Bioinformatics, 2025, **41(3)**, btaf105
https://doi.org/10.1093/bioinformatics/btaf105
Advance Access Publication Date: 11 March 2025
Applications Note

Genome analysis

PopGLen—a Snakemake pipeline for performing population genomic analyses using genotype likelihood-based methods

Zachary J. Nolen[1,*]

[1]Department of Biology, Lund University, Lund 22 362, Sweden

*Corresponding author. Department of Biology, Lund University, Kontaktvägen 10, Lund 22 362, Sweden. E-mail: zachary.nolen@biol.lu.se.

Associate Editor: Peter Robinson

Abstract

Summary: PopGLen is a Snakemake workflow for performing population genomic analyses within a genotype-likelihood framework, integrating steps for raw sequence processing of both historical and modern DNA, quality control, multiple filtering schemes, and population genomic analysis. Currently, the population genomic analyses included allow for estimating linkage disequilibrium, kinship, genetic diversity, genetic differentiation, population structure, inbreeding, and allele frequencies. Through Snakemake, it is highly scalable, and all steps of the workflow are automated, with results compiled into an HTML report. PopGLen provides an efficient, customizable, and reproducible option for analyzing population genomic datasets across a wide variety of organisms.

Availability and implementation: PopGLen is available under GPLv3 with code, documentation, and a tutorial at https://github.com/zjnolen/PopGLen. An example HTML report using the tutorial dataset is included in the Supplementary Material.

1 Introduction

Genomic resources are rapidly becoming available for non-model organisms due to efforts to generate reference genomes for the majority of the planet's species. These genomes make population genomic analyses possible for many wild species for the first time, addressing an increasing need for genetic estimates to address research questions and inform conservation action. However, the data types that go into these studies are not always suited for traditional genotype call based population genomics, in particular for studies with low sequencing depth (Lou *et al.* 2021), such as those utilizing degraded DNA from historical samples (Díez-del-Molino *et al.* 2018). Genotype likelihood-based methods are especially suited for these data types, utilizing a probabilistic framework that incorporates the uncertainty in base calling (Nielsen *et al.* 2011).

Here, I describe PopGLen, a pipeline I have developed to be an all-inclusive workflow that includes raw data processing and mapping, quality control, multiple filtering schemes, and a wide breadth of genotype likelihood-based analyses (Fig. 1). It is developed using the Snakemake workflow manager (Mölder *et al.* 2021), enabling it to be easily deployable, scalable, modular, and reproducible. Other workflows partially overlap these aims, such as processing historical DNA sequencing data (PALEOMIX—Schubert *et al.* 2014, nf-core/eager—Peltzer *et al.* 2016, Mapache—Neuenschwander *et al.* 2023), analysis of combined historical and modern DNA datasets (GenErode—Kutschera *et al.* 2022), and running genotype likelihood analyses from BAMs (loco-pipe—Zhou *et al.* 2024, see Supplementary Tables S1 and S2 for a comparison of PopGLen's features with these pipelines). PopGLen aims to incorporate all necessary steps to process raw sequencing data into population genomic results in a way that is flexible to datasets with both modern and historical DNA by performing alternate processing and filtering when required. To enable freely combining PopGLen with related workflows, including those cited above, the pipeline contains several modular components (Fig. 1), allowing users to substitute and extend portions of the workflow to their liking.

2 Approach and features

2.1 Preparation and input data

Sample data is required as either raw FASTQ files or aligned BAM files, and a local copy of the reference genome must be provided. Configuration is handled with (i) a YAML file where analyses are enabled and software configurations are set, (ii) a sample metadata list, assigning samples to populations, and (iii) a sequencing unit list, which links sample IDs to their input files. In the config file a name is given to the dataset, and all results will be placed in a folder under that name, allowing for multiple datasets to be managed within one working directory where they share relevant intermediate files (e.g. BAMs, repeat libraries).

2.2 Data processing

2.2.1 Reference genome preparation

To enable parallel processing, the pipeline groups contigs into similar sized "chunks" of a user-defined size, scalable to

Received: 4 November 2024; **Revised:** 11 February 2025; **Editorial Decision:** 3 March 2025; **Accepted:** 6 March 2025

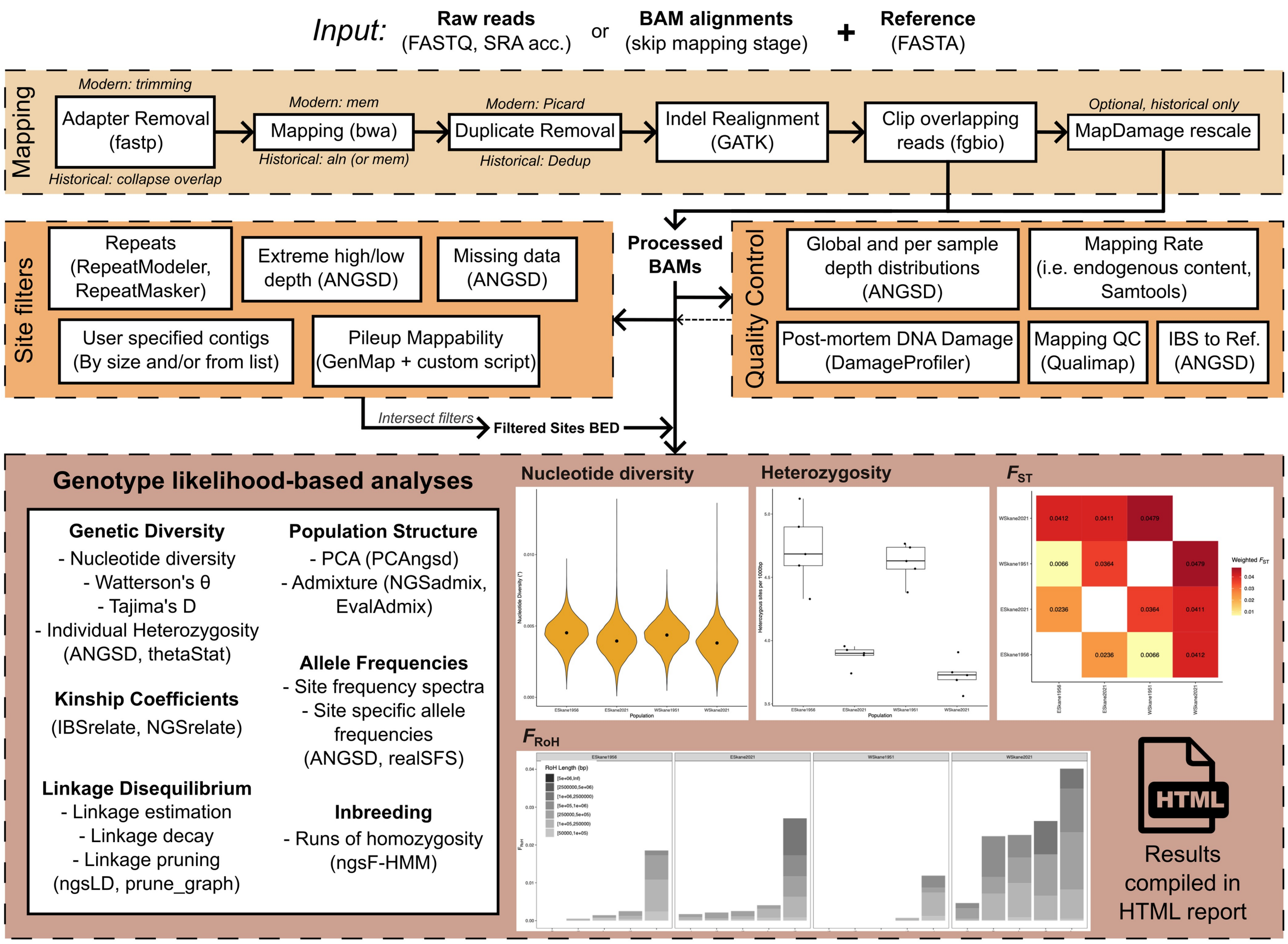

Figure 1. A broad overview of the workflow, analyses, and reporting available with PopGLen. Input requires a reference genome and sequencing reads provided as raw reads (FASTQ) or alignments (BAM). The workflow is divided into four modular components (dashed boxes): (i) mapping of sequence data to the reference genome (omitted for BAM input) with separate pathways for historical and modern samples, (ii) generation of a filtered sites file to limit analyses to, (iii) quality control to assess data quality and inform configuration, and (iv) population genomic analyses using genotype likelihood-based tools. Within each component, analyses are individually enabled and configured, allowing for high customizability. Once configured, all steps in the pipeline are automated using Snakemake and can be run using a single command. Figures and tables are prepared for most analyses and included in an HTML report. Four example plots are included here, based on data in the tutorial dataset, which examines declines in genetic diversity and increases in genetic differentiation and inbreeding in a pair of modern butterfly populations compared to their historical counterparts. The example plots have been resized but remain otherwise unmodified from the pipeline output. Higher resolution versions and other example outputs can be viewed in the example report included in the Supplementary Material.

many thousands of contigs. Mitochondrial, sex-linked, and other specified contigs can be excluded from these chunks.

2.2.2 Raw sequence data processing

Raw sequence data input in FASTQ format can be provided either as local paths or NCBI SRA accessions, the latter being downloaded automatically. Raw reads are processed and aligned to the reference genome for samples provided as raw reads rather than aligned BAMs.

Modern samples have adapters trimmed with fastp (Chen *et al.* 2018) and trimmed paired-end reads are mapped to the reference using BWA-MEM (Li 2013). If multiple libraries are provided for a sample, the subsequent BAM files are merged with Samtools (Danecek *et al.* 2021). Afterwards, duplicates are removed with Picard MarkDuplicates, reads are realigned around indels with GATK IndelRealigner (Van der Auwera and O'Connor 2020), and overlapping reads are trimmed by BamUtil (Jun *et al.* 2015) to avoid double counting of overlapping reads by ANGSD (Lou *et al.* 2021).

Historical samples will have pairs of overlapping reads collapsed by fastp, generating single reads. By default, collapsed reads are mapped using BWA-ALN using settings optimized for historical samples (Palkopoulou *et al.* 2015). Duplicates are removed with DeDup (Peltzer *et al.* 2016), which considers both read ends of the read. Reads are realigned around indels with GATK IndelRealigner and, optionally, base quality scores can be recalibrated using MapDamage 2 (Jónsson *et al.* 2013) to account for post-mortem DNA damage.

2.2.3 Data filtering

A sites-based filtering scheme is used to limit analyses to suitable sites (adapted from e.g. Pečnerová *et al.* 2021, Quinn *et al.* 2023). Each site filter is optional and generated independently. All filters are intersected using BEDTools (Quinlan and Hall 2010) to create a final usable sites list. Filters for mappability, repetitive content, extreme high or low global sequencing depth and missing data are included. Users can provide additional filters as BED files e.g. to restrict analyses to neutral sites or genic regions. The proportion of the genome passing each filter, and the combined filter set, is described in a summary table.

Low mappability regions are identified using mappability scores from GenMap (Pockrandt *et al.* 2020). These scores are converted to pileup mappability (Derrien *et al.* 2012), by averaging the mappability score of all fragments overlapping a position using BedOps (Neph *et al.* 2012). Both the fragment size (K) and allowed mismatches (E) used to estimate scores in GenMap, and the pileup mappability threshold, are set in the config file.

Repeat content identification and filtering is implemented through (i) building of a de-novo repeat library with RepeatModeler (Flynn *et al.* 2020), (ii) using a provided repeat library informing repetitive region identification with RepeatMasker (Smit *et al.* 2013), or (iii) removing repetitive regions provided in a BED/GFF file.

Upper and lower global depth filters are calculated across all samples, as well as for user-defined subsets of samples if configured. A filtering threshold can either be defined using upper/lower percentiles or multipliers to the median global sequencing depth.

2.2.4 Quality control

Several metrics to assess quality are compiled into the pipeline report. Adapter trimming reports are directly provided by fastp. Mapping rates are provided using Samtools. ANGSD is used to calculate mean and standard deviation sequencing depth per sample for (i) all positions without read filtering, (ii) all positions passing mapping and base quality thresholds, (iii) positions passing the main filter set, and (iv) positions passing any user-provided filter sets. Post-mortem DNA damage reports are provided by MapDamage2 and/or DamageProfiler (Neukamm *et al.* 2021). Reports containing general mapping statistics such as average mapping quality and GC content are provided by Qualimap (García-Alcalde *et al.* 2012). Identity by state similarity to the reference is estimated using ANGSD.

2.3 Population genomic analyses

Each of the analyses described below can be individually enabled in the config file. Analyses utilize genotype likelihoods for SNPs in Beagle format or site allele frequencies in SAF format, which are generated as needed and shared between analyses where appropriate. Transition removal in ANGSD to account for post-mortem DNA damage can be enabled across all analyses in the configuration file.

Estimation of linkage disequilibrium (LD) and LD decay are available using ngsLD (Fox *et al.* 2019), both across the whole dataset and for each population separately. Before running analyses that assume independence of positions, LD estimates are used to prune SNPs using prune_graph (https://github.com/fgvieira/prune_graph).

Two approaches for illustrating and assessing population structure are implemented—principal component analysis (PCA) using PCAngsd (Meisner and Albrechtsen 2018) and admixture analysis using NGSadmix (Skotte *et al.* 2013), both based on pruned SNPs. For each value of K, multiple independent NGSadmix replicates are performed to assess convergence, following customizable criteria described by Pečnerová *et al.* (2021). EvalAdmix (Garcia-Erill and Albrechtsen 2020) is run to provide an additional assessment of model fit. A list of individuals to exclude from these two analyses can be provided in the config to enable removing close relatives.

Kinship is inferred using IBSrelate (Waples *et al.* 2019) and/or ngsRelateV2 (Hanghøj *et al.* 2019). For IBSrelate, population allele frequencies are not required and both the IBS-based method implemented in ANGSD and the SFS-based method implemented in both ANGSD and ngsRelateV2 can be performed for all pairs of samples, with results provided in a table of R0, R1, and KING-robust kinship coefficients. The allele frequency-based methods in ngsRelateV2 can additionally be performed on each population, with allele frequencies estimated automatically by ANGSD and used by ngsRelateV2.

As a basis for population genomic estimates, folded and unfolded site frequency spectra (SFS) for single populations and population pairs are produced by ANGSD as needed for downstream analyses or as a requested output. Bootstrapped SFS can be additionally estimated, with the bootstrap count defined in the config file. Site frequency spectra are polarized to the reference allele, unless ancestral states are provided.

The pipeline has various options for estimating genetic diversity and neutrality statistics using ANGSD. Per-population estimates are performed in sliding windows, using user-defined size and step. Pairwise nucleotide diversity (π), Watterson's estimator (θ_W) and Tajima's D are also summarized into genome-wide means and confidence intervals per population. Individual heterozygosity is estimated from a single sample SFS by dividing the number of heterozygous positions by the total number of sites using ANGSD. Confidence intervals are inferred using the bootstraps of the single sample SFS.

Genetic differentiation, pairwise FST, is estimated both globally and in windows using ANGSD for all population pairs. The estimator used can be set in the config file, with the default set to the Hudson-Bhatia estimator (Bhatia *et al.* 2013), which is suited for small sample sizes.

Inbreeding coefficients are estimated from identical by descent (IBD) tracts using ngsF-HMM (Vieira *et al.* 2016). IBD tracts are estimated for pruned SNPs called within each population. Inbreeding coefficients are estimated both by ngsF-HMM and using the inferred IBD tracts to estimate F_{RoH}, the proportion of the autosomal genome covered by runs of homozygosity greater than a user-defined length (McQuillan *et al.* 2008).

An identity by state (IBS) matrix between all samples is estimated using SNPs called across the dataset in ANGSD.

Population allele frequencies are estimated using ANGSD. Each population has allele frequencies estimated for (i) variable sites within the population, assigning the minor allele to the population specific minor and (ii) variable sites within the dataset, including invariable sites within the population, assigning the minor allele to the minor of the entire dataset. Alternatively, the "major" allele can be fixed to the reference or provided ancestral allele.

Variation in sequencing depth can influence the outputs of many analyses. Users can define one or more target subsampled sequencing depths to subsample all samples to using Samtools. Each analysis can be separately enabled to run on subsampled data, allowing for comparisons between full and subsampled depth outputs to assess the influence of sequencing depth variation.

2.4 Reporting

This pipeline uses Snakemake's native report features to compile the results of the pipeline into a single HTML report. Quality control statistics summarized with MultiQC

(Ewels *et al.* 2016) and figures are generated for several analyses (Fig. 1) in R (R Core Team 2017). The report can be compiled for any successful run of the pipeline, allowing for easy assessment of partial runs to inform later analyses. An example report from the tutorial dataset, with all analyses enabled, is included in the Supplementary Material.

3 Availability, execution, and use cases

PopGLen is available on GitHub with documentation and a tutorial dataset (https://zjnolen.github.io/PopGLen) and can be deployed as a module in a single Snakefile using Snakedeploy. In this modular format, users can customize the workflow by adding additional rules that use PopGLen outputs as input, allowing extension or replacement of specific analyses. Required software are defined as Conda environments (Grüning *et al.* 2018) or Singularity containers (da Veiga Leprevost *et al.* 2017), requiring no pre-installed software aside from Conda and Singularity. After configuration, the pipeline can be run with a single Snakemake command, which will infer the required steps to generate the requested output, prepare the required software, and execute the workflow. Using Snakemake's executor plugins, it is compatible with a variety of high-performance computing job queue systems, enabling high levels of parallelization as jobs are submitted and monitored automatically to multiple nodes. I have ensured compatibility with Snakemake's code and parameter monitoring features, meaning changes to settings in the config or addition/removal of samples will trigger re-runs of relevant steps.

PopGLen is an efficient, flexible, and reproducible way for population genomic projects to go from raw data to several common population genomic statistics. It is particularly well suited for historical or low sequencing depth samples and comparisons of such samples to contemporary ones with higher coverage. It has applicability in a variety of research fields utilizing the implemented methods, including conservation, speciation, and evolutionary ecology.

Acknowledgements

I would like to thank Verena Kutschera and the Swedish Bioinformatics Advisory Program for advice on the development of the workflow and Anna Runemark for helpful comments on the manuscript. The computations for development and testing were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden and the Swedish National Infrastructure for Computing at UPPMAX partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973.

Author Contributions

ZJN: Conceptualization, Investigation, Software, Validation, Writing—Original Draft.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by a grant from the Strategic Research Area "Biodiversity and Ecosystem Services in a Changing Climate".

Data availability

PopGLen and its documentation is available at https://github.com/zjnolen/PopGLen, where questions, bug reports, and feature requests can be submitted. Releases are archived at https://doi.org/10.5281/zenodo.13384125. The data required to follow the tutorial is available at https://doi.org/10.6084/m9.figshare.27453978.v1.

References

Bhatia G, Patterson N, Sankararaman S *et al.* Estimating and interpreting FST: the impact of rare variants. *Genome Res* 2013; **23**:1514–21.

Chen S, Zhou Y, Chen Y *et al.* Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90.

da Veiga Leprevost F, Grüning BA, Alves Aflitos S *et al.* BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* 2017;**33**:2580–2.

Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**:1–4.

Díez-Del-Molino D, Sánchez-Barreiro F, Barnes I *et al.* Quantifying temporal genomic erosion in endangered species. *Trends Ecol Evol* 2018;**33**:176–85.

Derrien T, Estellé J, Marco Sola S *et al.* Fast computation and applications of genome mappability. *PLoS One* 2012;**7**:e30377.

Ewels P, Magnusson M, Lundin S *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8.

Flynn JM, Hubley R, Goubert C *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* 2020;**117**:9451–7.

Fox EA, Wright AE, Fumagalli M *et al.* ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics* 2019; **35**:3855–6.

García-Alcalde F, Okonechnikov K, Carbonell J *et al.* Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 2012;**28**:2678–9.

Garcia-Erill G, Albrechtsen A. Evaluation of model fit of inferred admixture proportions. *Mol Ecol Resources* 2020;**20**:936–49.

Grüning B, Dale R, Sjödin A *et al.*; Bioconda Team. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;**15**:475–6.

Hanghøj K, Moltke I, Andersen PA *et al.* Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *Gigascience* 2019;**8**:giz034.

Jónsson H, Ginolhac A, Schubert M *et al.* mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 2013;**29**:1682–4.

Jun G, Wing MK, Abecasis GR *et al.* An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* 2015; **25**:918–25.

Kutschera VE, Kierczak M, van der Valk T *et al.* GenErode: a bioinformatics pipeline to investigate genome erosion in endangered and extinct species. *BMC Bioinformatics* 2022;**23**:228–17.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013; 1303.3997.

Lou RN, Jacobs A, Wilder AP *et al.* A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol Ecol* 2021; **30**:5966–93.

McQuillan R, Leutenegger A-L, Abdel-Rahman R *et al.* Runs of homozygosity in European populations. *Am J Hum Genet* 2008;**83**:359–72.

Meisner J, Albrechtsen A. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* 2018;**210**:719–31.

Mölder F, Jablonski KP, Letcher B *et al.* Sustainable data analysis with snakemake [version 2; peer review: 2 approved]. *F1000Res* 2021;**10**:33.

Neph S, Kuehn MS, Reynolds AP *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* 2012;**28**:1919–20.

Neuenschwander S, Cruz Dávalos DI, Anchieri L *et al.* Mapache: a flexible pipeline to map ancient DNA. *Bioinformatics* 2023;**39**:btad028.

Neukamm J, Peltzer A, Nieselt K *et al.* DamageProfiler: fast damage pattern calculation for ancient DNA. *Bioinformatics* 2021;**37**:3652–3.

Nielsen R, Paul JS, Albrechtsen A *et al.* Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;**12**:443–51.

Palkopoulou E, Mallick S, Skoglund P *et al.* Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr Biol* 2015;**25**:1395–400.

Pečnerová P, Garcia-Erill G, Liu X *et al.* High genetic diversity and low differentiation reflect the ecological versatility of the African leopard. *Curr Biol* 2021;**31**:1862–71.e5.

Peltzer A, Jäger G, Herbig A *et al.* EAGER: efficient ancient genome reconstruction. *Genome Biol* 2016;**17**:60–14.

Pockrandt C, Alzamel M, Iliopoulos CS *et al.* GenMap: ultra-fast computation of genome mappability. *Bioinformatics* 2020;**36**:3687–92.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.

Quinn L, Garcia-Erill G, Santander C *et al.* Colonialism in South Africa leaves a lasting legacy of reduced genetic diversity in Cape buffalo. *Mol Ecol* 2023;**32**:1860–74.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017.

Schubert M, Ermini L, Der Sarkissian C *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat Protoc* 2014;**9**:1056–82.

Skotte L, Korneliussen TS, Albrechtsen A *et al.* Estimating individual admixture proportions from next generation sequencing data. *Genetics* 2013;**195**:693–702.

Smit AFA, Hubley R, Green P. *RepeatMasker Open-4.0*. 2013–2015. http://www.repeatmasker.org.

Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. Sebastopol, CA, USA: O'Reilly Media, Inc. 2020.

Vieira FG, Albrechtsen A, Nielsen R *et al.* Estimating IBD tracts from low coverage NGS data. *Bioinformatics* 2016;**32**:2096–102.

Waples RK, Albrechtsen A, Moltke I *et al.* Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Mol Ecol* 2019;**28**:35–48.

Zhou ZT, Owens GL, Larson WA *et al.* loco-pipe: an automated pipeline for population genomics with low-coverage whole-genome sequencing. *Bioinform Adv* 2024;**4**:vbae098.

Bioinformatics, 2025, 41, 1–5
https://doi.org/10.1093/bioinformatics/btaf105
Applications Note