Genotype likelihoods incorporated in non-linear dimensionality reduction techniques infer fine-scale population genetic structure

F. Gözde Çilingir[1,2†*], Kerem Uzel[3,4†], Christine Grossen[2]

[1] University of Zurich, Department of Evolutionary Biology and Environmental Studies, Zurich, Switzerland
[2] Swiss Federal Research Institute WSL, Biodiversity and Conservation Biology Research Unit, Birmensdorf, Switzerland
[3] University of Zurich, Brain Research Institute, Laboratory of Neuroepigenetics, Zurich, Switzerland
[4] ETH Zurich, Institute for Neuroscience, Department of Health Sciences and Technology, Zurich, Switzerland

† These authors contributed equally to this study
* Corresponding author, email fgcilingir@gmail.com

Abstract

Understanding population structure is essential for conservation genetics, as it provides insights into population connectivity and supports the development of targeted strategies to preserve genetic diversity and adaptability. While Principal Component Analysis (PCA) is a common linear dimensionality reduction method in genomics, the utility of non-linear techniques like t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) for revealing population genetic structures has been largely investigated in humans and model organisms but less so in wild animals. Our study bridges this gap by applying UMAP and t-SNE, alongside PCA, to medium and low-coverage whole-genome sequencing data from the scimitar oryx, once extinct in the wild, and the Galápagos giant tortoises, facing various threats. By estimating genotype likelihoods from coverages as low as 0.5x, we demonstrate that UMAP and t-SNE outperform PCA in identifying genetic structure at reduced genomic coverage levels. This finding underscores the potential of these methods in conservation genomics, particularly when combined with cost-effective, low-coverage sequencing. We also provide detailed guidance on hyperparameter tuning and implementation, facilitating the broader application of these techniques in wildlife genetics research to enhance biodiversity conservation efforts.

1. Introduction

Population genetics is a rapidly expanding field that aims to elucidate the genetic relationships within and between populations of a given species, thereby transforming our understanding of biodiversity. Population genetic approaches enable the estimation of population size, determination of demographic history, and characterization of population structure, which are all critical in evaluating the long-term survivability of a population, in

addition to, for instance, identifying loci that contribute to adaptive capacity or uncovering the genetic basis of reduced fitness in populations (Hohenlohe et al., 2021). Understanding the population genetic structure is particularly important for endangered species, where identification of genetically isolated or distinct populations is essential for effective conservation efforts at implementing habitat restoration, establishing protected areas, facilitating genetic connectivity, or preserving genetic diversity while preventing the loss of unique genetic variation (Frankham, 2003; Hohenlohe et al., 2021). Therefore, the application of population genetics tools is crucial in assessing, monitoring, and maintaining ecosystem health and biodiversity.

The emergence of next-generation sequencing has been a pivotal force in transforming population and conservation genetics, facilitating the high-throughput acquisition of DNA sequencing data. However, this comes with the challenge of sequencing errors associated with the sequencing platform used. To address this, one strategy is to increase the sequencing depth per sample. Researchers often balance the higher costs by sampling fewer individuals, sequencing pools of individuals (Pool-seq [Schlötterer et al., 2014]), or employing reduced representation sequencing to sequence a smaller portion of the genome per individual (e.g., RAD-seq [Baird et al., 2008] and a range of related techniques [Andrews et al., 2016]). As an efficient alternative, low-coverage whole genome sequencing offers a cost-effective strategy for comprehensive population-scale genomic screening that, in many instances, is as economical as reduced representation methods (Lou et al., 2021). At low depths of coverage, individual genotypes cannot reliably be called; rather, probabilistic models utilizing genotype likelihoods have been introduced (Nielsen et al., 2011, 2012). These models integrate base quality scores and allele sampling errors (reflected as likelihoods of each of the possible genotypes at a particular site) to account for genotype uncertainty in subsequent analyses (Korneliussen et al., 2014). Recently, a growing number of tools have emerged that use genotype-likelihoods for various analyses. These tools perform different tasks such as analyzing population structure (Meisner & Albrechtsen, 2018; Skotte et al., 2013), scanning for selection (Meisner & Albrechtsen, 2018), estimating pairwise linkage disequilibrium (Fox et al., 2019; Fumagalli et al., 2014), carrying out genome-wide association studies, testing for introgression, quantifying genetic differentiation across populations, conducting neutrality tests within a single population (Korneliussen et al., 2014), as well as assessing within-population genetic diversity and individual heterozygosity (Korneliussen et al., 2014; Link et al., 2017).

Regardless of whether high- or low-coverage sequencing methods are employed, the task of unraveling the complex structure within population genomics datasets remains a constant

challenge. The primary goal is to determine whether the samples belong to a uniform population or distinct subgroups and to quantify the evidence for such subgroups. However, with the developments in high-throughput sequencing technologies, these analyses have become increasingly complex and challenging to perform on large datasets containing many sampled individuals, genetic markers, and populations (Shafer et al., 2015). To overcome the challenges of such complex and high-dimensional datasets, dimensionality reduction techniques have been used as the first step in analysis to visualize and identify relatedness patterns in the data. Principal component analysis (PCA) was first introduced to the study of genetic data 45 years ago by Cavalli-Sforza (Menozzi et al., 1978), and one of the wide application areas of this technique is inferring population genetic structure (Patterson et al., 2006). By generating orthogonal axes that capture the maximum variation within the high-dimensional space, PCA provides a low-dimensional representation of the genomic data. This technique has been applied successfully to summarize covariances among hundreds of thousands of loci, making it an essential tool in the field (Patterson et al., 2006). However, PCA has several limitations when dealing with complex genomic data. For example, it uses linear combinations of variables, potentially overlooking non-linear relationships between genetic markers (Alanis-Lobato et al., 2015). Also, PCA seeks directions of maximum variance and, in doing so, may neglect variation along other directions, leading to incomplete or biased representations of population structure.

Non-linear neighbor graph-based dimension reduction algorithms have been developed over the years to address these limitations. Using such methods on population genomics data can provide a more detailed and accurate representation of the population structure. One such method is t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008), which has gained popularity in recent years due to its ability to capture local structures in genomics and transcriptomics data (Kiselev et al., 2019; Li et al., 2017; Platzer, 2013). Another algorithm is uniform manifold approximation and projection (UMAP) (McInnes, Healy, & Melville, 2018), which can also preserve non-linear structure in high-dimensional data and has already been used to investigate patterns and relationships across different levels of dataset complexity and size (Becht et al., 2018; Diaz-Papkovich et al., 2019, 2020; Dorrity et al., 2020). t-SNE analyzes the similarity of points in high-dimensional space using a Gaussian distance (van der Maaten & Hinton, 2008), whereas UMAP is based on generating a weighted graph where data points in close proximity to each other are given greater weights (McInnes, Healy, & Melville, 2018); thus, both algorithms preserve the local topology of the neighborhood.

In standard applications of t-SNE and UMAP for dimensionality reduction, these techniques

can be applied directly to individual genotypes. However, due to the computational intensity of these methods, an alternative approach often employed is first to reduce the dimensionality of the data using PCA (Kobak & Berens, 2019; Kobak & Linderman, 2021). Subsequently, top PCs are used as the input for further dimensionality reduction using t-SNE or UMAP (referred to as PCA-t-SNE and PCA-UMAP, respectively) (Diaz-Papkovich et al., 2019; van der Maaten & Hinton, 2008). This two-step process not only helps in managing computational demands but also in enhancing the extraction of meaningful population structures by filtering out stochastic noise (Diaz-Papkovich et al., 2019). In contrast to using PCA alone, where typically only the first two principal components (PCs) are utilized for visualization, combining PCA with t-SNE or UMAP allows for the integration and representation of a greater portion of the data's variance (Gaspar & Breen, 2019). Additionally, in scenarios where genotype data is unavailable, such as in low-coverage whole-genome sequencing, PCs computed based on genotype likelihoods can be effectively incorporated into t-SNE and UMAP analyses. This adaptability further underscores the versatility of these dimensionality reduction techniques in various genomic data contexts.

Although at different computing costs between t-SNE and UMAP, both techniques integrating principal components of genotype data were previously applied to large datasets (e.g., [Diaz-Papkovich et al., 2020]), providing a more comprehensive view of inferring population genetic structure in plants (Fu et al., 2022; Ma et al., 2021), invertebrates (*Anopheles gambiae* 1000 Genomes Consortium, 2020; Schmidt et al., 2020, 2021; Simon et al., 2021), and extensively in humans (Černý et al., 2023; Chyleński et al., 2019; Diaz-Papkovich et al., 2019; Halldorsson et al., 2022; Margaryan et al., 2020; Sengupta et al., 2021; Sohail et al., 2023). Interestingly, despite being available for several years, t-SNE and UMAP have not been widely adopted in population genomics research for non-human vertebrates. This is likely due to researchers being unfamiliar with these techniques and the difficulties associated with determining the optimal parameter settings and interpreting the resulting visualizations in the context of wildlife population genomics. As a result, in the field of conservation genomics for species at risk, there exists a notable underutilization of non-linear dimensionality reduction methods. Additionally, the potential benefits of integrating these methods with genotype likelihood approaches (and hence low-coverage sequencing) in conservation genomics present a promising research direction that remains to be fully investigated.

To contribute to the current state of knowledge, this paper demonstrates the application of t-SNE and UMAP to two previously published medium to low-coverage whole genome resequencing datasets. These datasets originate from two distinct vertebrate groups: the

captive populations of the scimitar-horned oryx, a mammal species once extinct in the wild, and Galápagos giant tortoises, a reptilian taxa facing varying levels of threats. We explored the impact of reduced sequencing depths (down to 0.5x) on the discernibility of clustering patterns when employing these dimensionality reduction methodologies in comparison to classical PCAs. Additionally, we have developed a comprehensive guideline that includes Python-based codes, as well as a guide for optimizing parameters, to facilitate the adoption of these analytical techniques to a wide range of organisms. We see particular potential for these techniques in the field of conservation genomics, where identifying fine-scale population structure, coupled with traditional statistical approaches, can potentially serve for better management strategies of the species at risk.

2. Materials and Methods

2.1 Case system 1: Scimitar-horned oryx

The scimitar-horned oryx (*Oryx dammah*) is a large antelope that once roamed widely across North Africa (Bertram, 1988). However, the iconic animals experienced a precipitous decline in the 20th century due to drought, hunting, and land-use competition (Dixon et al., 1991), leading to their extinction in the wild (IUCN SSC Antelope Specialist Group, 2016). Before their disappearance, captive breeding began with fewer than 100 oryxes from Chad, expanding the ex-situ global population to about 15,000 individuals (Gilbert, 2019). While 1,000 are in coordinated breeding programs, many reside in places with minimal genetic management (Humble et al., 2023). Due to the success of reintroduction programs in Chad, the species was recently downlisted to Endangered by the IUCN Red List (IUCN SSC Antelope Specialist Group, 2023). To understand the genetic consequences of different conservation management strategies, (Humble et al., 2023) sequenced the whole genomes of 49 oryxes from populations with varying genetic management levels, notably from EAZA Ex Situ Programmes (EEP, n = 8), USA (n=17), and two unmanaged collections in the United Arab Emirates (EAD A, n = 9 & EAD B, n = 15).

2.1.1 Whole-genome resequencing data analysis

We obtained the raw whole-genome re-sequencing data of 46 genetically unrelated scimitar-horned oryx individuals from (Humble et al., 2023) (NCBI BioProject PRJEB37295), which were also utilized as the final dataset in the PCA analysis presented in the referenced study ((Humble et al., 2023); Table S1). Subsequent data processing was performed using the ATLAS Pipeline v7 (Link et al., 2017; Marchi et al., 2022; https://atlaswiki.netlify.app/atlas-pipeline), involving the Gaia, Rhea, and Perses workflows. In Gaia, individual sequencing data underwent quality trimming using TrimGalore v0.6.6 (Krueger, 2016) with default settings and was aligned to the *Oryx dammah* reference genome, SCBI_Odam_1.1 (NCBI

RefSeq GCF_014754425.2) (Humble et al., 2020) employing BWA-MEM v0.7.17 (Li & Durbin, 2009) in paired-end mode. Within the Rhea workflow, local re-alignment was conducted using GATK v3.8 (McKenna et al., 2010) (Table S1). Subsequently, the Perses workflow facilitated the merging of forward and reverse reads into single reads per fragment using the tool ATLAS v0.9 (Link et al., 2017), avoiding pseudo-duplicated bases where original reads overlapped.

To reproduce the PCA results as detailed in (Humble et al., 2023), we implemented downsampling with ATLAS (Link et al., 2017) (task=downsample) to standardize the coverage across individuals, setting high-coverage samples to an average coverage of 6x while preserving the original coverage levels for low-coverage individuals in accordance with the methodology outlined in the cited study. We designated the primary dataset as "SO_6x". In this set, we maintained the 20 samples with coverage below the 6x average and downsampled the remaining 26 to the target of ~6x (Table S1). For subsequent analyses to explore the effect of lower sequencing coverage, we created "SO_2x" and "SO_0.5x", where we downsampled all samples to about 2x and 0.5x, respectively.

2.1.2 Genotype likelihood estimation and PCA

We estimated genotype likelihoods using ANGSD v0.940 (Korneliussen et al., 2014) with the three distinct datasets described above, each characterized by varying levels of average coverage (6x, 2x, 0.5x). Following the methodology of (Humble et al., 2023), we employed ANGSD (Korneliussen et al., 2014) using the GATK model (-GL 2) to infer major and minor alleles (-doMajorMinor 1, -doMaf 1). We restricted this analysis to the 28 chromosome-length autosomes (Table S1) and included only regions with Phred quality and mapping scores exceeding 30. We utilized properly paired (-only_proper_pairs 1) and unique reads (-uniqueOnly 1) while retaining only biallelic sites (-skipTrialleleic 1). Sites with read coverage in less than 60% of the samples were excluded (-minInd 30, allowed missingness 60%), and we retained only polymorphic sites with a genotype likelihood p-value less than 1e-6 (-SNP_pval 1e-6). Lastly, we applied thinning for the variant sites, ensuring a minimum distance of 1Kb between two sites with a custom bash script. Using thinned sites only, we reestimated genotype likelihoods (with the -sites option in ANGSD).

Then, we performed principal component analysis with PCAngsd v1.11 (Meisner & Albrechtsen, 2018) by using default settings where the minor allele frequency cut-off is 5%.

2.1.3 Non-linear dimensionality reduction analyses

We used Jupyter Notebook (Kluyver et al., 2016), Python v3.11.4 (Van Rossum & Drake, 2009), and a set of Python libraries for the t-SNE and UMAP analyses. We decomposed the covariance matrix output of PCAngsd (Meisner & Albrechtsen, 2018) by using NumPy (v1.24.4) (Harris et al., 2020) to obtain eigenvectors and eigenvalues for each dataset. Then, we calculated principal components and performed PCA-t-SNE and PCA-UMAP in libraries scikit-learn (v1.3.0) (Pedregosa et al., 2016) and umap-learn (v0.5.3) (McInnes, Healy, & Melville, 2018), respectively. From this point on, for readability, we will refer to the techniques of PCA-t-SNE and PCA-UMAP as t-SNE and UMAP, respectively.

We optimized hyperparameters for t-SNE and UMAP using grid search to examine dimensionality reduction trends. For t-SNE, we adjusted the perplexity (*perp*), which influences the number of effective neighbors and typically ranges from 5 to 50 (van der Maaten & Hinton, 2008). In UMAP, we changed two hyperparameters: the number of neighbors (*NN*, default 15) and the minimum distance (*MD*, default 0.1) (McInnes, Healy, & Melville, 2018). Both *perp* and *NN* balance local versus global data structure representation, with smaller values focusing on local and larger ones on the global structure. The *MD* parameter controls data point separation, with lower values tightening clustering and higher values dispersing data points for global structure preservation.

We tested various combinations of *perp*, *NN*, and *MD* values across three datasets. For t-SNE, we used *perp* values of 5, 10, and 23 (half the total number of individuals), while for UMAP, we employed *NN* values of 5, 10, and 23, along with *MD* values of 0.01, 0.1 (the default), and 0.5. These parameters were combined with a set of top PCs. For each dataset, we established the set of PCs to be used by determining the minimum as well as the maximum number of available PCs, as suggested by (Diaz-Papkovich et al., 2019). We employed the elbow method to identify the minimum number of PCs using the 'KneeLocator' function from the Python package 'kneed' v0.8.5 (Satopaa et al., 2011) with a polynomial fit method. This technique identifies the ‘elbow point’ by smoothing the PCs’ explained variance curve with polynomial interpolation and selecting the point where the rate of increase significantly diminishes (Figure S1). For our datasets, this approach suggested using between 5 to 46 PCs. For instance, in the SO_6x dataset, we applied 6 (the elbow point) and 46 PCs (the maximum available, as shown in Figure S1A); similarly, for both SO_2x and SO_0.5x datasets, we used 5 and 46 PCs (Figure S1B & C). Employing this method allowed us to systematically assess the impact of different hyperparameters and PC selections on the non-linear dimensionality reduction results, thereby enhancing the robustness and reproducibility of our analyses.

All visuals of dimensionality reduction analyses were plotted with the seaborn library (v0.11.2) (Waskom, 2021). To ensure accessibility in visual data representation, we used a color palette generated by the online tool, which is available at https://medialab.github.io/iwanthue/.

2.2 Case system 2: Galápagos giant tortoises

Galápagos giant tortoises are endemic to the geologically young Galápagos Archipelago (Van Denburgh, 1914). There are 14 named taxa, most of which originate from distinct islands, and each of them is a subject of conservation concern, with listings on the IUCN Red List ranging from Vulnerable to Critically Endangered, and some even declared Extinct (IUCN, 2023). Their population decline during the 19th and 20th centuries can be attributed to various factors, including human exploitation, habitat degradation, and the detrimental influence of invasive species (Hamann, 1993; Pritchard, 1996). While recent taxonomic revisions have merged these tortoises into a single species (Turtle Taxonomy Working Group, 2021), findings from a whole-genome sequencing approach (Gaughran et al., 2023) suggest multiple species within the genus *Chelonoidis*. To minimize confusion, we mostly employed English common names and scientific epithets for each taxon when necessary throughout the text (e.g., *darwini* for Chelonidis darwini / *Chelonidis niger darwini*), following (Gaughran et al., 2023).

Galápagos giant tortoises exhibit a continuum of carapace shapes, spanning from domed forms prevalent in humid high-altitude environments to saddleback variants typically residing in drier, lower elevation habitats; notably, two taxa occupy an intermediate position on this morphological spectrum, known as "semi-saddleback" (Chiari, 2021). A recent genomic investigation aimed to provide genomic evidence for the rediscovery of the Fernandina Island Galápagos giant tortoise (*phantasticus)* and to infer the whole genome phylogeny of extinct and extant 13 Galápagos giant tortoises (Jensen et al., 2022). This study also revealed that the Fernandina tortoises form a monophyletic group, clustering together all lineages exhibiting a saddleback carapace morphology, along with one displaying a semi-saddleback morphology (Jensen et al., 2022). The remaining taxa constituted a distinct group, wherein all lineages characterized by a domed-shaped carapace morphology clustered together along with another semi-saddleback tortoise. The overall clustering pattern within each carapace morphology group appeared to be largely influenced by geographical factors, mainly the islands inhabited by these tortoises (Jensen et al., 2022).

2.2.1 Analysis of whole-genome resequencing data

We acquired whole-genome resequencing data from (Jensen et al., 2021), where demographic history and patterns of molecular evolution of all extant Galápagos giant tortoises were explored. Additionally, we downloaded the resequencing data of the contemporary *phantasticus* sample from (Jensen et al., 2022) (NCBI BioProject PRJNA761229). Overall the acquired dataset comprised three individuals per each of 10 Galápagos giant tortoise taxa, six individuals from *becki* (three from each of the two lineages, PBL and PBR [Garrick et al., 2014]), one from the extinct *abingdonii*, and another one from the recently rediscovered *phantasticus*. In total, the dataset encompasses 38 individuals, of which eight are saddlebacked, 24 are domed, and six are semi-saddlebacked (Table S2, Figure 4A).

To mirror the PCA of (Jensen et al., 2022), with genotype likelihood approaches, we mostly followed our approach in the Scimitar-horned oryx. Namely, adapter trimming, reference genome alignment, indel realignment, and merging of the read groups were performed with ATLAS Pipeline (Link et al., 2017; Marchi et al., 2022; https://atlaswiki.netlify.app/atlas-pipeline). To decrease the computational complexity of downstream analyses caused by high numbers of scaffolds (particularly in indel realignment), we aligned sequencing reads to the chromosome-level assembly of the Aldabra giant tortoise (AldGig_1.0, NCBI GenBank GCA_026122505.1) (Çilingir et al., 2022a) instead of using the Galápagos giant tortoise reference genome (ASM359739v1, NCBI RefSeq GCF_003597395.1) (Quesada et al., 2019). Then, we performed downsampling with ATLAS (Link et al., 2017) and produced three datasets with average sequencing coverages of 8x (close to minimal coverage), 2x, and 0.5x. We named the primary dataset "GT_8x". Here, we kept a single sample with coverage less than ~8x and adjusted the other 37 to approximate the 8x target (Table S2). Next, we created datasets "GT_2x" and "GT_0.5x" by downsampling all sample coverage levels to ~2x and ~0.5x, respectively.

2.2.2 Genotype likelihood estimation and dimensionality reduction analyses

We performed genotype likelihood analysis as described above for the Oryx dataset; we restricted the whole analysis to the 26 pseudo-chromosomes (Table S2) and allowed 40% missingness (-minInd 30).

For the three datasets with varying sequencing coverages, we conducted PCA, UMAP, and t-SNE analyses as described above, but with some adjustments due to differences in the number of individuals and populations sampled. For t-SNE, we applied *perp* values of 3, 5, and 10 (~a quarter of the total number of individuals). For UMAP, we used the *NN* values of 3, 5, and 10; and *MD* values were set at 0.01, 0.1, and 0.5. A *NN* value of 3 instead of 5 was

chosen because the sample size per species was, in most cases, three. We systematically combined these parameter configurations with varying numbers of PCs (minimal number of PCs again assessed using the elbow method as implemented in the function 'KneeLocator' in package kneed): 9 and 38 for the GT_8x dataset; 7 and 38 for GT_2x and; 4 and 38 for GT_0.5x (Figures S2A-C). This approach also mirrors the strategy we employed for the Oryx datasets to determine the appropriate range of PCs.

All intermediate files and codes required to reproduce our analyses are available at https://github.com/fgcilingir/lcUMAPtSNE.

3. Results

3.1 Scimitar-horned oryx

After quality filtering, we retained an average of 99% of the raw sequencing data coming from 46 scimitar-horned oryx individuals. After mapping and deduplication, an average of 81% (range: 77-87%) of these high-quality reads mapped to the reference genome. The approximate individual-level coverage varied from ~4.6x to ~20.6x (Table S1).

SO_6x (dataset with coverage 4.6x-6x) yielded a total of 1,415,641 variant sites with a minor allele frequency (MAF) of greater than 0.05 and a minimal pairwise distance of 1kb. SO_2x yielded 1,516,230, and SO_0.5x yielded 23,129 variant sites with MAF > 0.05 and distance > 1kb. The PCA performed was insensitive to coverage, with all three coverage datasets mirroring the PCA results reported in Humble et al. (2023), where PC1 separated the genetically unmanaged population EAD_B from the rest, PC2 separated the other genetically unmanaged population EAD_A from the cluster of genetically managed populations (EEP + USA) and PC3 separated the genetically managed EEP population from the others (Figures 1A-C, Figures S3A-C). The cumulative variance explained by PC1 and PC2 for the SO_6x, SO_2x, and SO_0.5x datasets were 21.9%, 20.3%, and 19.2%, respectively.

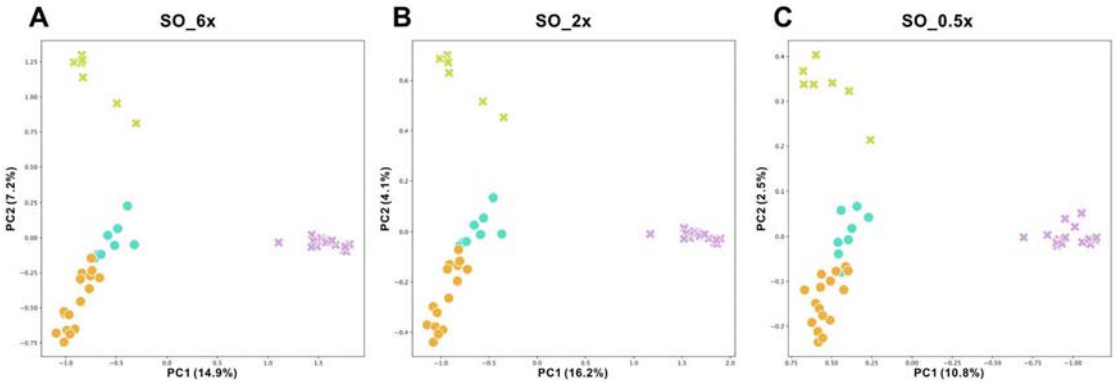

Figure 1 The PCA plot of the SO_6x (A), SO_2x (B), and SO_0.5x (C) datasets using the first two PCs. Each dot represents a single individual. Colors indicate populations as follows: Light green: EAD_A, purple: EAD_B (both genetically unmanaged, depicted with the crosses), turquoise: EEP, and orange: USA (genetically managed, depicted with circles).

The minimum (N=5) and maximum number of PCs (N=46) incorporated in the t-SNE and UMAP analyses of SO_6x, SO_2x, and SO_0.5x corresponded to 35.3-99.8%, 26.6-99.3%, and 24.0-98.8% cumulative variance explained, respectively. To ensure clarity, we present in the main figures results acquired using the mid-range parameters selected for t-SNE and UMAP (*perp* 10, *NN* 10, *MD* 0.1) coupled with the top PCs obtained with the elbow method unless stated otherwise, as these parameters effectively capture the overall trends observed across the hyperparameter space we explored. See the Supplementary Material for a comprehensive overview of the parameter space we explored.

The t-SNE and UMAP analyses of all three datasets of varying coverages confirmed the two distinct groups formed by the genetically unmanaged populations and their separation from the unmanaged populations (Figure 2 & 3; Figures S4-9). Additionally, the SO_6x and SO_2x datasets allowed the differentiation between the two genetically managed, thus outperforming the corresponding top two PC projections (as in PCA) (Figure 2; Figures S4-7). For both datasets, as anticipated, the cohesiveness of the clusters improved with combinations of local parameters. Lowering the *perp*, *NN,* or *MD* values led to clustering at a finer scale, while increasing these values helped to visualize more global patterns, suggesting higher genetic similarity between one unmanaged population (EAD_B) and the two managed populations than each of these with the other genetically unmanaged population (EAD_A) (Figures S4–7). The increase in the number of PCs did not influence the clustering patterns obtained using either technique (Figures S4–7).

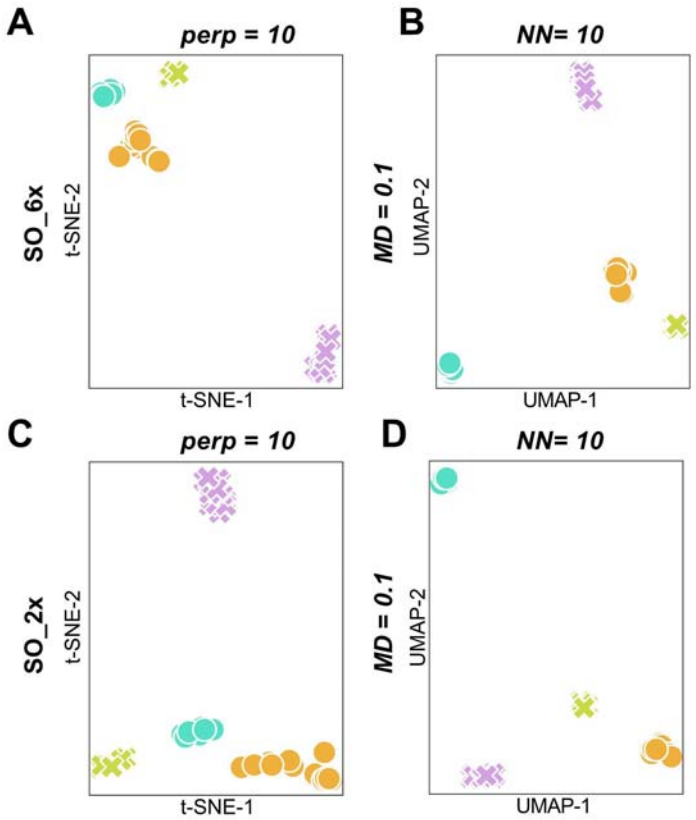

Figure 2 t-SNE and UMAP representations of SO_6x (A, B) with 6 PCs and of SO_2x (C, D) with 5 PCs (minimum number of significant PCs based on elbow method). Each dot represents a single individual. Colors indicate populations as in Figure 1: Light green: EAD_A, purple: EAD_B (both genetically unmanaged, depicted with the crosses), turquoise: EEP, and orange: USA (genetically managed, depicted with circles).

For the SO_0.5x dataset, both t-SNE and UMAP separated the two genetically managed populations (Figure 3A & B; Figure S8 & S9) as opposed to the corresponding PCA (Figure 1C). However, t-SNE analyses provided poorer local structure when compared to UMAP (Figure 3A & C). Notably, when we increased the number of PCs, the clustering efficacy of UMAP was reduced as opposed to the trends observed with SO_6x and SO_2x (Figure 3C & D; Figure S9).

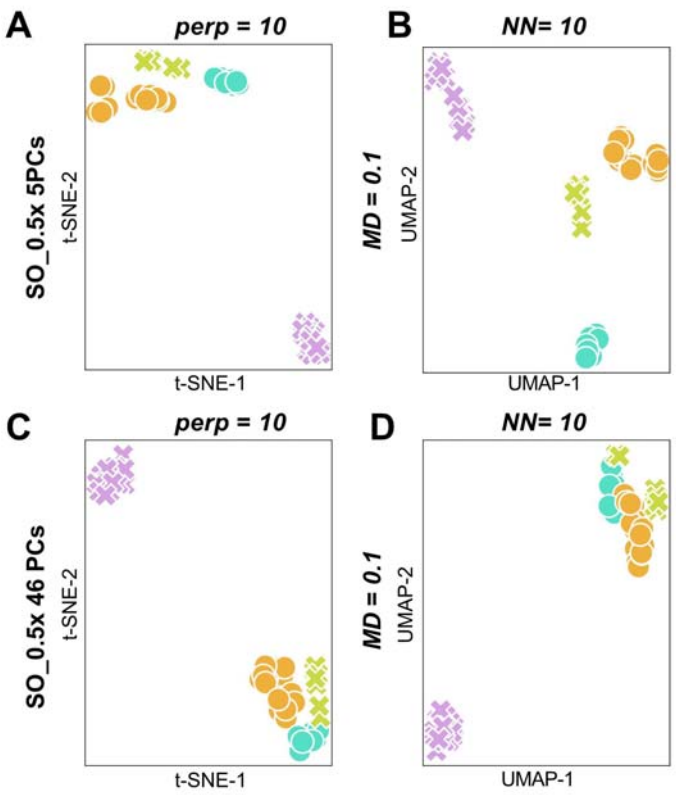

Figure 3 t-SNE and UMAP representations of SO_0.5x with 5 PCs (minimum number of significant PCs, A, B) and with 46 PCs (maximum number of PCs, C. D). Each dot represents a single individual. Colors indicate populations as in Figures 1 and 2: Light green: EAD_A, purple: EAD_B (both genetically unmanaged, depicted with the crosses), turquoise: EEP, and orange: USA (genetically managed, depicted with circles).

3.2 Galápagos giant tortoises

After processing the raw sequencing data of 38 Galápagos giant tortoise individuals to remove adapters and improve quality, 99.97% of the initial data was preserved. Post mapping and deduplication, about 79.5% (with a range of 57.4-84.2%) of these reads aligned to the reference genome. Individual coverage varied between ~2.6x and ~40.2x (Table S2).

GT_8x yielded a total of 854,967 variant sites with a minor allele frequency (MAF) greater than 0.05. GT_2x yielded 957,108, and GT_0.5x yielded 14,219 variant sites with MAF > 0.05. The PCA performed with GT_8x and GT_2x mirrored the PCA results reported by Jensen et al. (2022), where PC1 mostly separated the Galápagos giant tortoise taxa from each other, and PC2 improved the distinction mainly by separating two taxa from the rest (Figure 4B & C). One point of difference was *darwini* individuals clustered with *becki-PBR*

instead of *becki-PBL*; however, this was also the case in the phylogenetic trees reported in the same study (Jensen et al., 2022) and in a more recent species delimitation study (Gaughran et al., 2023). On the contrary, GT_0.5x showed remarkably decreased clustering power, revealing three loose clusters, separating Isabela Island taxa into two groups and grouping Santa Cruz taxa with all saddlebacked and one semi-saddlebacked taxa (Figure 4D).

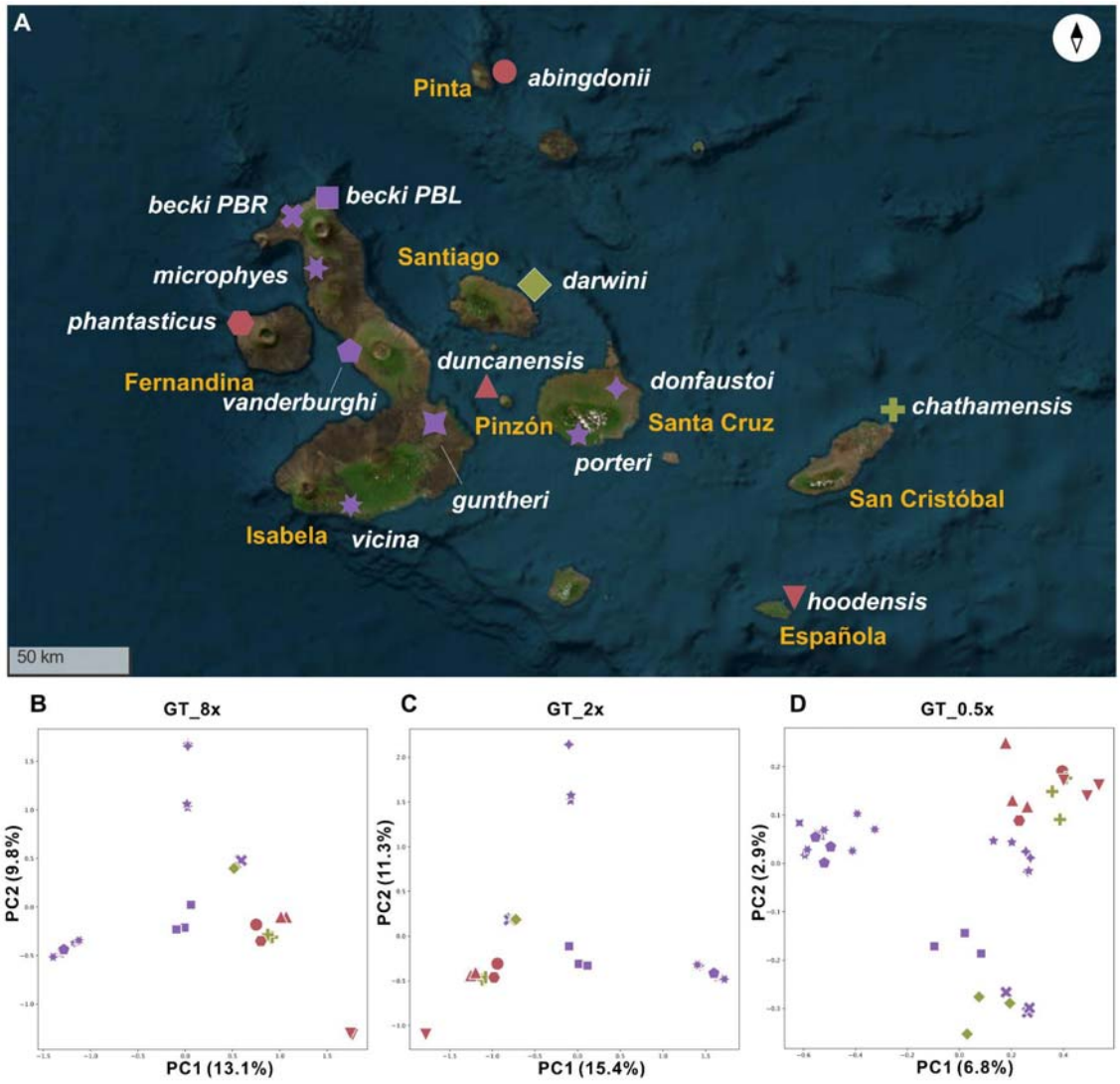

Figure 4 A) Map of Galápagos Archipelago showing the locations of each Galápagos giant tortoise lineage in the sample set used in this study (adapted from Jensen et al. 2021; 2022). Island and taxon names are colored yellow and white, respectively. B) PCA plot of the GT_8x, (C) GT_2x, and (D) GT_0.5x datasets using the first two PCs. Each marker represents a single individual. Each shape represents a single taxon, and the color depicts the shell shape: red for saddlebacked, green semi-saddlebacked, and purple for domed-shell lineages.

The minimum (N=7) and maximum number of PCs (N=38) incorporated in the t-SNE and UMAP analyses of GT_8x, GT_2x, and GT_0.5x corresponded to 35.3-99.8%, 26.6-99.3%, and 24.0-98.8% cumulative variance explained, respectively. When applying t-SNE and UMAP to the GT_8x and GT_2x with local parameter settings (with lower *perp*, *MD*, *NN* values), UMAP was able to discern all taxa except for *abingdonii*, *phantasticus,* and *chathamensis* (Figure 5A & B; Figures S10-13). Note that the former two have only one sample each in our sample set, and also, the PCA was unable to discern *phantasticus* and *chathamensis* clearly. On the other hand, t-SNE, with the most local parameter setting (*perp* 3), was less efficient in separating the different taxa (Figure 5C & D; Figures S10-13). Increasing the number of PCs typically did not have a significant effect on the form of clusters but their relative position to each other (Figures S10-13).

When we allowed the t-SNE and UMAP analyses to capture more global clustering patterns (with higher *perp*, *MD*, *NN* values) in GT_8x (Figure 5E & G) and GT_2x (Figure 5F & H), Santa Cruz taxa and all Isabela Island taxa except both lineages of *becki* formed two distinct clusters. Interestingly, the third cluster consisted of all domed-shell, two semi-saddlebacked, and two lineages of *becki,* a group of taxa that was also discerned from the other dome-shelled taxa in the latest phylogenomics study using all extant Galápagos giant tortoise taxa (Gaughran et al., 2023).

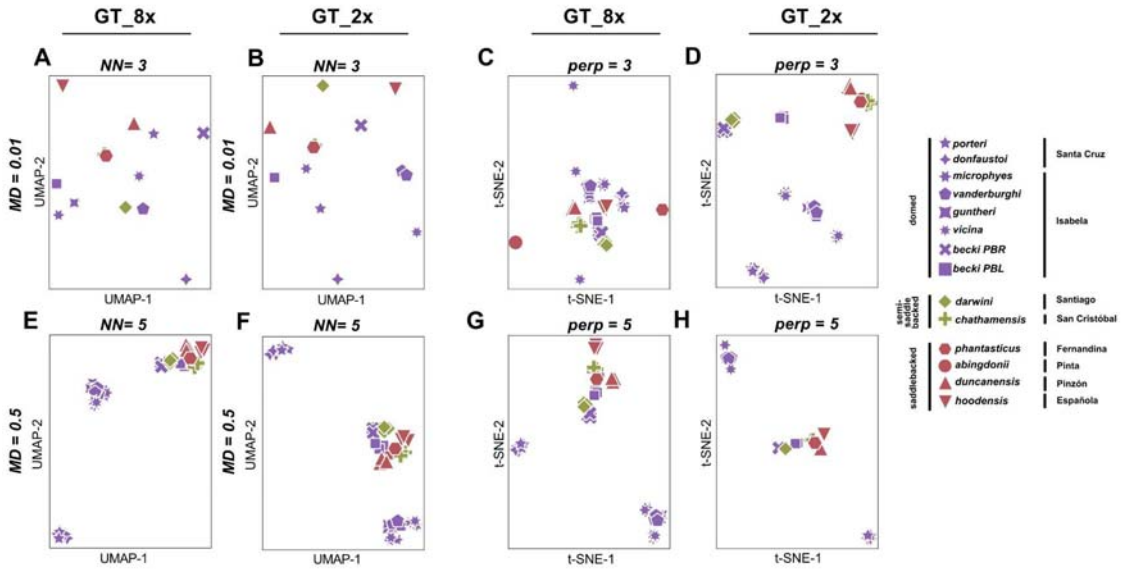

Figure 5 UMAP (A, B, E, F) and t-SNE analyses (C, D, G, H) of GT_8x and GT_2x emphasizing the most local (top row) and more global (bottom row) parameter settings in the hyperparameter space explored.

Our UMAP and t-SNE analyses with GT_0.5x showed a loss of local resolution compared to GT_8x and GT_2x (Figure 6A & B; Figure S14 & 15). However, the results obtained with global parameter settings were still comparable to the PCAs with 8x and 2x, suggesting genetic similarities among the domed shelled taxa on Isabella Island (except *becki*, which grouped with *darwini*), among the two domed shelled taxa on Santa Cruz and among the saddleback taxa together with the semi-saddlebacked *chathamensis* (Figure 6C & D). In particular, both t-SNE and UMAP successfully distinguished the Santa Cruz taxa from others, a task where the PCA with 0.5x coverage failed (Figure 6 & 4D; Figure S14). Overall, t-SNE and UMAP outperformed the PCA at 0.5x coverage and allowed comparable resolution as the PCAs of higher coverage. Also, as opposed to our findings with GT_8x and GT_2x, in all t-SNE trials, adding more PCs into the analyses led to a tendency of over-clustering the groups that were separated when fewer PCs were used, while UMAP primarily exhibited a similar pattern with global settings (Figure S15).

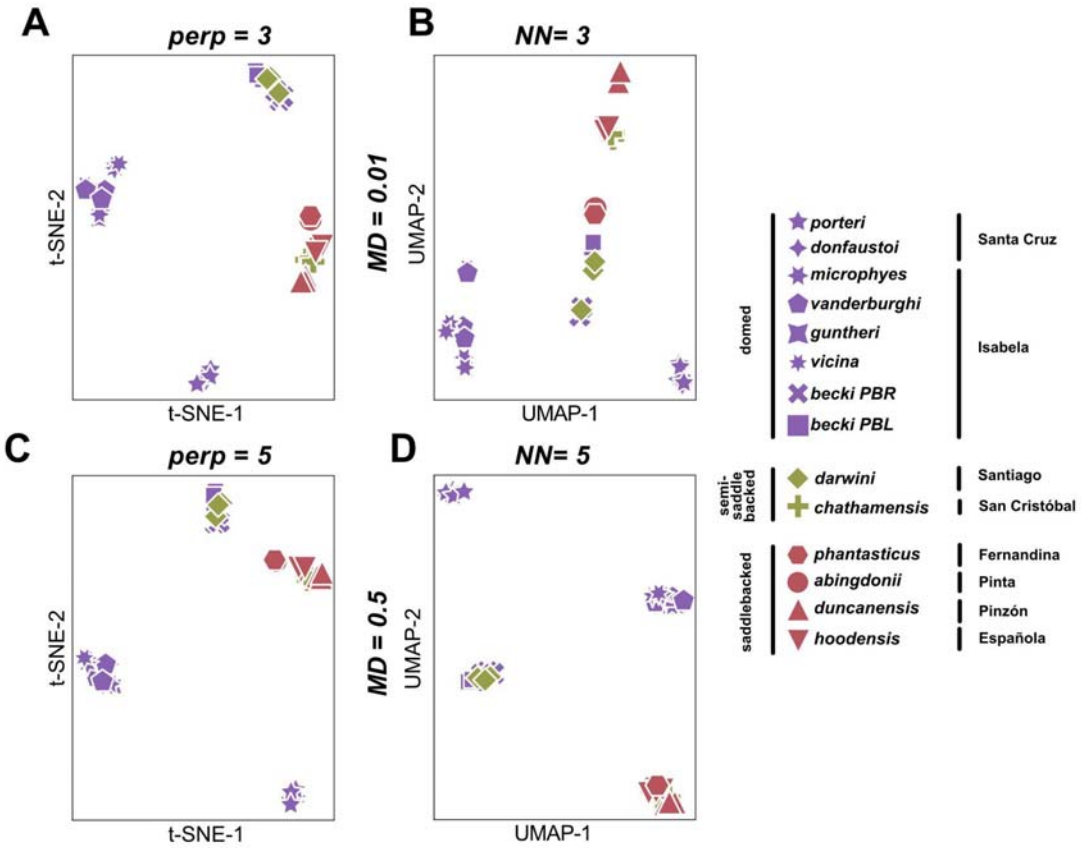

Figure 6 t-SNE (A, C) and UMAP projections (B, D) of GT_0.5x focusing on emphasizing local clustering patterns (top row) and glocal patterns (bottom row).

3.3 Overview of the online guidelines

Our online guideline initiates with the process of estimating genotype likelihoods and conducting PCA for the SO_2x dataset. This step incorporates the generation of the corresponding scree plots to facilitate the visualization of the cumulative explained variation by each PC of the data. It progresses to detailing the rationale behind selecting specific hyperparameters for this dataset (e.g., considering the number of individuals in the populations, etc.) in conjunction with determining the number of PCs derived from the scree plot. To aid practical application, we offer several Jupyter notebooks enabling users to input their covariance matrix, compute principal components, and subsequently apply t-SNE and UMAP techniques using pre-determined hyperparameters. These notebooks were crafted to manage multiple input files simultaneously, streamlining the analysis of diverse datasets with varied parameter configurations, thereby enhancing efficiency in handling extensive genomic datasets.

4. Discussion

In the context of conservation genetics, analyzing the genetic structure of populations is crucial for understanding how populations are interconnected, what their demographic histories looked like, and what genotype-phenotype associations may exist. This knowledge is essential for developing effective conservation strategies, managing genetic diversity, and identifying adaptive traits essential for species survival. High-coverage sequencing, while comprehensive, can be prohibitively expensive for such studies. Simulation studies have shown that low-coverage sequencing across a larger number of individuals can yield more accurate estimates of many population parameters, despite genotype uncertainty (Alex Buerkle & Gompert, 2013; Fumagalli, 2013). This approach not only reduces sequencing costs but also allows for more representative sampling; thereby, genotype likelihood-based analyses have been increasingly used in conservation genomics projects to address questions related to population genetic structure, demographic and evolutionary history, phylogeny/species delimitation, mutation load, and the genomic basis of trait variation (e.g., giant water lily [Smith et al., 2022], honey bees [Qiu et al., 2023] reef-building corals [Cooke et al., 2020], Pacific salmon [Prince et al., 2017], Aldabra giant tortoises [Çilingir et al., 2022a; 2022b], setophaga warblers [Baiz et al., 2021], Seychelles paradise flycather brown hyenas [Westbury et al., 2018], polar bears [Liu et al., 2014], muskox [Pečnerová et al., 2024], African and Asiatic cheetahs [Prost et al., 2022], Eurasian lynx [Mueller et al., 2022]).

PCA is widely recognized as a fundamental tool for analyzing and visualizing population genetic structure. Its utility extends beyond direct analysis, as PCA often serves as a preparatory step for non-linear dimensionality reduction techniques, such as initialization of t-SNE and UMAP (Kobak & Linderman, 2021), by denoising data and lowering computational

burden (Diaz-Papkovich et al., 2019). While the application of these techniques has predominantly been explored within the context of single-cell transcriptomics (Becht et al., 2018; Kobak & Berens, 2019) and human population genetic structure analyses (Diaz-Papkovich et al., 2020), their use in the study of wildlife species remains limited. Remarkably, there is a lack of research that leverages these techniques in conjunction with genotype likelihoods derived from low to medium-coverage sequencing data, particularly in the context of conservation genomics.

In this study, we focused on two case studies: the endangered scimitar oryx and the Galápagos giant tortoises, which face varying degrees of threat. Previous work on these species combined population structure analysis with mutation load analysis (Humble et al., 2023) and phylogenetic techniques (Gaughran et al., 2023; Jensen et al., 2022), yielding insights critical for their conservation. Building on these biologically relevant foundational findings, our project investigated the effectiveness of UMAP and t-SNE in analyzing low-coverage genetic data. By comparing our results with those of earlier studies, we assessed the biological relevance of the non-linear dimensionality reduction outcomes for conservation purposes. We showed that t-SNE and UMAP outperform PCA in terms of population/species discernment both at medium and low coverage. In the case of the Oryx data, UMAP and t-SNE discerned genetically managed populations with all three tested coverages (6x, 2x, 0.5x), outperforming their PCA. For the Galápagos data at medium coverage, t-SNE and UMAP were able to clearly discern taxa, which were more weakly clustered by PCA. At low coverage (0.5x), both t-SNE and UMAP were again able to separate taxa, resulting in a visualization comparable to PCAs of higher coverage for both systems, while the PCA allowed only poor resolution (Figure 4D). More global parameters revealed clustering that could be explained partly by geography and partly by phenotype, but the scope of this work was not to reinterpret previous studies on Galápagos phylogeny but to show the potential of t-SNE and UMAP.

In the case of the oryx example, it may be argued that plotting PC3 allowed the same additional resolution; hence, the added benefit of using t-SNE or UMAP may be questioned. It is indeed also possible to apply 3D PCAs, plot PC1 vs PC3 and PC4, or perform local PCAs (e.g., [Manjón et al., 2013]) to have a more resolved picture. However, 3D plots are often difficult to read, and the two methods described here provide more resolution beyond PC4, all summarized in one plot, and grid search allows exploring the clustering patterns. We would like to point out that t-SNE and UMAP are not always performing better than PCA, especially when the first two principal components explain a large proportion of variance (Diaz-Papkovich et al., 2019). We further observed that the relative positions of clusters vary

with parameters, and this, in the case of local parametrization, may not be a good proxy for genetic distance. However, it is essential to recognize that genomic data often have only a small proportion of variance explained by the first principal components, making the ability of t-SNE and UMAP to incorporate a broader variance spectrum and thus enhance resolution at lower coverage which could be considered as an advantage of these approaches.

Importance of parameter optimization/grid search and recommendations

Previous studies have highlighted default values for parameters such as *perp*, *NN*, and *MD*, which were effective in analyses involving human datasets worked well for their datasets, but most of these studies were working with human datasets characterized by large sample sizes (Diaz-Papkovich et al., 2019, 2020). However, in conservation genomics studies, sample sizes are generally much smaller, necessitating a tailored approach in using t-SNE and UMAP techniques. Recognizing this need, we offer a Jupyter Notebook to assist in conducting a grid search, recommending a systematic exploration of parameter ranges, as we have done in our study. This exploration is crucial for understanding the impact of local (e.g., a low number of neighbors, lower perplexity) versus global (e.g., a high number of neighbors, higher perplexity) parameter settings, each offering unique insights into the data structure. Furthermore, while previous studies advise using as many PCs as computational resources allow (Diaz-Papkovich et al., 2019, 2020), we found that the top critical number of PCs provided consistent results in our analysis, particularly for the 0.5x datasets, where adding more PCs even included noise (Figure S9 & S15), likely due to the added uncertainty of genotype likelihoods with 0.5x (Meisner & Albrechtsen, 2018). Therefore, we also recommend experimenting with a range of PCs and employing the elbow method as a preliminary step to determine the most effective number of PCs for capturing the essential variance without incorporating excess noise, thus optimizing the analysis for conservation genomics studies with varying data characteristics.

Limitations

An important caveat to note is the necessity of maintaining focus on the biological relevance of clustering results while cautioning against the potential for "over-separation" of genetic groups. Techniques such as UMAP and t-SNE try to find a lower-dimensional representation that preserves the distances between the points in the neighborhood; in other words, they are prone to highlighting similarities while exaggerating differences, which may lead to an “over-separation” of genetically similar individuals. Accordingly, it is essential to consider the biological relevance of the observed clustering patterns. Also, when population structure is detected, a thorough evaluation of the influences of genetic drift versus local adaptation is crucial for deriving appropriate conservation strategies. Additionally, although PCA

initialization may provide insights into the global structure of clusters, the interpretation of global structure with these non-linear techniques remains highly sensitive to parameter adjustments, rendering distances between groups not directly interpretable. Finally, the inherent limitations of PCA, a linear dimensionality reduction method, when used in conjunction with non-linear methods, may sometimes exacerbate data distortion (Chari & Pachter, 2023). An alternative approach may involve directly employing a distance matrix based on genotype likelihoods rather than relying on PCA. The Jupyter notebook that we provided is designed for flexible adaptation to incorporate such methodologies, enhancing the analysis of population genetic structure.

Conclusion and general implications

In our study, we explored the potential of non-linear dimensionality reduction techniques for exploring population structure from low-coverage genomic data based on genotype likelihoods. Our findings reveal that these methods can outperform the classical PCA in revealing subtle genetic structure, particularly when low read depth results in the loss of local information in classical PCAs. Our results demonstrate that t-SNE and UMAP are valuable supplements to PCAs rather than substitutes. By gaining resolution in low-coverage studies, these two techniques can lead to cost and computational savings or support the analysis of larger sample sizes, thereby increasing the information content (Alex Buerkle & Gompert, 2013; Fumagalli, 2013). However, it is crucial to recognize that both linear and non-linear approaches to dimensionality reduction can result in data distortion, each presenting unique benefits and challenges. As such, careful handling of data and biologically relevant interpretation of the results are imperative.

Acknowledgments

We would like to thank all WSL Ecological Genetics Group members who provided their comments on the manuscript. We extend our gratitudes to the authors who have made their genomic datasets available for open access. Specifically, we thank Evelyn Jensen et al. for the Galápagos giant tortoise dataset, Emily Humble et al. for the scimitar-horned oryx dataset, and all others who contributed to the resources that significantly supported this research. F. Gözde Çilingir was funded by the University of Zurich Postdoc Grant #FK-22-109.

References

Alanis-Lobato, G., Cannistraci, C. V., Eriksson, A., Manica, A., & Ravasi, T. (2015). Highlighting nonlinear patterns in population genetics datasets. *Scientific Reports*, *5*,

8140.

Alex Buerkle, C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, *22*(11), 3028–3035.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92.

Anopheles gambiae 1000 Genomes Consortium. (2020). Genome variation and population structure among 1142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*. *Genome Research*, 30(10), 1533–1546.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, *3*(10), e3376.

Baiz, M. D., Wood, A. W., Brelsford, A., Lovette, I. J., & Toews, D. P. L. (2021). Pigmentation genes show evidence of repeated divergence and multiple bouts of introgression in Setophaga warblers. *Current Biology: CB*, *31*(3), 643–649.e3.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*. https://doi.org/10.1038/nbt.4314

Bertram, B. C. R. (1988). *Conservation and Biology of Desert Antelopes* (A. Dixon & D. M. Jones (Eds.); pp. 136–145). Christopher Helm.

Černý, V., Priehodová, E., & Fortes-Lima, C. (2023). A population genetic perspective on subsistence systems in the Sahel/Savannah belt of Africa and the historical role of pastoralism. *Genes*, *14*(3). https://doi.org/10.3390/genes14030758

Chari, T., & Pachter, L. (2023). The specious art of single-cell genomics. *PLoS Computational Biology*, *19*(8), e1011288.

Chiari, Y. (2021). Morphology. In James P. Gibbs, Linda J. Cayot, Washington Tapia Aquilera (Ed.), *Galapagos giant tortoises*. Academic Press.

Chyleński, M., Ehler, E., Somel, M., Yaka, R., Krzewińska, M., Dabert, M., Juras, A., & Marciniak, A. (2019). Ancient mitochondrial genomes reveal the absence of maternal kinship in the burials of Çatalhöyük people and their genetic affinities. *Genes*, *10*(3). https://doi.org/10.3390/genes10030207

Çilingir, F. G., A'Bear, L., Hansen, D., Davis, L. R., Bunbury, N., Ozgul, A., Croll, D., & Grossen, C. (2022a). Chromosome-level genome assembly for the Aldabra giant tortoise enables insights into the genetic health of a threatened population. *GigaScience*, *11*, giac090.

Çilingir, F. G., Hansen, D., Bunbury, N., Postma, E., Baxter, R., Turnbull, L., Ozgul, A., & Grossen, C. (2022b). Low-coverage reduced representation sequencing reveals subtle

within-island genetic structure in Aldabra giant tortoises. *Ecology and Evolution*, *12*(3), e8739.

Cooke, I., Ying, H., Forêt, S., Bongaerts, P., Strugnell, J. M., Simakov, O., Zhang, J., Field, M. A., Rodriguez-Lanetty, M., Bell, S. C., Bourne, D. G., van Oppen, M. J., Ragan, M. A., & Miller, D. J. (2020). Genomic signatures in the coral holobiont reveal host adaptations driven by Holocene climate change and reef specific symbionts. *Science Advances*, *6*(48). https://doi.org/10.1126/sciadv.abc6318

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*, *15*(11), e1008432.

Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2020). A review of UMAP in population genetics. *Journal of Human Genetics*, *66*(1), 85–91.

Dixon, A. M., Mace, G. M., Newby, J. E., & Olney, P. J. S. (1991). Planning for the re-introduction of scimitar-horned oryx (*Oryx dammah*) and addax (*Addax nasomaculatus*) into Niger. *Symposia of the Zoological Society of London*, 62:201–216.

Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, *11*(1), 1537.

Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, *35*(19), 3855–3856.

Frankham, R. (2003). Genetics and conservation biology. *Comptes Rendus Biologies*, *326 Suppl 1*, 22–29.

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PloS One*, *8*(11), e79667.

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* , *30*(10), 1486–1487.

Fu, R., Zhu, Y., Liu, Y., Feng, Y., Lu, R.-S., Li, Y., Li, P., Kremer, A., Lascoux, M., & Chen, J. (2022). Genome-wide analyses of introgression between two sympatric Asian oak species. *Nature Ecology & Evolution*, *6*(7), 924–935.

Garrick, R. C., Benavides, E., Russello, M. A., Hyseni, C., Edwards, D. L., Gibbs, J. P., Tapia, W., Ciofi, C., & Caccone, A. (2014). Lineage fusion in Galápagos giant tortoises. *Molecular Ecology*, *23*(21), 5276–5290.

Gaspar, H. A., & Breen, G. (2019). Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. *BMC Bioinformatics*, *20*(1), 116.

Gaughran, S. J., Gray, R., Jones, M., Fusco, N., Ochoa, A., Miller, J. M., Poulakakis, N., de

Queiroz, K., Caccone, A., & Jensen, E. L. (2023). Whole-genome sequencing confirms multiple species of Galapagos giant tortoises. *bioRxiv*, https://doi.org/10.1101/2023.04.05.535692

Gilbert, T. (2019). *International studbook for the scimitar-horned oryx Oryx dammah*. Marwell Wildlife, Winchester.

Halldorsson, B. V., Eggertsson, H. P., Moore, K. H. S. et. al. (2022). The sequences of 150,119 genomes in the UK Biobank. *Nature*, *607*(7920), 732–740.

Hamann, O. (1993). On vegetation recovery, goats and giant tortoises on Pinta Island, Galápagos, Ecuador. *Biodiversity & Conservation*, *2*(2), 138–151.

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Molecular Ecology*, *30*(1), 62–82.

Humble, E., Dobrynin, P., Senn, H., Chuven, J., Scott, A. F., Mohr, D. W., Dudchenko, O., Omer, A. D., Colaric, Z., Lieberman Aiden, E., Al Dhaheri, S. S., Wildt, D., Oliaji, S., Tamazian, G., Pukazhenthi, B., Ogden, R., & Koepfli, K.-P. (2020). Chromosomal-level genome assembly of the scimitar-horned oryx: Insights into diversity and demography of a species extinct in the wild. *Molecular Ecology Resources*, *20*(6), 1668–1681.

Humble, E., Stoffel, M. A., Dicks, K., Ball, A. D., Gooley, R. M., Chuven, J., Pusey, R., Remeithi, M. A., Koepfli, K.-P., Pukazhenthi, B., Senn, H., & Ogden, R. (2023). Conservation management strategy impacts inbreeding and mutation load in scimitar-horned oryx. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(18), e2210756120.

IUCN. (2023). *The IUCN Red List of Threatened Species. Version 2023-1*. Accessed https://www.iucnredlist.org.

IUCN SSC Antelope Specialist Group. (2016). *Oryx dammah*. The IUCN Red List of Threatened Species 2016: e.T15568A50191470. https://dx.doi.org/10.2305/IUCN.UK.2016-2.RLTS.T15568A50191470.en

IUCN SSC Antelope Specialist Group. (2023). *Oryx dammah. The IUCN Red List of Threatened Species 2023: e.T15568A197393805*. IUCN Red List. https://dx.doi.org/10.2305/IUCN.UK.2023-1.RLTS.T15568A197393805.en

Jensen, E. L., Gaughran, S. J., Fusco, N. A., Poulakakis, N., Tapia, W., Sevilla, C., Málaga, J., Mariani, C., Gibbs, J. P., & Caccone, A. (2022). The Galapagos giant tortoise Chelonoidis phantasticus is not extinct. *Communications Biology*, *5*(1), 546.

Jensen, E. L., Gaughran, S. J., Garrick, R. C., Russello, M. A., & Caccone, A. (2021). Demographic history and patterns of molecular evolution from whole genome sequencing in the radiation of Galapagos giant tortoises. *Molecular Ecology*, *30*(23),

6325–6339.

Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Publisher Correction: Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews. Genetics*, *20*(5), 310.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks -- a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90). IOS Press.

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, *10*(1), 5416.

Kobak, D., & Linderman, G. C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, *39*(2), 156–157.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356.

Krueger, F. (2016). *TrimGalore*. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , *25*(14), 1754–1760.

Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., & Wegmann, D. (2017). ATLAS: Analysis tools for low-depth and ancient samples. *bioRxiv*. https://doi.org/10.1101/105346

Liu, S., Lorenzen, E. D., Fumagalli, M. et al. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, *157*(4), 785–794.

Li, W., Cerise, J. E., Yang, Y., & Han, H. (2017). Application of t-SNE to human genetic data. *Journal of Bioinformatics and Computational Biology*, *15*(4), 1750017.

Lou, R. N., Jacobs, A., Wilder, A., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology Resources*, *30*, 5966–5993.

Manjón, J. V., Coupé, P., Concha, L., Buades, A., Collins, D. L., & Robles, M. (2013). Diffusion weighted image denoising using overcomplete local PCA. *PloS One*, *8*(9), e73021.

Marchi, N., Winkelbach, L., Schulz, I. et al. (2022). The genomic origins of the world's first farmers. *Cell*, *185*(11), 1842–1859.e18.

Margaryan, A., Lawson, D. J., Sikora, M. et al. (2020). Population genomics of the Viking world. *Nature*, *585*(7825), 390–396.

Ma, Y., Tian, J., Chen, Y., Chen, M., Liu, Y., & Wei, A. (2021). Volatile oil profile of prickly

ash (*Zanthoxylum*) pericarps from different locations in China. *Foods (Basel, Switzerland)*, *10*(10), 2386.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*. http://arxiv.org/abs/1802.03426

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731.

Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science*, *201*(4358), 786–792.

Mueller, S. A., Prost, S., Anders, O., Breitenmoser-Würsten, C., Kleven, O., Klinga, P., Konec, M., Kopatz, A., Krojerová-Prokešová, J., Middelhoff, T. L., Obexer-Ruff, G., Reiners, T. E., Schmidt, K., Sindičič, M., Skrbinšek, T., Tám, B., Saveljev, A. P., Naranbaatar, G., & Nowak, C. (2022). Genome-wide diversity loss in reintroduced Eurasian lynx populations urges immediate conservation management. *Biological Conservation*, *266*, 109442.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PloS One*, *7*(7), e37558.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–451.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190.

Pečnerová, P., Lord, E., Garcia-Erill, G. et al. (2024). Population genomics of the muskox' resilience in the near absence of genetic variation. *Molecular Ecology*, *33*(2), e17205.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2016). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*.

Platzer, A. (2013). Visualization of SNPs with t-SNE. *PloS One*, *8*(2), e56883.

Prince, D. J., O'Rourke, S. M., Thompson, T. Q., Ali, O. A., Lyman, H. S., Saglam, I. K., Hotaling, T. J., Spidle, A. P., & Miller, M. R. (2017). The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. *Science Advances*, *3*(8), e1603198.

Pritchard, P. C. H. (1996). The Galápagos Tortoises: Nomenclatural and Survival Status.

Chelonian Research Monographs, *1*, 1–85.

Prost, S., Machado, A. P., Zumbroich, J., Preier, L., Mahtani-Williams, S., Meissner, R., Guschanski, K., Brealey, J. C., Fernandes, C. R., Vercammen, P., Hunter, L. T. B., Abramov, A. V., Plasil, M., Horin, P., Godsall-Bottriell, L., Bottriell, P., Dalton, D. L., Kotze, A., & Burger, P. A. (2022). Genomic analyses show extremely perilous conservation status of African and Asiatic cheetahs (Acinonyx jubatus). *Molecular Ecology*, *31*(16), 4208–4223.

Qiu, L., Dong, J., Li, X., Parey, S. H., Tan, K., Orr, M., Majeed, A., Zhang, X., Luo, S., Zhou, X., Zhu, C., Ji, T., Niu, Q., Liu, S., & Zhou, X. (2023). Defining honeybee subspecies in an evolutionary context warrants strategized conservation. *Zoological Research*, *44*(3), 483–493.

Quesada, V., Freitas-Rodríguez, S., Miller, J. et al. (2019). Giant tortoise genomes provide insights into longevity and age-related disease. *Nature Ecology & Evolution*, *3*(1), 87–95.

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. *2011 31st International Conference on Distributed Computing Systems Workshops*, 166–171.

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature Reviews. Genetics*, *15*(11), 749–763.

Schmidt, T. L., Chung, J., Honnen, A.-C., Weeks, A. R., & Hoffmann, A. A. (2020). Population genomics of two invasive mosquitoes (*Aedes aegypti* and *Aedes albopictus*) from the Indo-Pacific. *PLoS Neglected Tropical Diseases*, *14*(7), e0008463.

Schmidt, T. L., Swan, T., Chung, J., Karl, S., Demok, S., Yang, Q., Field, M. A., Muzari, M. O., Ehlers, G., Brugh, M., Bellwood, R., Horne, P., Burkot, T. R., Ritchie, S., & Hoffmann, A. A. (2021). Spatial population genomics of a recent mosquito invasion. *Molecular Ecology*, *30*(5), 1174–1189.

Sengupta, D., Choudhury, A., Fortes-Lima, C., Aron, S., Whitelaw, G., Bostoen, K., Gunnink, H., Chousou-Polydouri, N., Delius, P., Tollman, S., Gómez-Olivé, F. X., Norris, S., Mashinya, F., Alberts, M., AWI-Gen Study, H3Africa Consortium, Hazelhurst, S., Schlebusch, C. M., & Ramsay, M. (2021). Genetic substructure and complex demographic history of South African Bantu speakers. *Nature Communications*, *12*(1), 2080.

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C. et al. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, *30*(2), 78–87.

Simon, A., Fraïsse, C., El Ayari, T., Liautard-Haag, C., Strelkov, P., Welch, J. J., & Bierne, N.

(2021). How do species barriers decay? Concordance and local introgression in mosaic hybrid zones of mussels. *Journal of Evolutionary Biology*, *34*(1), 208–223.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693–702.

Smith, L. T., Magdalena, C., Przelomska, N. A. S., Pérez-Escobar, O. A., Melgar-Gómez, D. G., Beck, S., Negrão, R., Mian, S., Leitch, I. J., Dodsworth, S., Maurin, O., Ribero-Guardia, G., Salazar, C. D., Gutierrez-Sibauty, G., Antonelli, A., & Monro, A. K. (2022). Revised species delimitation in the giant water lily genus *Victoria* (*Nymphaeaceae*) confirms a new species and has Implications for its conservation. *Frontiers in Plant Science*, *13*, 883151.

Sohail, M., Palma-Martínez, M. J., Chong, A. Y. et a. (2023). Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature*, *622*(7984), 775–783.

Turtle Taxonomy Working Group, [Rhodin, A. G. J., Iverson, J. B., Bour, R., Fritz, U., Georges, A., Bradley Shaffer, H., & van Dijk, P. P.] (2021). *Turtles of the World: Annotated Checklist and Atlas of Taxonomy, Synonymy, Distribution, and Conservation Status (9th Ed.). In* van Dijk, P. P. , Stanford, C. B. , Goode, E. V. , Buhlmann, K. A. , and Mittermeier, R. A., Rhodin, A. G. J. , Iverson, J. B. (eds). *Chelonian Research Monographs*, 8, 472.

Van Denburgh, J. (1914). *The Gigantic Land Tortoises of the Galapagos Archipelago*. The Academy.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research: JMLR*, *9*, 2579–2605.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual: (Python Documentation Manual Part 2)*. CreateSpace Independent Publishing Platform.

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Westbury, M. V., Hartmann, S., Barlow, A., Wiesel, I., Leo, V., Welch, R., Parker, D. M., Sicks, F., Ludwig, A., Dalén, L., & Hofreiter, M. (2018). Extended and continuous decline in effective population size results in low genomic diversity in the world's rarest hyena species, the brown hyena. *Molecular Biology and Evolution*, *35*(5), 1225–1237.

Data accessibility statement

All raw sequencing data we used in this study were downloaded from public databases, and no new data were generated. All intermediate files we produced using these datasets, all bioinformatic codes used for generating the results we presented, and guidelines are available at https://github.com/fgcilingir/lcUMAPtSNE.

Benefit-sharing statement

Benefits Generated: Benefits from this research accrue from the sharing of the intermediate files we produced by using publicly available sequencing data, our results on these, and the codes and guidelines we provide as described above.

Author contributions

F. Gözde Çilingir conceived the project idea and designed the study with input from Christine Grossen and Kerem Uzel. F. Gözde Çilingir and Kerem Uzel performed the bioinformatic analyses and prepared the online guidelines. F. Gözde Çilingir wrote the manuscript with substantial input from Christine Grossen. All authors reviewed the manuscript.