Article

Not peer-reviewed version

PSMC-FAC: A Statistical Framework for Correcting Loss of Heterozygosity in Low-Coverage Genomic Demographic Inference

Francisco Iglesias-Santos , Alba Nieto , Sònia Casillas , Antonio Barbadilla , Carlos Sarabia *

Posted Date: 9 March 2026

doi: 10.20944/preprints202603.0681.v1

Keywords: demographic inference; Sequential Markovian Coalescent; PSMC; Hausdorff distance; Fréchet distance; low coverage genome

 doi:10.20944/preprints202603.0681.v1

Article

PSMC-FAC: A Statistical Framework for Correcting Loss of Heterozygosity in Low-Coverage Genomic Demographic Inference

Francisco Iglesias-Santos [1,2,3,†], Alba Nieto [4,5,6,†], Sònia Casillas [1,7], Antonio Barbadilla [1,7] and Carlos Sarabia [1,4,*]

[1] Institut de Biotecnologia i Biomedicina (IBB), Universitat Autònoma de Barcelona (UAB), Bellaterra, Spain
[2] Department of Mathematics, University of Vienna, Vienna, Austria
[3] Department of Evolutionary Biology, University of Vienna, Vienna, Austria
[4] Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain
[5] École Pratique des Hautes Études (EPHE), PSL Research University, Paris, France
[6] iTHEMS, RIKEN, Wako, Japan
[7] Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Spain
* Correspondence: carlos.sarabia@upf.edu
† These authors contributed equally to this work.

Simple Summary

Understanding how animal and human populations changed over time helps scientists explain evolution, protect biodiversity, and predict how species may respond to environmental change. Modern DNA sequencing makes it possible to study population history, but collecting very detailed genetic data is expensive and often not practical. Researchers therefore frequently use lower-quality genetic data, which can lead to inaccurate results because some genetic differences are missed. In this study, we present a new automated method that corrects these errors by estimating how many genetic differences were overlooked and adjusting the results accordingly. The approach uses objective mathematical comparisons to find the best correction, avoiding subjective choices by researchers. We tested this method using genetic data from humans, wolves, and cattle, representing different population histories. This method allows researchers to obtain reliable information while reducing sequencing costs. Overall, the method improves consistency and makes population history studies more accessible for conservation, ecology, and evolutionary research across many species.

Abstract

Inferring demographic history from whole-genome data is a fundamental objective in evolutionary and conservation genomics. However, the Pairwise Sequentially Markovian Coalescent (PSMC) framework, the most widely used demographic inference method for whole-genome sequence data, is highly sensitive to sequencing coverage, with low coverage producing systematic underestimation of heterozygosity and biased effective population size trajectories. Here, we present PSMC-FAC, an automated method designed to optimize false-negative rate correction in low-coverage genomes by minimizing geometric distances between corrected and high-coverage demographic trajectories. Whole-genome datasets from humans, gray wolves, and cattle were downsampled across multiple coverage levels and processed through standard demographic inference pipelines. Corrected trajectories were compared using Hausdorff and discrete Fréchet distance metrics projected onto a common temporal grid, and optimal correction factors were modeled as a function of sequencing depth using polynomial regression. Across species and demographic contexts, PSMC-FAC markedly improved concordance between low- and high-coverage trajectories and revealed highly predictable coverage-dependent correction patterns. Overall, PSMC-FAC provides a reproducible and mathematically grounded alternative to subjective correction approaches, enabling reliable

 doi:10.20944/preprints202603.0681.v1

demographic inference from moderate-coverage genomes and facilitating broader population-scale genomic analyses.

Keywords: demographic inference; Sequential Markovian Coalescent; PSMC; Hausdorff distance; Fréchet distance; low coverage genome

1. Introduction

Reconstructing species' demographic histories is a main goal in evolutionary biology and conservation genetics, providing insight into how populations responded to past climatic fluctuations, geological events, and anthropogenic pressures, and helping predict future responses. Classical genetic markers, such as mitochondrial DNA and microsatellites, yield valuable information about recent population dynamics, but their limited temporal resolution restrict demographic inference largely to the most recent thousands of years [1–4]. In contrast, whole-genome sequencing (WGS) has driven the emergence of population genomic approaches capable of inferring demographic change across hundreds of thousands of years. Among these, the Pairwise Sequentially Markovian Coalescent (PSMC) method represents a landmark advance, allowing the reconstruction of effective population size (Ne) trajectories from a genome across deep evolutionary timescales [5].

PSMC infers changes in Ne through time by modeling genome-wide patterns of heterozygosity, which reflect variation in the time to the most recent common ancestor (TMRCA) between homologous genomic regions. Because demographic contractions, expansions, and periods of stability leave distinct signatures in these patterns, PSMC provides a powerful framework for investigating long-term population history under a coalescent-based model. The method is particularly informative for timescales spanning approximately 10,000 to 1–3 million years ago and has been widely applied across diverse taxa, including humans and both wild and domesticated species [5–9].

The theoretical foundation of PSMC lies in coalescent theory, which provides a retrospective description of genealogical relationships by tracing sampled lineages backward in time until they merge at common ancestors [10,11]. In recombining genomes, genealogies vary along the sequence, forming a complex structure known as the ancestral recombination graph (ARG). Although the ARG provides a complete representation of ancestry, its high dimensionality makes it computationally intractable for most practical applications [12]. To overcome this limitation, approximations based on Hidden Markov Models have been developed, leading to the Sequentially Markov Coalescent (SMC), which assumes that genealogical changes along the genome follow a Markov process [13]. Building on this framework, PSMC models coalescent time at each genomic position as a hidden state and the observed homozygous or heterozygous genotype as the emission, enabling inference of Ne trajectories from a single genome that, under ideal assumptions, can approximate the history of an entire population [14,15].

Despite its broad utility, largely due to its non-parametric nature, PSMC is sensitive to data quality, demographic complexity, and analytical choices. PSMC further assumes neutrally evolving regions within a panmictic population. Therefore, background selection or the inclusion of overlooked constrained genomic elements can bias coalescent rate estimates and inferences of effective population size trajectories [16,17]. Population structure can be another source of bias. Past coalescent rate depends on migration and deme configuration in addition to changes in Ne [18,19], making interpretation of inference non-trivial. Beyond this change in expected curve due to population structure, PSMC may generate spurious, abrupt peaks under structured scenarios that do not reflect the underlying coalescent dynamics [20]. Similar instabilities can also arise from misspecification of the PSMC time-interval discretization ("time vector") [21]. Although extensions such as MSMC and SMC++ improve resolution by incorporating multiple genomes [22,23], they retain the panmixia assumption, along with biases linked to it [20]; only more recent developments explicitly accommodate structured demographic scenarios [24].

 doi:10.20944/preprints202603.0681.v1

Inferring population-wide demographic history in natural populations is especially difficult when population structure is ubiquitous [25,26]. Typically, multiple genomes are required to characterize underlying demographic processes and disentangle biological signals from methodological artefacts [20]. Consequently, empirical studies frequently adopt low- to medium-coverage sequencing strategies and prioritize sampling large numbers of individuals to better capture population-level variation. When considering genome data quality, sequencing errors, false-negative heterozygous calls, and insufficient coverage can bias heterozygosity estimates and distort inferred demographic histories [6,9]. A particularly important limitation of the PSMC method arises when using low-coverage genomes. Reduced coverage decreases detection of heterozygous sites, leading to systematic underestimation of individual heterozygosity and consequently biased inference of Ne and coalescent times. PSMC trajectories therefore become displaced downward and toward more recent times, while sharp demographic changes appear attenuated and temporally shifted [4]. One strategy to mitigate this bias during plotting of the PSMC curve is to estimate the false-negative rate (FNR), a statistical correction for lost heterozygosity through rescaling of inferred coordinates from the first estimation of genetic diversity. Traditionally, FNR has been estimated by downsampling a high-coverage genome to lower coverage, plotting both together, and visually adjusting FNR values of the low-coverage curve until it partially overlaps with the high-coverage reference curve [7–9,27]. Although this approach allows incorporation of more genomes and improves PSMC-based inferences, visual corrections are subjective, time-consuming, and difficult to reproduce. Because generating large numbers of high-coverage genomes remains cost-prohibitive for many laboratories, population genomics will increasingly rely on medium- and low-coverage datasets. Consequently, there is a growing need for fast, objective, and mathematically grounded approaches to estimate and evaluate FNR corrections and enable robust comparisons of demographic trajectories across samples.

Here, we present PSMC FNR-Automatized Correction (PSMC-FAC), a novel method that introduces a statistical framework for optimizing FNR correction of low-coverage PSMC trajectories by minimizing graphical distances between candidate FNR-corrected low-coverage curves and their corresponding high-coverage reference curves. We assessed this approach using downsampled high-coverage publicly available WGS data from multiple populations of humans (*Homo sapiens*), a wild species (European gray wolves, *Canis lupus*), and a domesticated species (cattle, *Bos taurus*), representing diverse demographic histories and varying levels of genome-wide heterozygosity. We show that PSMC-FAC provides reproducible and robust FNR estimates through an automated pipeline applicable to any WGS data in BAM format, independent of genome-wide heterozygosity and demographic history. By enabling accurate demographic inference from low-coverage genomes, PSMC-FAC substantially lowers sequencing cost requirements and broadens access to population genomic analyses, thereby facilitating large-scale and comparative studies of demographic history across a wide range of taxa.

2. Materials and Methods

2.1. Dataset Preparation

We downloaded previously published WGS data for multiple populations of interest from publicly available repositories, including humans, cattle, and wolves. Wolf genomes were obtained in FASTQ format [28] and aligned to the CanFam3.1 reference genome [29] using bwa-mem v0.7.17 [30]. Reads were filtered to remove spurious alignments and PCR duplicates and were sorted following the protocol of Sarabia et al. [27]. Human WGS data were downloaded in CRAM format [31] from the 1000 Genomes Project [32] and included nine individuals from three different geographical origins: three Han Chinese (CHB), three Yoruba from Ibadan, Nigeria (YRI), and three Tuscan from Italy (TSI). Files in CRAM format were decompressed and transformed to BAM format using samtools [33]. Cattle genomes were obtained in BAM format from the public Agricultural Research Service of the United States Department of Agriculture [34,35]. Genome-wide sequencing depth for all samples was calculated using the samtools depth function [36]. Only samples with mean

 doi:10.20944/preprints202603.0681.v1

coverage >15× were retained to enable downsampling. Metadata for all downloaded samples are provided in **Table A1**.

Downsampling was performed on all BAM files to multiple target depths (5×–15×) using the samtools view -s function. Genome-wide depth of coverage for each downsampled BAM file was verified using the samtools depth function (**Table A1**). Following the PSMC manual specifications [5], downsampled BAM files were converted to variant call format (VCF) using bcftools mpileup and call functions [36]. To run this step, the CanFam3.1 reference genome [29] was used for wolves, the GRCh38 (hg38) reference genome [37] for humans, and the bosTau9 reference genome [38] for cattle. Variant Call Format (VCF) files [39] were then converted to FASTQ format using the bcftools view function and the vcfutils.pl vcf2fq utility [5], filtering out variants with coverage lower than 5 (-d 5) and higher than twice the average depth of coverage (-D), retaining most heterozygous sites while filtering out gene duplicates. Finally, FASTQ files were converted to PSMC input format (PSMCFA) using fq2psmcfa [5] with a minimum phred-scaled base quality threshold of 20.

We obtained a PSMC inference from each individual sample and its downsamples and followed the literature to define time interval patterns. For wolves, we applied a customized time interval pattern of 64 atomic intervals arranged as six intervals of size one followed by 58 intervals of size one (*16+581*), as described in Freedman et al. [40]. For humans and cattle, we used the default PSMC time interval pattern (4+25*2+4+6), following Li and Durbin [5] for humans and Mei et al. [41] and Liu et al. [42] for cattle (see Appendix A for specific commands and the pipeline in Figure 1).

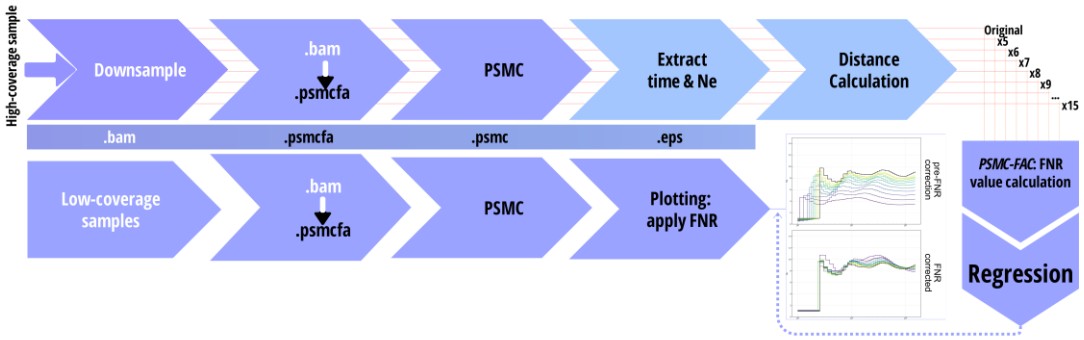


Figure 1. Schematic overview of the PSMC-FAC pipeline. A high-coverage genome (e.g., 20×) in BAM format from a given population is downsampled to multiple lower coverages (5×–15×), and each dataset is processed through the standard PSMC workflow. PSMC-FAC then estimates the optimal False Negative Rate (FNR) for each coverage level by minimizing the Hausdorff and discrete Fréchet distances between the original and downsampled trajectories. The resulting optimal FNR values are plotted against sequencing coverage, and a polynomial regression is fitted. This regression can subsequently be used to infer appropriate FNR corrections for additional genomes from the same population sequenced at varying depths. .

For each PSMC trajectory inferred from a downsampled genome, we applied False Negative Rate (FNR) correction across values ranging from 0 (no correction) to 0.99 in increments of 0.01, generating 100 corrected PSMC trajectories per downsample. Each corrected trajectory was then compared to the PSMC curve inferred from the original (non-downsampled) sequence. Because PSMC trajectories are defined over discretized time intervals that differ between inferences, direct pointwise comparison is not straightforward. To standardize comparisons, we defined a custom vector of 60 logarithmically spaced time points (in years) and projected all trajectories onto these shared temporal coordinates.

For cattle and wolves, the custom vector spanned 10 kya to 1.5 Mya. For humans, we repeated the analysis restricting the range to 50 kya to 1.5 Mya to avoid recent-time instabilities in PSMC. Early analyses attributed these peaks to Out-of-Africa demographic history [5], whereas more recent work has shown that they can arise from population structure [20] and from sensitivity to time-interval

 doi:10.20944/preprints202603.0681.v1

discretization [21]. Restricting the lower bound from 10 to 50kya was done to mitigate the influence of these artifacts.

2.2. A Mathematical Approach to Compute FNR for Low Coverage Samples

To quantify differences between PSMC trajectories, we employed two complementary distance-based metrics: (i) Hausdorff distance [43] and (ii) the discrete Fréchet distance [44]. Each PSMC trajectory is represented as an ordered set of points in the plane, where the x-coordinate corresponds to log-scaled time and the y-coordinate corresponds to inferred effective population size (Ne). Let

$$P = \{p_1, p_2, \ldots, p_m\}; Q = \{q_1, q_2, \ldots, q_m\} (1)$$

denote two PSMC curves, where P represents the reference trajectory obtained from a high-coverage genome (for example, 20×) and Q represents a trajectory derived from a downsampled genome after FNR correction. Before distance computation, trajectories are projected onto a shared temporal grid so that corresponding demographic histories can be compared consistently while preserving the piecewise-constant nature of PSMC inference.

2.2.1. Hausdorff Distance

The undirected Hausdorff distance measures the maximal local discrepancy between two curves. It is defined as:

$$H(P,Q) = max\left\{\underset{p_i \in P}{max}\ \underset{q_j \in Q}{min}\|p_i - q_j\|, \underset{q_j \in Q}{max}\ \underset{p_i \in P}{min}\|q_j - p_i\|\right\} (2),$$

where $||.||$ denotes the Euclidean norm [43].

For each point in one curve, the minimum distance to the other curve is computed. The Hausdorff distance corresponds to the largest of these minimal distances and therefore captures the worst-case disagreement between trajectories. This metric is sensitive to localized deviations, such as abrupt expansions or bottlenecks, and is useful for identifying regions where curves diverge strongly.

2.2.2. Discrete Fréchet Distance

While the Hausdorff distance ignores the sequential ordering of points, demographic trajectories represent ordered processes through time. To account for temporal progression, PSMC-FAC additionally uses the discrete Fréchet distance, which preserves ordering along both curves and measures similarity in overall shape. The discrete Fréchet distance between two polygonal curves is defined as:

$$\delta_F(P,Q) = \underset{\delta,\tau}{min}\ \underset{\kappa}{max}\|P_{\sigma(k)} - Q_{\tau(k)}\| (3),$$

where σ and τ are non-decreasing index sequences that traverse the points of P and Q, respectively, from start to end, preserving ordering [44].

Intuitively, the Fréchet distance measures the minimum leash length required for two entities to traverse both curves without backtracking. Because temporal ordering is maintained, this metric captures global similarity in demographic trajectory shape rather than isolated local differences.

2.2.3. A Combination of Both Methods

By using both metrics, we can detect localized deviations and global differences in curve geometry, providing a more comprehensive and robust comparison than either metric alone (Appendix A). A custom distance-calculation script computed both the discrete Fréchet distance and the Hausdorff distance between the original path and each FNR-corrected path.

After evaluating the accuracy of both the Hausdorff and Fréchet distances for our samples, we defined the optimal FNR value for each downsampled genome as the FNR corresponding to the minimum discrete Fréchet distance relative to the original (high-coverage) trajectory. The Fréchet

 doi:10.20944/preprints202603.0681.v1

metric was selected as the primary optimization criterion because it preserves point ordering and thus reflects similarity in overall demographic trajectory shape and temporal progression. Hausdorff distances were retained to evaluate localized maximal deviations.

The set of optimal FNR values obtained across coverage levels was subsequently modeled using a polynomial regression of degree 2 to generate continuous correction curves. These fitted polynomial functions describe the relationship between sequencing depth and the FNR value that minimizes trajectory divergence, thereby enabling depth-dependent calibration of Ne estimates.

2.3. FNR and Heterozygosity as a Function of Coverage

We evaluated whether the optimal FNR corrections depend not only on genome-wide heterozygosity and sequencing coverage, but also on the demographic history of the population under study. To assess this, we systematically examined the relationship between sequencing depth and the most optimal FNR factor inferred for each individual. To model the relationship between sequencing depth and optimal FNR, we fitted a second-degree polynomial regression in which coverage was treated as the independent variable, and the most optimal FNR value as the dependent variable. This polynomial model was chosen to capture non-linear scaling of FNR with sequencing depth. The regression was fitted independently for each individual.

Because optimal FNR values were determined using two alternative distance metrics, we repeated the polynomial fitting procedure separately for FNR estimates obtained through minimization of the Hausdorff distance [43] and for those obtained through minimization of the Fréchet distance [44], allowing comparison between both metrics. Following Nadachowska-Brzyska et al. [9], we assumed that the optimal FNR is equal to zero for samples with mean coverage greater than 15×. Accordingly, for regression purposes, high-coverage genomes exceeding this threshold were treated as having FNR = 0.

2.4. Sum-of-Least-Squares Assessment of Goodness-of-Fit

To evaluate the effectiveness of FNR-based correction across coverage levels, we quantified the similarity between PSMC trajectories using an independent sum-of-least-squares criterion. For each downsampled individual and each coverage level (5×–15×), we first identified the most optimal FNR value by minimizing either the Hausdorff distance [43] or the Fréchet distance [44] between the FNR-corrected trajectory and the corresponding high-coverage (non-downsampled) reference trajectory. This procedure yielded, for each downsample and for each distance metric, a single corrected curve with an optimal FNR value.

Subsequently, and independently of the optimization step, we computed the Sum of Squared Errors (SSE) between the best FNR-corrected trajectory and the high-coverage reference trajectory. SSE was therefore used exclusively as an evaluation metric and was not involved in the selection of the optimal FNR value. To calculate SSE, at each time point of the custom vector, we obtained the difference in effective population size (Ne) between the low-coverage and high-coverage trajectories, squared, and summed across all time points. This procedure was repeated for each downsampled coverage level and for each individual separately, generating a set of SSE values corresponding to the best FNR correction at each depth. These SSE values were then used to assess how closely the optimally corrected trajectories approximated the high-coverage reference across coverage levels. Comparisons were based on direct inspection of SSE trends across depths and individuals.

3. Results

3.1. PSMC-FAC Enables Accurate FNR-based Correction Across Species and Coverages

Across all three species (cattle, wolves, and humans) and populations analyzed, PSMC-FAC produced mathematically consistent corrections of the false negative rate (FNR), substantially improving concordance between downsampled and original (~20×) PSMC trajectories (**Figure 2**, **Fig.**

 doi:10.20944/preprints202603.0681.v1

A1). After projection onto a common logarithmic time grid with a custom time vector and comparison to the corresponding high-coverage inference, FNR-corrected trajectories generally converged toward the reference curve across the evaluated temporal window (**Fig. A1**).

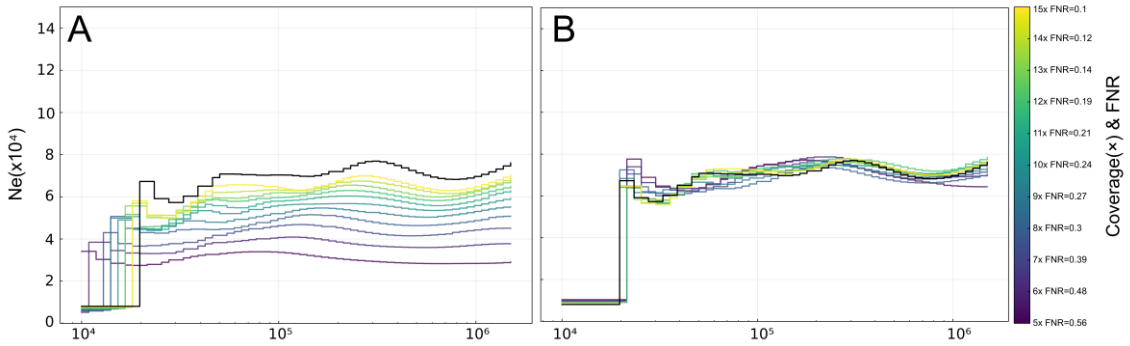


Figure 2. Overlapping PSMC trajectories of sample Iberian grey wolf (SAMN43221691) before (A) and after (B) PSMC-FAC-based FNR correction. Low coverage reduces heterozygote detection, shifting inferred coalescent events toward more recent times (leftward displacement) and lowering Ne estimates (downward displacement). FNR correction substantially restores trajectory concordance at moderate coverages (6–15×), whereas at very low coverages (<5×) correction becomes less reliable due to increased stochastic noise.

Correction performance varied depending on both sequencing depth and demographic profile. In general, reconstructions from genomes of higher coverages (10–15×) showed close agreement with the original trajectory after FNR correction. In contrast, medium-to-low coverage datasets (5–6×) showed greater instability and reduced correction accuracy (**Fig. A1**). At 5× coverage, several corrected trajectories of gray wolves displayed slight increases in population size when the original 20× curve showed a slight decrease (**Fig. A1 panels D2, E2, G2**). This pattern, although apparently paradoxical, is consistent with expectations from the variant-calling pipeline, since a minimum depth filter of 5 (-d 5) was applied during PSMC preprocessing, causing sites with depth lower than 5× to be ignored. In addition, the reference genome used for wolves (CanFam3.1) is known to exhibit relatively low heterozygosity [29].

PSMC-FAC was also able to adjust downsampled curves toward high-coverage trajectories in demographic scenarios characterized by abrupt population size changes, such as strong bottlenecks or rapid expansions (e.g., **Fig. A1 panels A1–A2, D1–D2, F1–F2, G1–G2**). In such cases, corrected curves tended to smooth extreme transitions, suggesting that strong local curvature in inferred coalescent rates is especially sensitive to heterozygosity loss caused by downsampling.

3.2. Appropriate FNR-based Correction Depends on Recent Demographic History

Correction performance differed among human populations and appeared to depend on recent demographic history. While Yoruba (YRI) trajectories were consistently well corrected across coverages (**Fig. A1 panels M–O**), PSMC-FAC showed reduced performance for Han (CHB; **Figure S1 panels J–L**) and Toscani (TSI; **Fig. A1 panels P–R**) genomes. This discrepancy coincides with a pronounced population size increase inferred in non-African samples between ~50–40 kya, where PSMC trajectories display a large, sharp elevation in effective population size during this interval. Because FNR optimization relies on minimizing Hausdorff and Fréchet distances between corrected and original trajectories [42,43], extreme local deviations can disproportionately influence the fitting procedure. If the last fragment of the demographic trajectory presents a sharp elevation, the algorithm will be biased to optimize correction parameters to better match this last maximum rather than the overall shape of the curve across the remaining temporal range.

To evaluate how PSMC-FAC adjusts downsampled curves to the original if the last peak is ignored, we repeated the FNR correction restricting the custom vector for the comparison to 50 kya–

 doi:10.20944/preprints202603.0681.v1

1.5 Mya for all genomes (**Fig. A2**). This modification produced no substantial change for wolves or cattle (**Fig. A2 panels A–I**), nor for Yoruba genomes (**Fig. A2 panels M2–O2**), where the discrepancy was not found previously. In contrast, correction performance for Han (**Fig. A2 panels J2–L2**) and Toscani (**Fig. A2 panels P2–R2**) genomes improved markedly. When the recent sharp elevation in the curve was excluded from the fitting interval, corrected trajectories showed substantially better global concordance with the original high-coverage curves.

3.3. FNR Corrections Are Robust Across Diverse Demographic Histories

Across species and populations, optimization of FNR values produced highly concordant results between the two distance metrics used. When corrections were computed over the 10 kya–1.5 Mya window, Hausdorff and Fréchet distances were strongly correlated across the full panel (**Figures 3, A3, Table A2**), indicating that both metrics identified nearly identical optimal FNR values. The only clear exceptions were the non-African human populations (Han and Toscani), consistent with the influence of the pronounced recent-time Ne peak described above. For all genomes, the relationship between sequencing coverage and inferred optimal FNR followed a highly regular polynomial trend. In every case, the fitted models yielded $R^2 > 0.99$, demonstrating a very close relationship between depth of coverage and the magnitude of FNR correction required. This pattern was consistent across cattle, wolves, and humans, despite their distinct demographic histories and differences in reference genomes and time interval parametrization. Importantly, Hausdorff and Fréchet distances performed nearly equivalently in identifying optimal FNR values. Differences in the selected FNR were minimal and resulted in only marginal variation in goodness-of-fit statistics for the coverage-FNR regression curves. Thus, correction accuracy does not depend strongly on the specific choice of distance metric, reinforcing the robustness of the PSMC-FAC optimization framework.

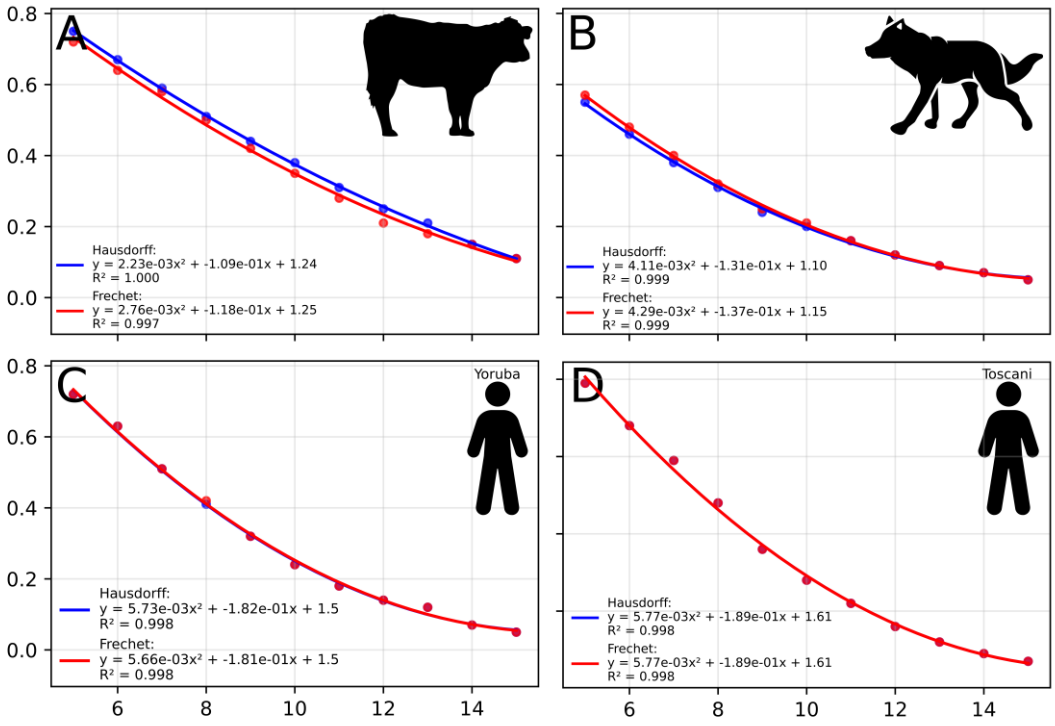


Figure 3. Polynomial regressions of optimal FNR as a function of sequencing coverage for cattle (A), wolves (B), and humans (C–D). Regressions were derived from PSMC-FAC-based FNR corrections using demographic trajectories between 10 kya and 1.5 mya (A, B, C) or between 50 kya and 1.5 mya (D), following the standard psmc_plot.pl output specifications. Strong and consistent correlations between FNR and coverage are observed across species and populations.

 doi:10.20944/preprints202603.0681.v1

After restricting the optimization window to 50 kya–1.5 Mya and re-estimating FNR values, the previously observed discrepancy in non-African human populations largely disappeared (Fig. A4; Table A3). Han and Toscani genomes then showed near-perfect correlations between genomic coverage and optimal FNR, comparable to those observed in Yoruba, wolves, and cattle. In contrast, only negligible changes were observed for the other genomes, whose trajectories lack the extreme recent-time Ne increase. These results strongly suggest that deviations observed in specific populations were not due to intrinsic limitations of the correction framework but rather to localized, high-amplitude features in the recent portion of the trajectory that disproportionately influenced distance-based optimization. Once these features were excluded, the coverage-FNR relationship became uniformly stable and highly predictable across all species and populations analyzed, making it possible to use it as a mathematical tool to predict the FNR needed for other coverages.

Although FNR optimization behaved consistently within species and populations, the shape of the polynomial relationship between sequencing coverage and optimal FNR differed among taxa (**Figs. A3-A5**). These differences highlight the distinct demographic histories and genome-wide heterozygosity levels observed across cattle, wolves, and humans (**Table A1**). Thus, while the coverage-FNR relationship is highly predictable within a population ($R^2 > 0.99$ in all cases), it is not universally transferable across populations with divergent demographic backgrounds, so running a PSMC-FAC pipeline per species or per population is highly advisable.

4. Discussion

In conservation and population genomics studies, there is an inherent trade-off between sequencing depth and sampling breadth [45,46]. High-coverage whole-genome sequencing remains costly, and generating large numbers of genomes at ~20× depth is often financially prohibitive for many research groups. Conversely, sequencing a small number of individuals at high coverage risks failing to capture the demographic heterogeneity of structured populations [47]. Although low-coverage genomes are frequently considered unsuitable for demographic inference, previous work has shown that meaningful demographic signals can be retained even under reduced representation approaches such as RAD sequencing under specific conditions [48]. Nadachowska-Brzyska et al. [9] demonstrated that PSMC performance declines substantially below approximately 10× coverage and recommended higher depths for stable inference. Building on these observations, we show that low- and medium-coverage genomes can still yield informative demographic inference when the loss of heterozygosity is quantitatively corrected using a False Negative Rate (FNR) factor. By deriving population-specific calibration curves linking FNR and sequencing coverage, the PSMC-FAC framework reduces uncertainty and allows demographic analyses to rely on a combination of high- and low-coverage genomes, thereby lowering overall sequencing costs for sound population-based demographic inference.

4.1. FNR Correction in Low- and Mid-Depth Genomes: Reference Genome Effect

At low coverage (e.g., ~5×), error rates remain substantial and FNR correction cannot fully compensate for stochastic loss of information. In contrast, between approximately 8× and 15× coverage, corrected trajectories converge closely toward high-coverage references, substantially expanding the practical usability of medium-depth sequencing data. The reduced performance observed at very low coverage reflects limitations of the variant-calling pipeline rather than a failure of the correction framework itself. During the PSMC preprocessing, a minimum depth threshold of five reads is applied (parameter -d 5). When mean coverage is ~5×, a substantial proportion of true heterozygous sites inevitably falls below this threshold and is therefore excluded. These sites are subsequently treated as homozygous reference during consensus generation, leading to systematic underestimation of heterozygosity. Because PSMC infers demographic history from the spatial distribution of heterozygous tracts along the genome, such undercalling directly alters the inferred coalescent rate. The resulting distortion is therefore not random noise but a predictable shift in trajectory shape, affecting inferred effective population size and potentially generating spurious

 doi:10.20944/preprints202603.0681.v1

demographic features. Similar coverage-dependent biases have been documented previously, particularly by Nadachowska-Brzyska et al. [9], who showed that decreasing depth disproportionately impairs heterozygote detection and destabilizes demographic inference. Our results are consistent with these observations. Importantly, while PSMC-FAC corrects for heterozygosity loss attributable to false negatives, at sufficiently low coverage both stochastic and systematic variation introduce an irreducible source of error. This is reflected in the elevated sum-of-squared errors observed even at the optimal FNR value at the 5× downsamplings. Thus, while the framework substantially mitigates coverage-related bias, it cannot completely overcome the intrinsic limitations of very shallow sequencing data.

These effects may become particularly pronounced when the reference genome does not adequately represent the genome-wide diversity of the population under study. For example, the human reference assembly (hg19/GRCh37) was constructed from a limited number of individuals and does not capture global human genetic diversity [49,50]. Similarly, the domestic dog reference genome, although widely used in canid studies, does not fully represent the diversity present across wolf and dog populations [51,52]. Under such circumstances, mapping bias toward the reference allele can further exacerbate heterozygote undercalling, amplifying the apparent loss of heterozygosity at low coverage. Consequently, the accuracy of FNR correction depends not only on sequencing depth but also on the appropriateness of the chosen reference genome. This highlights the importance of considering reference bias when interpreting demographic reconstructions derived from low-coverage data.

4.2. Effects of Biases Introduced by PSMC Assumptions on Optimal FNR Calculation

The pronounced peak observed in non-African human populations between approximately 50 and 40 thousand years ago provides a revealing example of how features of a PSMC trajectory can influence FNR optimization. When a broad temporal window starting at 10 kya was used for distance minimization, this sharp recent-time elevation disproportionately influenced both Fréchet and Hausdorff distance metrics. As a consequence, optimization tended to favor FNR values that improved agreement around this localized feature rather than across the trajectory as a whole. Similar peaks were already noted in the original PSMC publication by Li and Durbin (2011), and their biological interpretation has remained controversial. More recent work has demonstrated that population structure alone can generate such sharp biasing peaks in SMC-based inference due to coalescent signatures (Nieto et al. 2025). Tournebize et al. (2025) further demonstrated that structure and admixture can produce nearly indistinguishable coalescent patterns. Admixture with Neanderthals between approximately 50 and 43 kya remains a plausible biological contributor to this pattern, as interspecific gene flow is known to generate transient increases in inferred effective population size under SMC-based frameworks (Cahill et al 2017; Sarabia et al. 2021). However, because these peaks are also sensitive to the temporal discretization parameters used in PSMC (Hilgers et al., 2025), this feature is interpreted conservatively as a methodological artefact lacking straightforward demographic interpretation. When the optimization window was restricted to exclude this recent interval (starting at 50kya), FNR–coverage relationships stabilized markedly across Toscani and Han populations, whereas little change was observed for other populations and species. This indicates that the instability did not originate from the correction framework itself but rather from biases present in the reference trajectory. In practice, adjustment of PSMC hyperparameters and careful selection of the temporal window should therefore precede FNR optimization in order to reduce artefact-driven bias in the reference curve.

This observation leads to a central conceptual clarification: PSMC-FAC optimizes trajectory similarity rather than recovering true demography. The method minimizes geometric distance between a low-coverage trajectory and a chosen high-coverage reference; consequently, any artefacts contained in the reference trajectory are inherently propagated through the correction process. If the reference curve contains artefacts arising from structure (Nieto et. al., 2025), background selection (Cousins et al 2024), time discretization (Hilgers et al 2025), inclusion of functional genomic elements

 doi:10.20944/preprints202603.0681.v1

in the problem sequence (Sellinger et al 2021), or model misspecification (Cousins et al 2025), the FNR correction will reproduce them. PSMC-FAC therefore corrects coverage-dependent heterozygosity loss but does not address the theoretical or methodological limitations intrinsic to the sequentially Markovian coalescent framework. Sellinger et al. (2021) demonstrated that SMC-based methods have intrinsic convergence properties sensitive to mutation–recombination ratio assumptions, potentially limiting applicability in some populations. The present framework does not resolve such limitations; instead, it isolates and corrects one specific and pervasive source of bias: false-negative heterozygous calls caused by limited sequencing depth.

4.3. Polynomial Relationship Between Coverage and Optimal FNR

The strong and consistent polynomial relationship observed between sequencing coverage and optimal FNR indicates that heterozygosity loss scales predictably with depth under stable analytical conditions. Within individual populations, this relationship is highly regular, allowing coverage-dependent correction curves to be estimated with high confidence. However, the shape and parameters of these polynomial functions differ across species and populations, reflecting variation in genome-wide heterozygosity levels, recombination landscapes, and demographic history.

Demographic features themselves also influence optimization behavior. Sharp events, such as rapid bottlenecks or expansions, generate regions of high local curvature in PSMC trajectories. These features can disproportionately influence distance-based comparisons and, consequently, shift the inferred optimal FNR values. Thus, FNR estimation is determined not only by sequencing coverage but also by the geometric properties of the underlying demographic trajectory. The combined use of Hausdorff and Fréchet distances [42,43] provides complementary perspectives on this problem. The Hausdorff distance is sensitive to localized maximum deviations, whereas the Fréchet distance captures overall trajectory similarity while preserving temporal ordering. Together, these metrics offer a robust framework for estimating coverage-dependent correction parameters while accounting for variation in demographic trajectory shape.

4.4. Empirical Applications and Future Implications

In practical terms, this framework enables demographic analyses in conservation and population genomics contexts where sequencing large numbers of individuals at high coverage is economically unfeasible. By calibrating coverage-dependent bias using a high-coverage reference genome, additional individuals sequenced at moderate depth can be incorporated into demographic analyses without relying on subjective or visually determined FNR adjustments. This substantially expands the range of sampling designs compatible with PSMC-based inference while maintaining methodological consistency across individuals.

These results should also be interpreted in light of the known temporal limits of PSMC inference. Patton et al. [55] demonstrated that SMC-based methods achieve their highest resolution at intermediate timescales, whereas accuracy declines for very recent demographic events. More recently, Peede et al. [56] emphasized that temporal resolution depends strongly on the distribution of coalescent and recombination events along the genome. To accommodate these constraints, PSMC-FAC allows the specification of custom temporal windows during FNR optimization, enabling the exclusion of intervals known to contain unstable or artefactual signals. Consequently, low-coverage genomes corrected using PSMC-FAC should not be interpreted at very recent or fine-scale temporal resolutions. Still, they can provide robust information for intermediate and deep-time demographic inference where PSMC performs most reliably.

PSMC-FAC provides a reproducible and mathematically grounded procedure for calibrating coverage-dependent bias. By replacing subjective visual adjustment of FNR values with distance-based optimization metrics, the framework improves reproducibility and allows explicit quantification of correction performance. Although currently implemented as a correction layer for PSMC trajectories, the underlying conceptual approach could potentially be extended to other SMC-

 doi:10.20944/preprints202603.0681.v1

based methods that rely on heterozygosity patterns, provided that suitable high-coverage reference data are available.

Overall, these results demonstrate that low-coverage genomes are not inherently unsuitable for demographic inference. When appropriately calibrated, they can approximate high-coverage demographic trajectories within predictable limits. By correcting the component of distortion attributable specifically to sequencing depth, PSMC-FAC lowers the practical barrier to demographic reconstruction and enables broader sampling strategies, making population-scale PSMC analyses more accessible to laboratories for which large-scale high-coverage sequencing remains economically unfeasible.

5. Conclusions

This study introduces PSMC-FAC, an automated and reproducible framework that corrects coverage-dependent biases in demographic reconstructions derived from low-coverage genomes. By replacing subjective visual adjustments with objective, distance-based optimization, the method enables consistent estimation of false-negative rate (FNR) corrections and substantially improves concordance between low- and high-coverage demographic trajectories across diverse species and demographic scenarios.

Our results show that low- and medium-coverage genomes can provide reliable demographic information when appropriately calibrated using population-specific correction curves. Although the framework does not address inherent theoretical limitations of Sequentially Markovian Coalescent models, it effectively corrects a major practical source of bias associated with sequencing depth. By reducing reliance on high-coverage datasets, PSMC-FAC lowers sequencing cost constraints, supports broader sampling designs, and expands the practical applicability of demographic inference in evolutionary biology, conservation genomics, and comparative population studies.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Table A2: Optimal FNR value per coverage and sample according to Hausdorff and Frechet distances. FNR was calculated in PSMC files with a custom vector spanning between 10kya and 1.5Mya; TableA3: Optimal FNR value per coverage and sample according to Hausdorff and Frechet distances. FNR was calculated in PSMC files with a custom vector spanning between 50kya and 1.5Mya.

Author Contributions: Conceptualization, F.I.S. and C.S.; methodology, F.I.S., A.N. and C.S.; software, F.I.S. and A.N.; resources, S.C. and A.B.; writing - original draft preparation, F.I.S., A.N. and C.S.; writing - review and editing, F.I.S., A.N., S.C., A.B. and C.S.; visualization, F.I.S., A.N. and C.S.; supervision, C.S.; project administration, S.C., A.B. and C.S. All authors have read and agreed the published version of this manuscript.

Funding: This research was partially funded by the Spanish Ministry of Science and Innovation (PID2021-127107NB-I00), RES-Red Española de Supercomputación (DATA-2022-1-0015) and the Catalan Agency for Management of University and Research Grants (2021-SGR-00526). C.S. was funded by projects PID2021-127107NB-I00 and PID2023-147621NB-I00.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: PSMC-FAC, its scripts and some example data used in this work can be found at https://github.com/franiiss/PSMC_FAC.

Acknowledgments: The authors are thankful to Carles Acosta and the Port d'Informació Científica (PIC) of the RES-Red Española de Supercomputación for their technical support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

 doi:10.20944/preprints202603.0681.v1

PSMC	Pairwise Sequentially Markovian Coalescent
FNR	False-Negative Rate
PSMC-FAC	PSMC False-Negative Rate Automatized Correction
LD	Linear dichroism
WGS	Whole-Genome Sequencing
Ne	Effective Population Size
TMRCA	Time to the Most Recent Common Ancestor
ARG	Ancestral Recombination Graph
HMM	Hidden Markov Model
SMC	Sequentially Markov Coalescent
MSMC	Multiple Sequentially Markovian Coalescent
RAD	Restriction Site Associated DNA
BAM	Binary Alignment/Map format
CRAM	Compressed Reference-oriented Alignment Map format
VCF	Variant Call Format
PSMCFA	PSMC FASTA-like input format
PCR	Polymerase Chain Reaction
SSE	Sum of Squared Errors
CHB	Han Chinese in Beijing, China (1000 Genomes Project population)
YRI	Yoruba in Ibadan, Nigeria (1000 Genomes Project population)
TSI	Toscani in Italia (1000 Genomes Project population)
1000GP	1000 Genomes Project
ARS-USDA	Agricultural Research Service, United States Department of Agriculture
GRCh38	Genome Reference Consortium Human Build 38
Hg38	Human Genome version 38
CanFam3.1	Dog Reference Genome Assembly Version 3.1
BosTau9	Cattle Reference Genome Assembly Version 9
kya	Thousand Years Ago
Mya	Million Years Ago
DNA	Deoxyribonucleic Acid
R^2	Coefficient of Determination

Appendix A: Computational Workflow for PSMC Processing and FNR Estimation Using PSMC-FAC

A.1: Preparation of PSMC Input Files

PSMC-FAC is a bioinformatic pipeline designed to correct for False Negative Rate (FNR) reliably through the use of mathematical distances in a plane between a Pairwise Sequential Markovian Coalescent (PSMC) demographic curve of a high coverage genome and PSMC curves of downsamples of the same genome at different low-to-medium coverages. The goal is to get a trustable extrapolation of coverages with FNR, which allows researchers to run multi-sample PSMC graphs of a population without having to spend enormous resources on whole genome sequencing for multiple samples. PSMC-FAC needs only to have one sample at high coverage (>18X) which will be downsampled to several coverages. After FNR is corrected for every coverage, a FNR vs coverage plot relationship will be drawn and a polynomial regression can be calculated, after which low-to-medium coverage samples can be used. This procedure avoids using the same genome to construct a bootstrapped individual-level PSMC plot [5] and allows using entire populations, therefore allowing visualization of population structure and diverse demographic trajectories.

The initial processing steps follow the standard workflow described by [5] and outlined in their github page (https://github.com/lh3/psmc). Prior to running the pipeline, a high-coverage genome (typically >18×) must be aligned to a reference genome in FASTA format (ref.fa) or obtained as an aligned BAM file. PCR duplicates and spurious alignments should be removed before further analysis. Then, the .bam file is converted into variant call format (VCF) with bcftoolsv1.13 mpileup and call commands and later transformed into consensus FASTQ sequences as in [5]:

bcftools mpileup -C50 -f ref.fa -Ou highcovgenome.bam | bcftools call -c -Oz -o hcgenome.vcf.gz (step 1)

 doi:10.20944/preprints202603.0681.v1

*bcftools view hcgenome.vcf.gz | vcfutils.pl vcf2fq -d 5 -D {2*AverageCoverage} hcgenome.fq* (step 2)

In step (2), sites with depth below 5× are excluded, and positions exceeding twice the average genomic coverage are removed to reduce artifacts associated with duplicated regions and abnormal read depth.

Consensus FASTQ files are then converted into PSMC input format (PSMCFA) applying a minimum phred-scale base quality threshold of 20 using:

psmc/fq2psmcfa -q20 hcgenome.fq > hcgenome.psmcfa (step 3)

Then, demographic inference is performed using PSMC:

psmc -N20 -t10 -r5 -p "time_vector" -o hcgenome.psmc hcgenome.psmcfa (step 4)

where the time interval configuration (time_vector) is species-specific, as described in the section Dataset preparation of Materials and Methods. Although many studies use the default time_vector described in [5], this was designed originally for human populations. For non-human populations, the use of a custom-made time vector according to the specificities of the demography of the target population is recommended. In particular, adjusting the interval configuration can reduce artefacts associated with recent-time population size increases, as discussed in [20,21].

To construct the FNR–coverage correction curve, the high-coverage genome is downsampled to multiple target depths. In practice, approximately 8–10 coverage levels are sufficient to accurately estimate the coverage–FNR relationship. Downsampling is performed using samtools view -s, where the sampling fraction corresponds to the ratio between target and original coverage. For example, reducing a 20× genome to approximately 5× requires retaining 25% of reads:

samtools view -s 0.25 -b 20xgenome.bam > 5xgenome.bam (step 5)

This procedure should be repeated for each desired coverage level. After downsampling, steps (1)–(4) are repeated for every generated BAM file to obtain a set of PSMC trajectories spanning multiple coverages. These trajectories form the basis for FNR estimation and subsequent correction modeling within PSMC-FAC.

A2: Usage of PSMC-FAC

PSMC-FAC automatizes the process of finding the optimal FNR-corrected curve per downsampled genome using either the Hausdorff or Fréchet distances. For each downsampled genome, PSMC-FAC evaluates a discrete grid of candidate FNR values spanning the interval [0,0.99] in increments of 0.01. Application of each candidate FNR produces a corrected PSMC trajectory, resulting in a set of 100 demographic curves per downsampled dataset. These corrected trajectories are compared against the corresponding high-coverage reference trajectory obtained from the same individual.

Prior to comparison, all trajectories are projected onto a common logarithmically spaced temporal grid to ensure point-wise correspondence across time while preserving the stepwise structure inherent to PSMC inference. Each trajectory is therefore represented as an ordered sequence of coordinates in two-dimensional space:

$$P = \left(t_i, N_{e,i}^{ref}\right)_{i=1}^{n}; Q_f = \left(t_i, N_{e,i}^{(f)}\right)_{i=1}^{n}$$

According to user specifications, PSMC-FAC evaluates all 100 FNR-corrected trajectories by computing either the discrete Fréchet distance or the Hausdorff distance between each corrected curve and the corresponding high-coverage reference trajectory. The selected metric is used to systematically quantify the geometric discrepancy between trajectories across the full range of tested FNR values.

PSMC_FNR.py automatically applies FNR correction to a single (or a set of) PSMC(s) and extracts .tsv files with a custom time vector to facilitate comparison. In each tsv file, the first column is time (a vector with n logarithmically spaced timepoints between maximum and minimum time specified by the user), whereas the following columns represent Ne with the corresponding FNR correction, specified as a column title. For the reference (not downsampled) high coverage sequence,

 doi:10.20944/preprints202603.0681.v1

.tsv first column will be time (same custom vector as before) and Ne without any FNR correction (FNR=0). Users can use PSMC_FNR.py call to match their species and experiment.

Single file:

python PSMC_FNR.py --psmc_path path/to/file.psmc --species cow --mu 0.98e-8 --g 5 --FNR_min 0.74 --FNR_max 0.86 --svalue 100 --tmin 1e4 --tmax 1.5e6 --n_timepoints 60 (step 5)

Directory mode:

python PSMC_FNR.py --run_directory --input_dir 1.ALL_psmc_files_to_tsv/cattle --species cow --mu 0.98e-8 --g 5 --FNR_min 0.74 --FNR_max 0.86 --svalue 100

Options:

-h, --help show this help message and exit

--psmc_path PSMC_PATH Path to input .psmc file (single-file mode)

--run_directory If set, run on a directory tree (directory mode)

--input_dir INPUT_DIR Root directory to search for .psmc files (required if --run_directory)

--base_files BASE_FILES [BASE_FILES …]
List of base/original .psmc filenames to treat as baseline sample with high coverage (FNR is 0). Example: --base_files 801-Cattle.psmc 911-Cattle.psmc 927-Cattle.psmc

--species SPECIES Species name (default: cow)

--mu MU Mutation rate (default: 0.98e-8)

--g G Generation time (default: 5)

--FNR_max FNR_MAX Maximum FNR (default: 1.0)

--FNR_min FNR_MIN Minimum FNR (default: 0.0)

--svalue SVALUE svalue / number of sites scaling used by PSMC (int), parameter -s. Default is 100

--tmin TMIN Minimum time (years) desired for the custom time vector (default: 1e4)

--tmax TMAX Maximum time (years) desired for the custom time vector (default: 1.5e6)

--n_timepoints N_TIMEPOINTS
Number of timepoints desired in custom time vector (default: 60)

Execution of PSMC-FAC produces a tab-separated output (.tsv) containing the calculated distances between the reference PSMC trajectory (P) and each FNR-corrected trajectory (Qf). For each downsampled genome, the pipeline automatically identifies the corrected trajectory associated with the minimum distance value according to the chosen metric. The corresponding optimal FNR value and corrected curve are reported as output, providing a mathematically explicit framework for selecting FNR corrections and enabling robust mitigation of coverage-dependent bias in PSMC-based demographic inference. The procedure is repeated independently for each low- and medium-coverage downsampled genome (for example, 5×–15×), resulting in a set of optimal FNR estimates corresponding to each coverage level.

A3: Polynomial Regression and Visualization of Coverage-FNR Relationships.

A second script, FNR_curves.py, is a downstream component of the PSMC-FAC pipeline used to summarize and model the relationship between sequencing coverage and the optimal FNR estimated through trajectory distance minimization. The script processes the tab-separated .tsv output generated during FNR optimization as aforementioned, with distance measurements for all tested FNR values (typically 0–0.99) per each coverage and either Hausdorff distances, Fréchet distances, or both.

Using the command-line option --approach (“hausdorff”, “frechet” or “both”), the script selects the corresponding metric and identifies, for each sample and downsampled coverage, the FNR value that minimizes the chosen distance. This produces a reduced dataset containing one optimal FNR estimate per sample and coverage level.

The resulting optimal values are written as one or two .csv files, depending on the selected approach. These data are subsequently modeled using quadratic polynomial regression of the form

 doi:10.20944/preprints202603.0681.v1

FNR = $ax^2 + bx + c$, where x represents sequencing coverage. Regression models are fitted independently for each sample, and model fit is evaluated using the coefficient of determination (R^2). The script also generates a multi-panel .pdf figure showing the observed optimal FNR values, fitted regression curves, and corresponding equations and R^2 values for each sample. When both distance metrics are selected, Hausdorff- and Fréchet-based fits are displayed simultaneously for comparison. Finally, a .txt output file summarizes the regression equations and R^2 statistics in a machine-readable format.

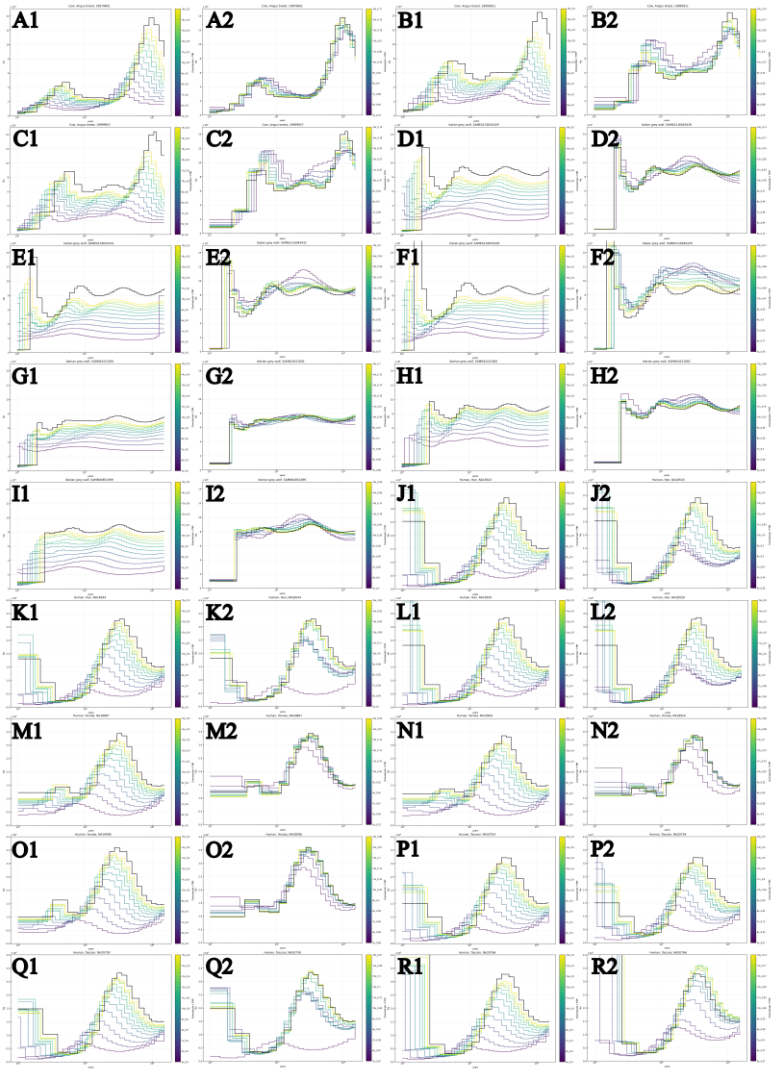
A4: Plotting Other Low-Coverage Genomes According to PSMC-FAC-assisted FNR Correction:

Once FNR values have been assigned to each genome, corrected trajectories can be visualized using the standard plotting utilities provided in the original PSMC package. FNR values are supplied through the -M option of psmc_plot.pl, which specifies the correction applied to each sample. For example, given three genomes with different sequencing coverages (genome1 = 20×, genome2 = 5.73×, genome3 = 7.5×) and corresponding FNR values inferred from the regression model (0, 0.53, and 0.22, respectively), plotting is performed as follows:

psmc_plot.pl -M "genome1=0,genome2=0.53,genome3=0.22" prefix genome1.psmc genome2.psmc genome3.psmc (step 6)

Following these steps, a bootstrapped PSMC plot can be generated using multiple genomes with heterogeneous sequencing coverages while applying coverage-specific FNR corrections.

Appendix B



 doi:10.20944/preprints202603.0681.v1

Figure A1. PSMC demographic reconstructions using the original (~20×) genome coverage and downsampled datasets ranging from 5× to 15×. Results are shown without (1) and with (2) PSMC-FAC–based false negative rate (FNR) corrections. FNR corrections were estimated using time intervals between 10 kya and 1.5 Mya and applied accordingly to the reconstructions. Panels depict three cattle breeds (A–C), Italian grey wolves (D–F), Iberian grey wolves (G–I), and human genomes from China (Han; J–L), Nigeria (Yoruba; M–O), and Italy (Toscani; P–R).

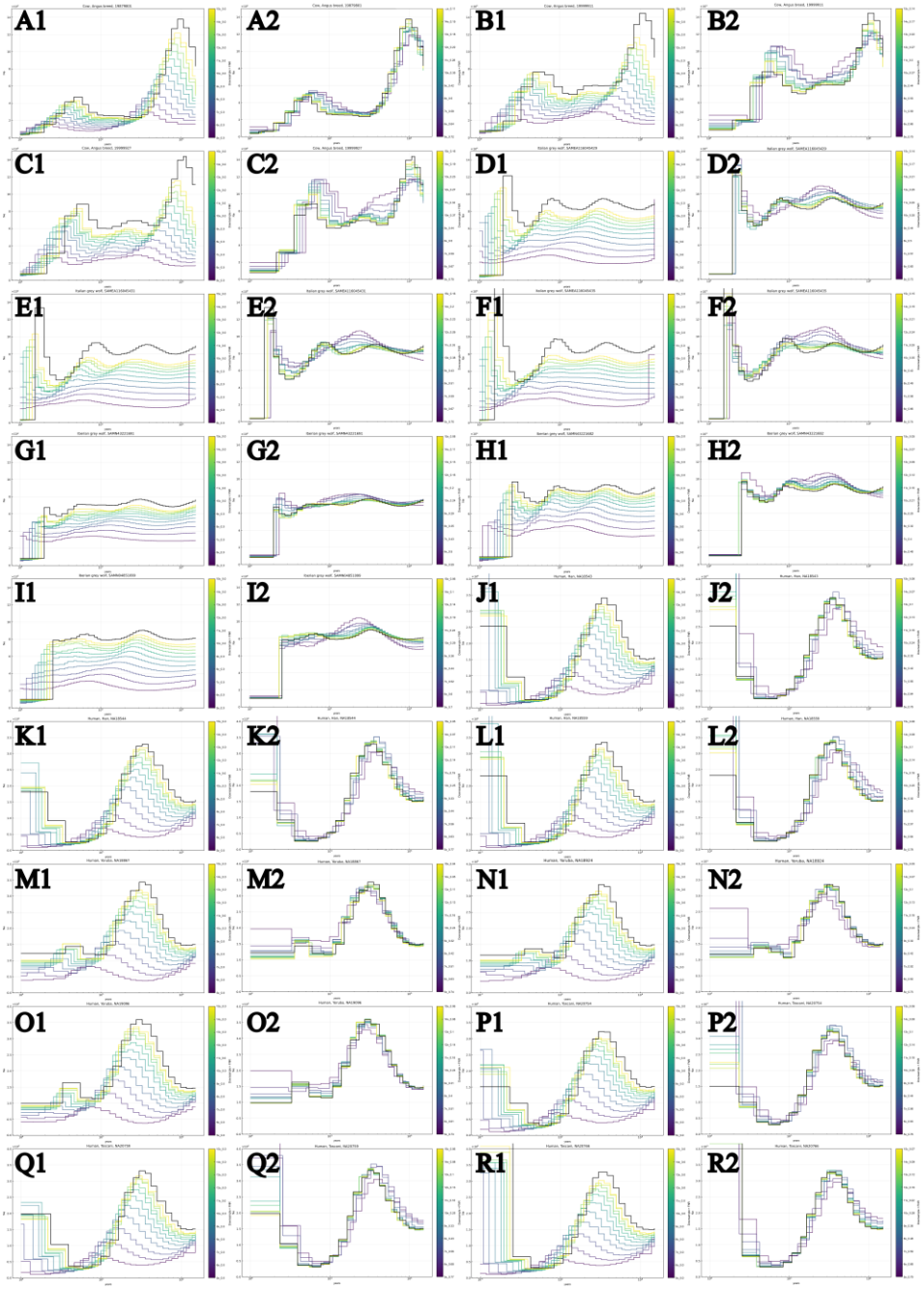


Figure A2. PSMC demographic reconstructions using the original (~20×) genome coverage and downsampled datasets ranging from 5× to 15×. Results are shown without (1) and with (2) PSMC-FAC–based false negative rate (FNR) corrections. FNR corrections were estimated using time intervals between 50 kya and 1.5 Mya and applied accordingly to the reconstructions. Panels depict three cattle breeds (A–C), Italian grey wolves (D–F), Iberian grey wolves (G–I), and human genomes from China (Han; J–L), Nigeria (Yoruba; M–O), and Italy (Toscani; P–R).

 doi:10.20944/preprints202603.0681.v1

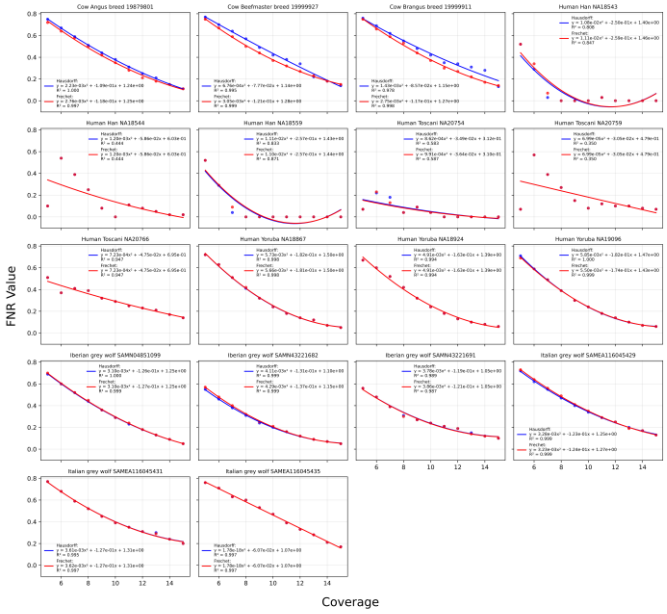
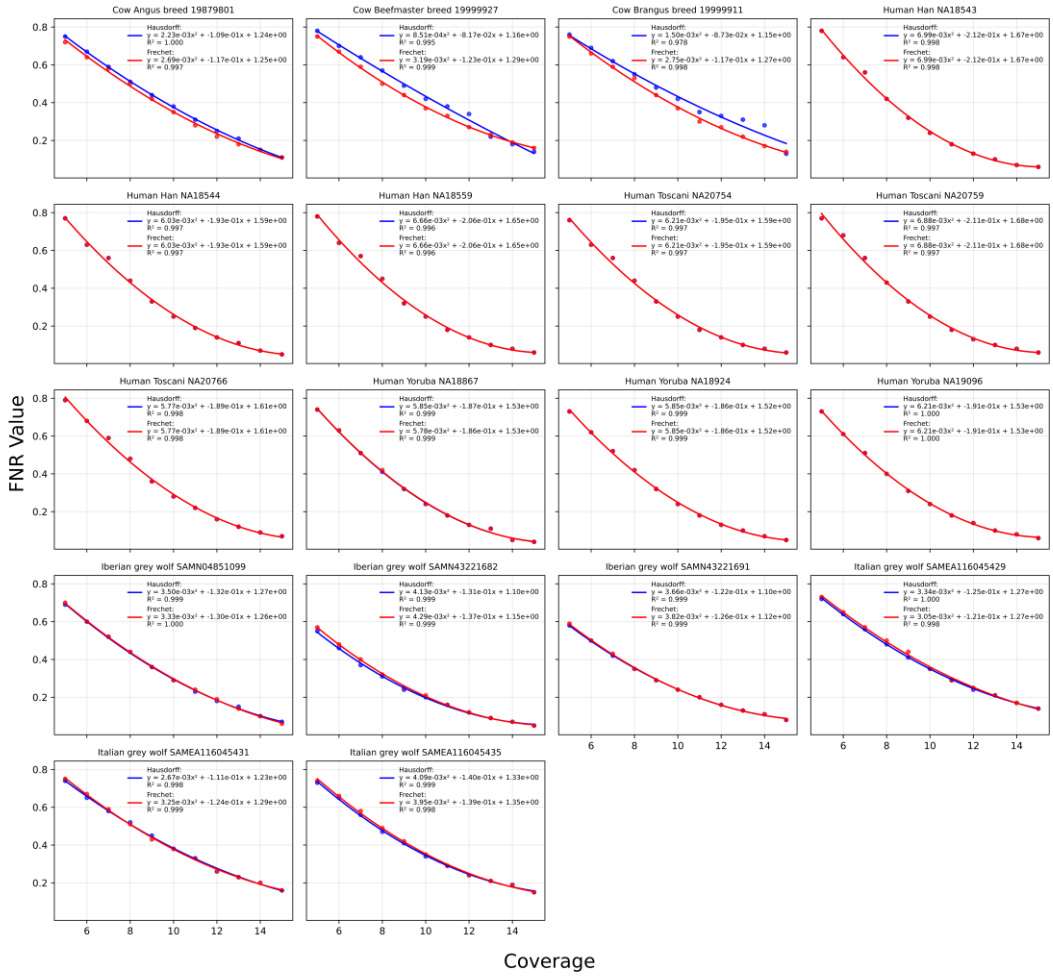


Figure A3. Quadratic polynomial regressions describing the relationship between false negative rate (FNR; y-axis) and sequencing coverage (x-axis) for all downsampled individuals. Polynomial models of degree 2 were fitted using both Hausdorff- and Fréchet-based optimization approaches. FNR values were estimated from comparisons performed across the time interval spanning 10 kya to 1.5 Mya.



 doi:10.20944/preprints202603.0681.v1

Figure A4. Quadratic polynomial regressions describing the relationship between false negative rate (FNR; y-axis) and sequencing coverage (x-axis) for all downsampled individuals. Polynomial models of degree 2 were fitted using both Hausdorff- and Fréchet-based optimization approaches. FNR values were estimated from comparisons performed across the time interval spanning 50 kya to 1.5 Mya.

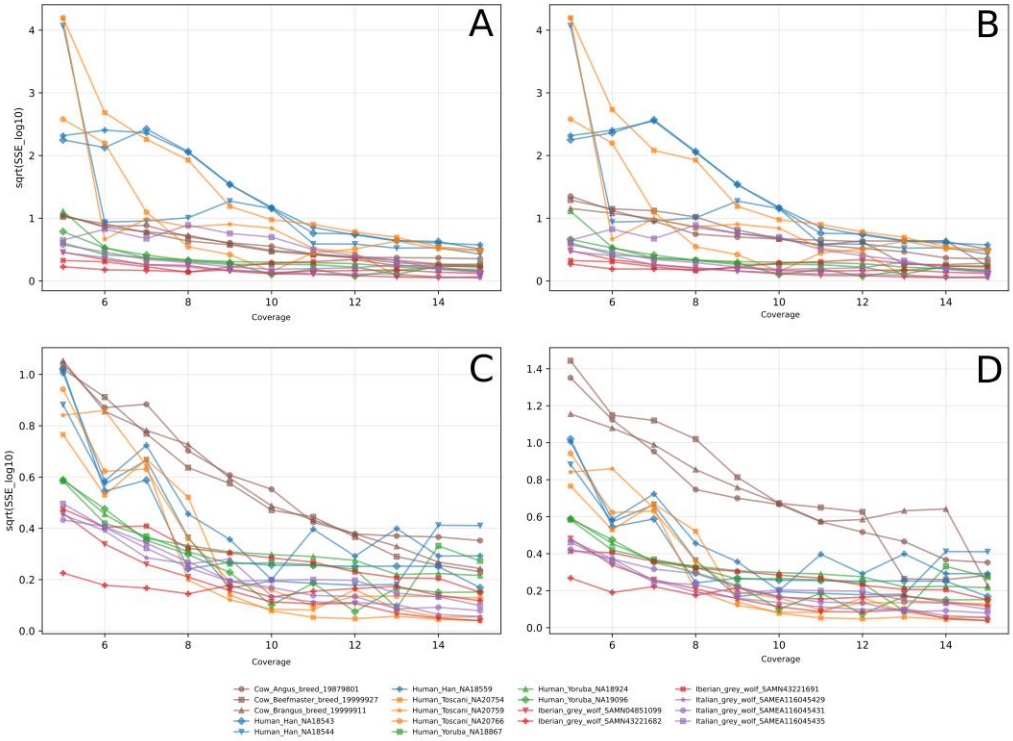


Figure A5. Log-scaled sum of squared errors per coverage and sample of every FNR calculation according to optimal Fréchet (A & C) or Hausdorff (B & D) distances. FNR values were estimated from comparisons performed across the time interval spanning 10 kya to 1.5 Mya (panels A & B) or 50 kya to 1.5 Mya (panels C & D).

Table A1. Summary of samples included in this study, comprising individuals from three species (Bos taurus, Canis lupus, and Homo sapiens), representing multiple populations and breeds. Data sources are as follows: (A) Daetwyler HD, et al. (2014) Nat Genet. 2014 Aug;46(8):858-65. doi: 10.1038/ng.3034. (B) Battilani D, et al. (2024) J Hered. 2025 Jan 3;116(1):10-23. doi: 10.1093/jhered/esae041. (C) Sarabia C, Salado I, et al. (2025) Mol Ecol. 2025 Jun;34(12):e17639. doi: 10.1111/mec.17639. (C) 1000 Genomes Project Consortium et al. (2015) Nature. 2015 Oct 1;526(7571):68-74. doi: 10.1038/nature15393.

Common name	Scientific name	Population	Sample number	Coverage	Heterozygosity	Source
Cow	*Bos taurus*	Angus breed	19879801	19.18X	$2.38*10^{-3}$	(A)
		Brangus breed	19999911	39.47X	$3.98*10^{-3}$	(A)
		Beefmaster breed	19999927	31.11X	$3.92*10^{-3}$	(A)
Grey wolf	*Canis lupus*	*C.l.italicus* (Italian wolf)	SAMEA116045429	23.67X	$1.68*10^{-3}$	(B)
			SAMEA116045431	27.19X	$1.48*10^{-3}$	(B)

 doi:10.20944/preprints202603.0681.v1

			SAMEA116045435	26.08X	1.44*10⁻³	(B)
		C.l.signatus (Iberian wolf)	SAMN43221691	20.42X	1.87*10⁻³	(C)
			SAMN43221682	19.03X	1.81*10⁻³	(C)
			SAMN04851099	18.08X	1.88*10⁻³	(C)
Human	*H. sapiens*	Han from China (CHB)	NA18543	29.59X	1*10⁻³	(D)
			NA18544	29.53X	9.82*10⁻⁴	(D)
			NA18559	33.34X	9.89*10⁻⁴	(D)
		Yoruba from Nigeria (YRI)	NA18867	30.18X	1.32*10⁻³	(D)
			NA18924	31.27X	1.32*10⁻³	(D)
			NA19096	31.35X	1.32*10⁻³	(D)
		Toscani from Italy (TSI)	NA20754	31.63X	1.04*10⁻³	(D)
			NA20759	32.65X	1.05*10⁻³	(D)
			NA20766	29.95X	1.03*10⁻³	(D)

References

1. Aimé, C.; Verdu, P.; Ségurel, L.; et al. Microsatellite data show recent demographic expansions in sedentary but not in nomadic human populations in Africa and Eurasia. *European Journal of Human Genetics* **2014**, 22, 1201–1207. https://doi.org/10.1038/ejhg.2014.2
2. Miller, E.F.; Manica, A.; Amos, W. Global demographic history of human populations inferred from whole mitochondrial genomes. *Royal Society Open Science* **2018**, 5(8), 180543. https://doi.org/10.1098/rsos.180543
3. Eddine, A.; Gomes Rocha, R.; Mostefai, N.; Karssene, Y.; De Smet, K.; Brito, J.C.; Klees, D.; Nowak, C.; Cocchiararo, B.; Lopes, S.; et al. Demographic expansion of an African opportunistic carnivore during the Neolithic revolution. *Biology Letters* **2020**, 16(1), 20190560. https://doi.org/10.1098/rsbl.2019.0560
4. Csapó, H.; Jabłońska, A.; Węsławski, J.M.; Mieszkowska, N.; Gantsevich, M.; Dahl-Hansen, I.; Renaud, P.; Grabowski, M. mtDNA data reveal disparate population structures and High Arctic colonization patterns in three intertidal invertebrates with contrasting life history traits. *Frontiers in Marine Science* **2023**, 10, 1275320. https://doi.org/10.3389/fmars.2023.1275320
5. Li, H.; Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **2011**, 475(7357), 493–496. https://doi.org/10.1038/nature10231
6. MacLeod, I.M.; Larkin, D.M.; Lewin, H.A.; Hayes, B.J.; Goddard, M.E. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Molecular Biology and Evolution* **2013**, 30(9), 2209–2223. https://doi.org/10.1093/molbev/mst125
7. Kim, H.; Ratan, A.; Perry, G.H.; Montenegro, A.; Miller, W.; Schuster, S.C. Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. *Nature Communications* **2014**, 5. https://doi.org/10.1038/ncomms6692
8. Hawkins, M.T.R.; Culligan, R.R.; Frasier, C.L.; Dikow, R.B.; Hagenson, R.; Lei, R.; Louis, E.E. Genome sequence and population declines in the critically endangered greater bamboo lemur (Prolemur simus) and implications for conservation. *BMC Genomics* **2018**, 19(1), 1–15. https://doi.org/10.1186/S12864-018-4841-4
9. Nadachowska-Brzyska, K.; Burri, R.; Smeds, L.; Ellegren, H. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology* **2016**, 25(5), 1058–1072. https://doi.org/10.1111/mec.13540
10. Kingman, J.F.C. On the genealogy of large populations. *Journal of Applied Probability* **1982**, 19(A), 27–43. https://doi.org/10.2307/3213548
11. Wakeley, J. Developments in coalescent theory from single loci to chromosomes. *Theoretical Population Biology* **2020**, 133, 56–64. https://doi.org/10.1016/j.tpb.2020.02.002

 doi:10.20944/preprints202603.0681.v1

12. McVean, G.A.T.; Cardin, N.J. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B* **2005**, 360(1459), 1387. https://doi.org/10.1098/rstb.2005.1673
13. Wiuf, C.; Hein, J. Recombination as a Point Process along Sequences. *Theoretical Population Biology* **1999**, 55(3), 248–259. https://doi.org/10.1006/tpbi.1998.1403
14. Mather, N.; Traves, S.M.; Ho, S.Y.W. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution* **2020**, 10(1), 579–589. https://doi.org/10.1002/ece3.5888
15. Peede, D.; Bañuelos, M.M.; Medina Tretmanis, J.; Miyagi, M.; Huerta-Sánchez, E. Recent advances in methods to characterize archaic introgression in modern humans. *Genome Research* **2026**, 36(2), 239–256. https://www.genome.org/cgi/doi/10.1101/gr.278993.124.
16. Sellinger, T.P.P.; Abu-Awad, D.; Tellier, A. Limits and convergence properties of the sequentially Markovian coalescent. *Molecular Ecology Resources* **2021**, 21(7), 2231–2248. https://doi.org/10.1111/1755-0998.13416.
17. Cousins, T.; Tabin, D.; Patterson, N.; Reich, D.; Durvasula, A. Accurate inference of population history in the presence of background selection. *BioRxiv* **2024.** https://doi.org/10.1101/2024.01.18.576291**.**
18. **Mazet, O.; Rodríguez, W.; Grusea, S.; et al. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference?** *Heredity* **2016, 116, 362–371.** https://doi.org/10.1038/hdy.2015.104
19. Chikhi, L.; Rodríguez, W.; Grusea, S.; Santos, P.; Boitard, S.; Mazet, O. The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity. *Heredity* **2018**, 120(1), 13–24. https://doi.org/10.1038/s41437-017-0005-6
20. Nieto, A.; Lao, O.; Mona, S. Performance of Sequential Markovian Coalescence Methods when Populations are Structured. *BioRxiv* **2025.** https://doi.org/10.1101/2025.10.09.681379**.**
21. **Hilgers, L.; Liu, S.; Jensen, A.; Brown, T.; Cousins, T.; Schweiger, R.; Guschanski, K.; Hiller, M. Avoidable false PSMC population size peaks occur across numerous studies.** *Current Biology* **2025, 35(4), 927–930.e3.** https://doi.org/10.1016/j.cub.2024.09.028**.**
22. **Schiffels, S.; Durbin, R. Inferring human population size and separation history from multiple genome sequences.** *Nature Genetics* **2014, 46, 919–925.** https://doi.org/10.1038/ng.3015
23. **Terhorst, J.; Kamm, J.A.; Song, Y.S. Robust and scalable inference of population history from hundreds of unphased whole genomes.** *Nature Genetics* **2016, 49(2), 303–309.** https://doi.org/10.1038/ng.3748
24. **Cousins, T.; Scally, A.; Durbin, R. A structured coalescent model reveals deep ancestral structure shared by all modern humans.** *Nature Genetics* **2025, 57, 856–864.** https://doi.org/10.1038/s41588-025-02117-1
25. **Hey, J.; Machado, C.A. The study of structured populations—new hope for a difficult and divided science.** *Nature Reviews Genetics* **2003, 4(7), 535–543.** https://doi.org/10.1038/nrg1112
26. **Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data.** *Genetics* **2000, 155(2), 945–959.** https://doi.org/10.1093/genetics/155.2.945
27. **Sarabia, C.; vonHoldt, B.; Larrasoaña, J.C.; Uríos, V.; Leonard, J.A. Pleistocene climate fluctuations drove demographic history of African golden wolves (Canis lupaster).** *Molecular Ecology* **2021, 30(23), 6101–6120.** https://doi.org/10.1111/mec.15784
28. **Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genomes Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* **2009, 25(16), 2078–2079.** https://doi.org/10.1093/bioinformatics/btp352
29. Lindblad-Toh, K.; Wade, C.; Mikkelsen, T.; et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **2005**, 438(7069), 803–819. https://doi.org/10.1038/nature04338
30. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **2010**, 26(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698
31. Bonfield, J.K. CRAM 3.1: advances in the CRAM file format. *Bioinformatics* **2022**, 38(6), 1497–1503. https://doi.org/10.1093/bioinformatics/btac010
32. 1000 Genomes Project Consortium; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; et al. A global reference for human genetic variation. *Nature* **2015**, 526(7571), 68–74. https://doi.org/10.1038/nature15393

 doi:10.20944/preprints202603.0681.v1

33. Li, H.; Handsaker, B.; Wysoker, A.; et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
34. USDA Agricultural Research Service (ARS). Bovine reference genome and whole-genome sequencing data. Agricultural Research Service, U.S. Department of Agriculture. Accessed November 2025. https://www.ars.usda.gov/plains-area/clay-center-ne/marc/wgs/bovref/
35. Heaton, M.P.; Smith, T.P.L.; Carnahan, J.K.; Basnayake, V.; Qiu, J.; Simpson, B.; Kalbfleisch, T.S. Using diverse U.S. beef cattle genomes to identify missense mutations in EPAS1, a gene associated with high-altitude pulmonary hypertension. *F1000Research* **2016**, 5, 2003. https://doi.org/10.12688/f1000research.9254.1
36. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; Li, H. Twelve years of SAMtools and BCFtools. *GigaScience* **2021**, 10(2), giab008. https://doi.org/10.1093/gigascience/giab008
37. Schneider, V.A.; Graves-Lindsay, T.; Howe, K.; Bouk, N.; Chen, H.C.; Kitts, P.A.; et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **2017**, 27(5), 849–864. https://doi.org/10.1101/gr.213611.116
38. Rosen, B.D.; Bickhart, D.M.; Schnabel, R.D.; Koren, S.; Elsik, C.G.; Tseng, E.; et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* **2020**, 9(3), giaa021. https://doi.org/10.1093/gigascience/giaa021
39. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330
40. Freedman, A.H.; Gronau, I.; Schweizer, R.M.; Ortega-Del Vecchyo, D.; Han, E.; Silva, P.M.; et al. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genetics* **2014**, 10(1), e1004016. https://doi.org/10.1371/journal.pgen.1004016
41. Mei, C.; Wang, H.; Liao, Q.; Wang, L.; Cheng, G.; Wang, H.; et al. Genetic Architecture and Selection of Chinese Cattle Revealed by Whole Genome Resequencing. *Molecular Biology and Evolution* **2018**, 35(3), 688–699. https://doi.org/10.1093/molbev/msx322
42. Liu, X.; Li, Z.; Yan, Y.; Li, Y.; Wu, H.; Pei, J.; et al. Selection and introgression facilitated the adaptation of Chinese native endangered cattle in extreme environments. *Evolutionary Applications* **2020**, 14(3), 860–873. https://doi.org/10.1111/eva.13168
43. Alt, H.; Behrends, B.; Blömer, J. Approximate matching of polygonal shapes. *Annals of Mathematics and Artificial Intelligence* **1995**, 13, 251–265. https://doi.org/10.1007/BF01530830
44. Ahn, H.K.; Knauer, C.; Scherfenberg, M.; Schlipf, L.; Vigneron, A. Computing the discrete Fréchet distance with imprecise input. *Lecture Notes in Computer Science* **2010**, 6507, 422–433. https://doi.org/10.1007/978-3-642-17514-5_36
45. Fuentes-Pardo, A.P.; Ruzzante, D.E. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Molecular Ecology 2017, 26(20), 5369–5406. https://doi.org/10.1111/mec.14264
46. Buerkle, C.A.; Gompert, Z. Population genomics based on low coverage sequencing: How low should we go? Molecular Ecology 2013, 22(11), 3028–3035. https://doi.org/10.1111/mec.12105
47. Hermosilla-Albala, N.; Silva, F.E.; Cuadros-Espinoza, S.; et al. Whole genomes of Amazonian uakari monkeys reveal complex connectivity and fast differentiation driven by high environmental dynamism. Communications Biology 2024, 7, 1283. https://doi.org/10.1038/s42003-024-06901-3
48. Liu, S.; Hansen, M.M. PSMC analysis of RAD sequencing data. Molecular Ecology Resources 2017, 17(4), 631–641. https://doi.org/10.1111/1755-0998.12606
49. Pan, B.; Kusko, R.; Xiao, W.; et al. Similarities and differences between variants called with human reference genome HG19 or HG38. BMC Bioinformatics 2019, 20(Suppl 2), 101. https://doi.org/10.1186/s12859-019-2620-0
50. Günther, T.; Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genetics 2019, 15(7), e1008302. https://doi.org/10.1371/journal.pgen.1008302

 doi:10.20944/preprints202603.0681.v1

51. Bergström, A.; Stanton, D.W.G.; Taron, U.H.; et al. Grey wolf genomic history reveals a dual ancestry of dogs. Nature 2022, 607, 313–320. https://doi.org/10.1038/s41586-022-04824-9
52. Battilani, D.; Gargiulo, R.; Caniglia, R.; et al. Beyond population size: Whole-genome data reveal bottleneck legacies in the peninsular Italian wolf. Journal of Heredity 2025, 116(1), 10–23. https://doi.org/10.1093/jhered/esae041
53. Tournebize, R.; Chikhi, L. Ignoring population structure in hominin evolutionary models can lead to the inference of spurious admixture events. Nature Ecology & Evolution 2025, 9, 225–236. https://doi.org/10.1038/s41559-024-02591-6
54. Cahill, J.A.; Soares, A.E.; Green, R.E.; Shapiro, B. Inferring species divergence times using pairwise sequentially Markovian coalescent modelling and low-coverage genomic data. Philosophical Transactions of the Royal Society B 2016, 371(1699), 20150138. https://doi.org/10.1098/rstb.2015.0138
55. Patton, A.H.; Margres, M.J.; Stahlke, A.R.; et al. Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in Tasmanian devils. Molecular Biology and Evolution 2019, 36(12), 2906–2921. https://doi.org/10.1093/molbev/msz191
56. Peede, D.; Cousins, T.; Durvasula, A.; et al. Not Just Ne No More: New Applications for SMC from Ecology to Phylogenies. Genome Biology and Evolution 2026, 18(1), evaf229. https://doi.org/10.1093/gbe/evaf229