19].

This method takes the differences in the number of derived alleles between two closely related subspecies individuals (H1 and H2) and one closely related outgroup sequence (H3). Specifically, the number of derived alleles seen in H1, which are not seen in either H2 or H3, will be assumed to have arisen on the branch leading to sister species one (H1) subsequent to the split from sister species two (H2). Assuming an equal mutation rate over all branches, we can estimate the split time between H1 and H2 using the known split time between the branch leading to the two sister subspecies and the common ancestor sequence, as well as the number of unique alleles on the common ancestor branch[19]. We estimated the T_{MRCA} between the quagga, Namibia, Botswana and Tanzania/Kenya populations using this method. We removed transitions for the quagga. Note that the T_{MRCA} is expected to be older than the population divergence, and more so if the ancestral population was large and/or structured[61].

Selection analyses. Preliminary analyses using the software pcadapt[62] identified a high number of outlier SNP candidates for local adaptation (1,657 or 7% of the included sites; Supplementary Note 3 'Selection scans across the genome'). This is in agreement with the expectation of a substantially increased rate of false positives (Supplementary Fig. 25) when the underlying history is a range expansion[62], probably due to the phenomenon of allelic surfing[63,64]. We therefore refrain from making conclusions about specific genomic regions or sites involved in local adaptation. Instead we reran a number of the analyses with the exclusion of these 1,657 potentially non-neutral sites to exclude any possible bias stemming from violating the neutrality assumption of most of the phylogeographic methods (Supplementary Note 3 'Selection scans across the genome', Supplementary Figs. 26–35 and Supplementary Tables 9 and 10). We also used another, more simplistic and conservative approach to assess selection patterns and local adaptation within the plains zebra. This approach is outlined in Supplementary Note 3 'Selection scans across the genome'.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. Called genotypes for the most inclusive dataset (dataset 1) and the list of SNPs putatively under selection have been deposited in the Figshare depository with the data https://doi.org/10.6084/m9.figshare.5678737.v1. R scripts used for all analyses are available upon request.

Received: 13 March 2017; Accepted: 14 December 2017;
Published online: 22 January 2018

References

1. Owen-Smith, N. & Cumming, D. H. M. Comparative foraging strategies of grazing ungulates in African savanna grasslands. In Proc. XVII Int. Grasslands Congress New Zealand, 691–698 (SIR Publishing, Wellington, 1993).
2. Lorenzen, E. D., Heller, R. & Siegismund, H. R. Comparative phylogeography of African savannah ungulates. *Mol. Ecol.* **21**, 3656–3670 (2012).
3. Bell, R. H. V. A grazing ecosystem in the Serengeti. *Sci. Am.* **225**, 86–93 (1971).
4. Hack, M. A., East, R., Rubenstein, D. I. & Moehlman, P. A. in Equids: Zebras, Asses and Horses: Status Survey and Conservation Action Plan (ed. Moehlman, P. D.) 43–60 (IUCN/SSC Equid Specialist Group, Gland and Cambridge, 2002).
5. Lorenzen, E. D., Arctander, P. & Siegismund, H. R. High variation and very low differentiation in wide ranging plains zebra (*Equus quagga*): insights from mtDNA and microsatellites. *Mol. Ecol.* **17**, 2812–2824 (2008).
6. Groves, C. P. & Bell, C. H. New investigations on the taxonomy of the zebras genus *Equus*, subgenus *Hippotigris*. *Mamm. Biol.* **69**, 182–196 (2004).
7. Leonard, J. et al. A rapid loss of stripes: the evolutionary history of the extinct quagga. *Biol. Lett.* **1**, 291–295 (2005).
8. Oakenfull, E. A., Lim, H. N. & Ryder, O. A. A survey of equid mitochondrial DNA: implications for the evolution, genetic diversity and conservation of *Equus*. *Conserv. Genet.* **1**, 341–355 (2000).
9. Rubenstein, D., Low Mackey, B., Davidson, Z. D., Kebede, F. & King, S. R. B. Equus grevyi. IUCN Red List of Threatened Species 2016: e.T7950A89624491 (IUCN, accessed 10 September 2017); https://doi.org/10.2305/IUCN.UK.2016-3.RLTS.T7950A89624491.en.
10. Moritz, C. Defining 'evolutionarily significant units' for conservation. *Trends Ecol. Evol.* **9**, 373–375 (1994).
11. Crandall, K. A., Bininda-Emonds, O. R. R., Mace, G. M. & Wayne, R. K. Considering evolutionary processes in conservation biology. *Trends Ecol. Evol.* **15**, 290–295 (2000).
12. Churcher, C. S. & Richardson, M. L. in Evolution of African Mammals (eds Maglio, V. J. & Cooke, H. B. S.) 379–422 (Harvard Univ. Press, Cambridge, 1978).
13. Klein, R. G. & Cruz-Uribe, K. Craniometry of the genus *Equus* and the taxonomic affinities of the extinct South African quagga. *S. Afr. J. Sci.* **95**, 81–86 (1999).

14. Orlando, L. et al. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).
15. Vilstrup, J. T. et al. Mitochondrial phylogenomics of modern and ancient equids. *PLoS. ONE* **8**, e55950 (2013).
16. Caro, T., Jones, T. & Davenport, T. R. B. Realities of documenting wildlife corridors in tropical countries. *Biol. Conserv.* **142**, 2807–2811 (2009).
17. Bradburd, G. S., Ralph, P. L., Coop, G. M. & Slatkin, M. A spatial framework for understanding population structure and admixture. *PLoS Genet.* 12, e1005703 (2016).
18. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
19. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
20. Jónsson, H. et al. Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proc. Natl. Acad. Sci. USA* **111**, 18655–18660 (2014).
21. Heller, R., Chikhi, L. & Siegismund, H. R. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS. ONE* **8**, e62992 (2013).
22. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. On the importance of being structured: instantaneous coalescence rates and a re-evaluation of human evolution. *Heredity* **116**, 362–371 (2016).
23. Orlando, L. et al. Revising the recent evolutionary history of equids using ancient DNA. *Proc. Natl. Acad. Sci. USA* **106**, 21754–21759 (2009).
24. Groves, C. & Grubb, P. *Ungulate Taxonomy* (Johns Hopkins Univ. Press, Baltimore, 2011).
25. Lorenzen, E. D., De Neergaard, R., Arctander, P. & Siegismund, H. R. Phylogeography, hybridization and Pleistocene refugia of the kob antelope (*Kobus kob*). *Mol. Ecol.* **16**, 3241–3252 (2007).
26. Siegismund, H. R., Lorenzen, E. D. & Arctander, P. in *Mammals of Africa: Volume VI. Pigs, Hippopotamuses, Chevrotain, Giraffes, Deer and Bovids* (eds Kingdon, J. & Hoffmann, M.) 373–379 (Bloomesbury, London, 2013).
27. Smitz, N. et al. Pan-African genetic structure in the African buffalo (*Syncerus caffer*): investigating intraspecific divergence. *PLoS ONE* **8**, e56235 (2013).
28. Lorenzen, E. D., Simonsen, B. T., Kat, P. W., Arctander, P. & Siegismund, H. R. Hybridization between subspecies of waterbuck (*Kobus ellipsiprymnus*) in zones of overlap with limited introgression. *Mol. Ecol.* **15**, 3787–3799 (2006).
29. Castañeda, I. S. et al. Hydroclimate variability in the Nile River Basin during the past 28,000 years. *Earth Planet. Sci. Lett.* **438**, 47–56 (2016).
30. Reynolds, D. J. et al. Reconstructing North Atlantic marine climate variability using an absolutely-dated sclerochronological network. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **465**, 333–346 (2017).
31. Reynolds, S. C. Mammalian body size changes and Plio–Pleistocene environmental shifts: implications for understanding hominin evolution in eastern and southern Africa. *J. Hum. Evol.* **53**, 528–548 (2007).
32. Lorenzen, E. D., Arctander, P. & Siegismund, H. R. Regional genetic structuring and evolutionary history of the impala *Aepyceros melampus*. *J. Hered.* **97**, 119–132 (2006).
33. Lorenzen, E. D., Masembe, C., Arctander, P. & Siegismund, H. R. A long-standing Pleistocene refugium in southern Africa and a mosaic of refugia in East Africa: insights from mtDNA and the common eland antelope. *J. Biogeogr.* **37**, 571–581 (2010).
34. Arctander, P., Johansen, C. & Coutellec-Vreto, M.-A. Phylogeography of three closely related African bovids (tribe Alcelaphini). *Mol. Biol. Evol.* **16**, 1724–1739 (1999).
35. Hewitt, G. M. Genetic consequences of climatic oscillations in the Quaternary. *Phil. Trans. R. Soc. Lond. B* **359**, 183–195 (2004).
36. Peter, B. M. & Slatkin, M. Detecting range expansions from genetic data. *Evolution* **67**, 3274–3289 (2013).
37. Städler, T., Haubold, B., Merino, C., Stephan, W. & Pfaffelhuber, P. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**, 205–216 (2009).
38. Davey, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510 (2011).
39. Puckett, E. E., Etter, P. D., Johnson, E. A. & Eggert, L. S. Phylogeographic analyses of American black bears (*Ursus americanus)* suggest four glacial refugia and complex patterns of postglacial admixture. *Mol. Biol. Evol.* **32**, 2338–2350 (2015).
40. Rašić, G., Filipović, I., Weeks, A. R. & Hoffmann, A. A. Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. *BMC Genom.* **15**, 275 (2014).
41. Sutherland, B. J. G. et al. Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps. *Genome Biol. Evol.* **8**, 3600–3617 (2016).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
44. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
45. Korneliussen, T., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
46. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
47. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
48. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2014).
49. Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Estimating individual admixture proportions from next generation sequencing data. *Genetics* **195**, 693–702 (2013).
50. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
51. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS. Genet.* **8**, e1002967 (2012).
52. Slatkin, M. & Excoffier, L. Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics* **191**, 171–181 (2012).
53. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
54. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
55. Willing, E. M., Dreyer, C. & Ova Oosterhout, C. Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. *PLoS. ONE* **7**, e42649 (2012).
56. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
57. McQuillan, R. et al. Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
58. Liu, X. & Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47**, 555–559 (2015).
59. Excoffier, L. & Foll, M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334 (2011).
60. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS. Genet.* **9**, e1003905 (2013).
61. Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G. & Schierup, M. H. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS. Genet.* **7**, 1–15 (2011).
62. Luu, K., Bazin, E. & Blum, M. G. B. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* **33**, 67–77 (2016).
63. Edmonds, C. A., Lillie, A. S. & Cavalli-Sforza, L. L. Mutations arising in the wave front of an expanding population. *Proc. Nat. Acad. Sci. USA* **101**, 975–979 (2004).
64. Klopfstein, S., Currat, M. & Excoffier, L. The fate of mutations surfing on the wave of a range expansion. *Mol. Biol. Evol.* **23**, 482–490 (2006).
65. Novellie, P. Equus zebra ssp. zebra. IUCN Red List of Threatened Species 2008: e.T7959A12876612(IUCN, accessed 10 September 2017); http://dx.doi.org/10.2305/IUCN.UK.2008.RLTS.T7959A12876612.en.

Acknowledgements

The authors thank A. al-Cher for laboratory work in connection with this study. The work was funded by research grants from the The Danish Council for Independent Research | Natural Sciences, the Lundbeck Foundation and the Villum Foundation.

Author contributions

C.-E.T.P and R.H. designed and performed the experiments, analysed the data and wrote the paper. A.A. developed the analytical tools and analysed the data. H.R.S., P.D.E., E.A.J., L.O. and L.C. analysed the data and wrote the paper.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-017-0453-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.-E.T.P. or R.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Casper-Emil Tingskov Pedersen & Rasmus Heller

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

See Supplementary Note 1: Population identification criteria. Also, see Supplementary Table 1 listing the samples, their origin and how we grouped them, and Supplementary Table 3, which states how the genetic clusters match between a concert of analyses. Furthermore, we have thoroughly described any analyses where a subset of the samples were used.

2. Data exclusions

Describe any data exclusions.

See Supplementary Table 2 listing different subsets of the full data and when they were used. Based on reviewer recommendations we excluded SNPs under selection, as they may distort the demographic signal. Information on these SNPs is available in Supplementary Note 3. Figures from these analyses based on a reduced dataset are available in the Supplementary Fig. 26-35 and Supplementary Tables 9 and Supplementary Table 10.

3. Replication

Describe whether the experimental findings were reliably reproduced.

No manipulative experiments were carried out. In MCMC and other optimization analyses we critically evaluated whether the results had converged by running several independent chains and checking for convergent output.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

See Supplementary note 1: Population identification criteria

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not applicable.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	We refer to the Materials and methods section, as we thoroughly go through all the software used.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	There are no restrictions to the samples used in this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	*For all antibodies, as applicable, provide supplier name, catalog number, clone name, and lot number. Also describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript OR state that no antibodies were used.*

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	*Provide information on cell line source(s) OR state that no eukaryotic cell lines were used.*
b. Describe the method of cell line authentication used.	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used have been authenticated OR state that no eukaryotic cell lines were used.*
c. Report whether the cell lines were tested for mycoplasma contamination.	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination OR state that no eukaryotic cell lines were used.*
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	*Provide a rationale for the use of commonly misidentified cell lines OR state that no commonly misidentified cell lines were used.*

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	We refer the Supplementary Table 1, as very specific information about each individual is available there.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	*Provide all relevant information on human research participants, such as age, gender, genotypic information, past and current diagnosis and treatment categories, etc. OR state that the study did not involve human research participants.*