Detectability of runs of homozygosity is influenced by analysis parameters and population-specific demographic history

Avril M. Harder[1], Kenneth B. Kirksey[2], Samarth Mathur[3], Janna R. Willoughby[1*]

1. College of Forestry, Wildlife and Environment, Auburn University, Auburn, Alabama, USA

2. Walker College of Business, Appalachian State University, Boone, North Carolina, USA

3. Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio, USA

*Author for correspondence

Abstract

Wild populations are increasingly threatened by human-mediated climate change and land use changes. As populations decline, the probability of inbreeding increases, along with the potential for negative effects on individual fitness. Detecting and characterizing runs of homozygosity (ROHs) is a popular strategy for assessing the extent of individual inbreeding present in a population and can also shed light on the genetic mechanisms contributing to inbreeding depression. Here, we analyze simulated and empirical data sets to demonstrate the downstream effects of program selection and long-term demographic history on ROH inference. We also apply a sensitivity analysis to evaluate the effects of various parameter values on ROH-calling results and demonstrate its utility for parameter value selection. We show that ROH inferences can be biased when sequencing depth and the distribution of ROH length is not interpreted in

light of demographic history as well as program-specific tendencies. This is particularly important for the management of endangered species, as underestimating inbreeding signals in the genome can substantially undermining conservation initiatives. Based on our observations, we suggest using a combination of ROH detection tools and ROH length-specific inferences to generate robust population inferences regarding inbreeding history. We outline these recommendations for ROH estimation at multiple levels of sequencing effort typical of conservation genomics studies.

Key words: inbreeding, population genomics, PLINK, BCFtools

Introduction

Climate change and expanding human land use are increasingly partitioning wild populations into smaller and smaller areas of available and suitable habitat, often leading to declining populations sizes [1,2]. Decreases in population size can lead to increased inbreeding, which has been reported to have negative fitness consequences for inbred individuals in many wild populations [3–6]. When inbreeding depression is sufficiently severe, populations can be threatened with extirpation; thus, assessing inbreeding extent is crucial for understanding and mitigating risk in small populations of conservation concern. Prior to widespread application of whole-genome sequencing strategies to non-model species, genetic estimates of inbreeding were obtained using allozyme or microsatellite data or inferred from known pedigrees [7–10]. These studies have been critically important to understanding the genetic dynamics of stable and shrinking populations and have led to increasing recognition of inbreeding depression's prevalence and ability to affect wild population persistence [4,11]. However, applying whole genome sequencing strategies to identify runs of homozygosity (ROHs; genomic regions where

43 both inherited haplotypes are identical) opens up lines of inquiry previously not accessible via
44 pedigree- or microsatellite-based studies [12].

45 The identification and quantification of inbreeding depression, largely a result of
46 homozygosity of strongly deleterious mutations, is critical to conservation as it guides effective
47 management strategies for myriad species. For example, the introduction of genetically diverse
48 individuals into declining populations to reduce the proportion of deleterious alleles present in a
49 population is expected to reduce genetic load and, as a result, reduce extinction risks. Further,
50 coupling genomic data with fitness outcomes has shown that longer runs of homozygosity
51 (ROHs) are associated with higher mutational loads [13,14] and ROH analyses in historical and
52 extant populations have identified genetic trajectories leading to species declines and extinctions
53 [15,16]. Thus, the integration of genetic data via ROH estimation and existing demographic data
54 can provide important clues about population stressors that support management decision
55 making, ultimately reducing extinction risks for target species.

56 Estimating ROHs can provide crucial insights into populations' evolutionary histories,
57 but these histories can in turn affect which ROH-calling software and combination of parameter
58 values are most appropriate. For example, the settings best suited for inferring ROHs in a small,
59 long-isolated population experiencing high levels of inbreeding would not be suitable for
60 individuals sampled from a large, genetically diverse population because underlying sources of
61 error in these two scenarios are very different (*e.g.*, differences in ROH length distributions,
62 numbers of variable sites, expected minor allele frequencies; [17]). While some studies include
63 comparisons of results from multiple programs or parameter value combinations (*e.g.,* [18–21]),
64 many more studies rely on default settings and do not explore the effects of varying these
65 parameter values on their results. This has the potential to be misleading because it is always

66 impossible to know how close the resulting estimates approximate reality, and yet ROH are an
67 alluring metric for decision makers based on the phenomenon that they illustrate.

68 Here, we address the challenge of interpreting ROH population patterns using simulated
69 and empirical genomic sequencing data to compare ROH identification patterns among different
70 demographic histories. Here, we focus on whole-genome sequencing data, building previous
71 studies that have examined ROH inference for data sets with lower marker densities [22–24].
72 Specifically, we test a wide array of setting combinations for two programs commonly used in
73 population genomic studies—PLINK and BCFtools/RoH—and, for PLINK, apply a sensitivity
74 analysis to evaluate the effects of parameter values on ROH inference. Based on these results, we
75 outline a set of recommendations for ROH estimation for population genomics studies.

76 **Methods**

77 *Part I: Simulated data*

78 Data generation and genotype calling

79 We used SLiM v4.0.1 and modifications of scripts published by Stoffel et al [6] to simulate four
80 distinct demographic scenarios: (i) a long-term large population, (ii) a long-term small
81 population, (iii) a bottlenecked population, and (iv) a long-term declining population. Each
82 simulation was initiated as a population of 10,000 individuals, wherein each individual consisted
83 of a homologous pair of 30-Mb chromosomes [25,26]. Each population was simulated for 10,000
84 generations, followed by scenario-specific demographic changes (Fig. 1). The large population
85 scenario involved an additional 1,000 generations at a population size of 1,000 individuals,
86 whereas the small population scenario involved 1,000 additional generations at a population size
87 of 250 individuals. The bottlenecked population persisted with 1,000 individuals for 900
88 additional generations, followed by a 50-generation interval with 50 individuals (*i.e.*, the

89 bottleneck) and a final 50-generation interval with 250 individuals. The declining population
90 continued for 850 generations with a population size of 1,000, followed by three 50-generation
91 intervals of decreasing population sizes, from 500 to 250 and finally to 50 individuals.
92 Recombination rate (9.15×10^{-9} per site per generation) and base mutation rate (1.39×10^{-8} per
93 site per generation) were chosen based on estimates calculated for Tasmanian devils
94 (*Sarcophilus harrissii*)—the focal species for our empirical tests [27,28]. Each VCF file output
95 from SLiM was converted to FASTA sequence files using a custom script in R v4.0.3 and a
96 haploid ancestral sequence comprising a 30-Mbp segment randomly sampled from the
97 Tasmanian devil genome [29].
98 Using the known genotypes for all individuals, we generated a coordinate file for each
99 demographic scenario recording the start and end coordinates for all true ROHs ≥ 100 kb in
100 length (*i.e.*, ≥ 100 kb of sequential homozygous loci). We imposed this lower limit on ROH
101 length because ROHs less than 100 kb in length likely originated in a single common ancestor
102 approximately 500 generations ago (assuming a recombination rate of 1 cM/1 Mb; [30]), and
103 would not be expected to influence contemporary individual fitness as strongly as more recently
104 acquired autozygous segments [13]. This threshold is widely applied in population genetics
105 studies of non-model species [5,31–33], and we follow this convention for all downstream
106 analyses.
107 We randomly selected 50 individuals from the final generation of each population
108 simulation and generated FASTQ read files from each of the two FASTA files representing
109 homologous chromosomes using ART (version MountRainier-2016-06-05) [34]. We simulated
110 150-bp paired-end reads using the HiSeq 2500 error model to a depth of 50X per individual (*i.e.*,
111 25X per homologous chromosome). Each FASTQ file was quality-checked using FASTQC

v0.11.9 [35]. We aligned reads to the ancestral sequence using the BWA-MEM algorithm implemented in BWA v2.0 and downsampled the resulting BAM files using SAMtools v1.17 to simulate four additional levels of coverage per individual: 5X, 10X, 15X, and 30X [36,37].

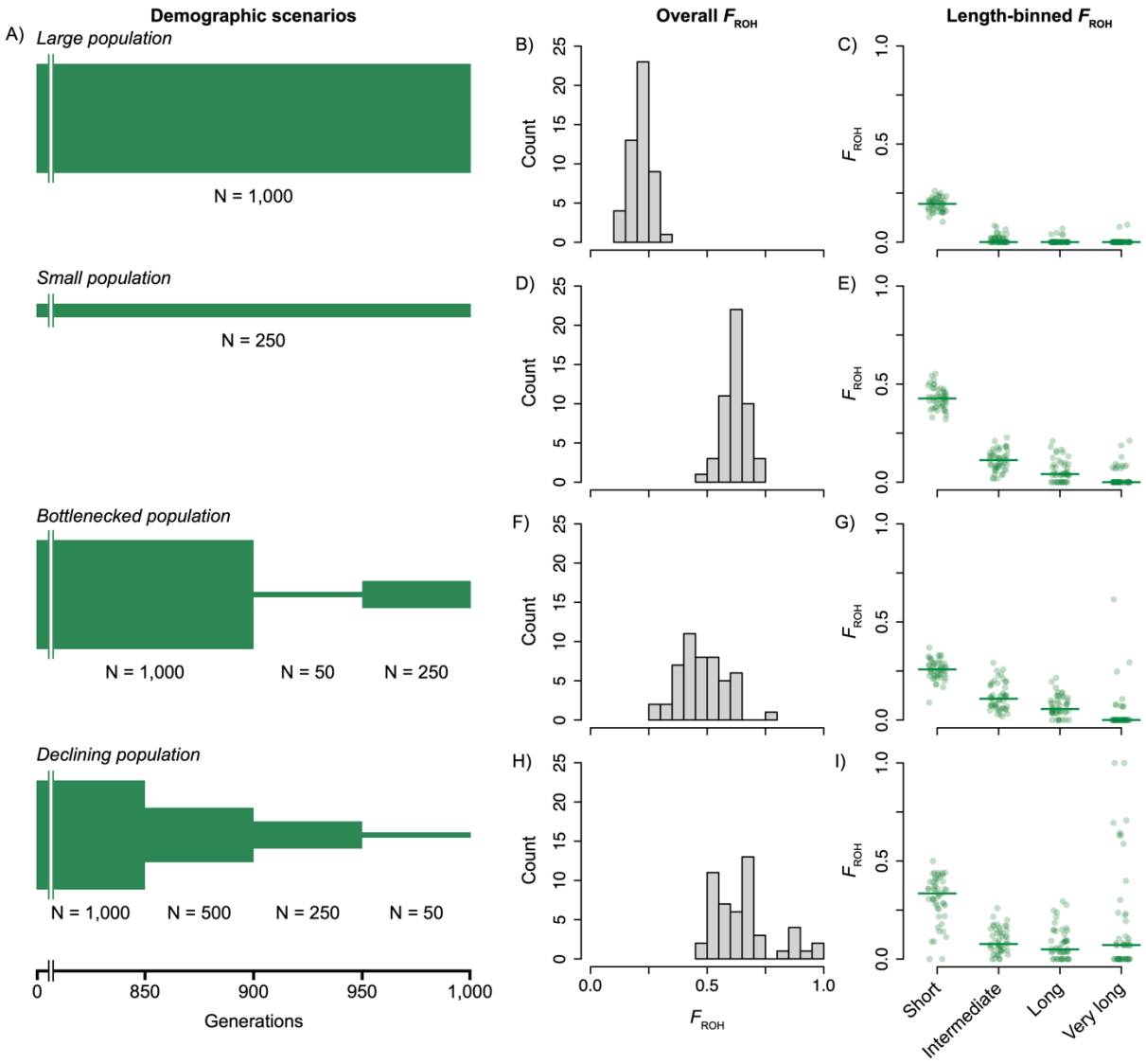


Fig. 1. A) Diagrams describing population size changes for each of the simulated demographic scenarios. Each diagram shows the final 1,000 generations of each scenario, which was preceded in each by a burn-in of 10,000 generations with a population size of 10,000 diploid individuals. Population sizes for each interval are noted for each scenario. B,D,F,H) True overall F_{ROH} frequencies. C,E,G,I) Length bin-specific true F_{ROH} values; horizontal lines correspond to bin median values.

123 For each sorted BAM file, we called genotypes using the 'HaplotypeCaller' algorithm in
124 Genomic Variant Call Format (GVCF) mode as implemented in GATK v4.1.9.0 [38]. For each
125 level of coverage, individual GVCF files were combined using 'CombineGVCFs' and genotyped
126 using 'GenotypeGVCFs'. We applied 'VariantFiltration' to these VCF files in GATK to flag
127 SNPs with low variant confidence (QualByDepth < 2), exhibiting strand bias (FisherStrand >
128 40), or with low mapping quality (RMSMappingQuality < 20). Finally, SNPs failing these filters
129 and indels were removed using 'SelectVariants.' Although population history and genetic dataset
130 characteristics can influence SNP filtering decisions, we chose to apply consistent filtering
131 across all demographic scenarios to instead focus on how changing software parameters affect
132 ROH inference. The effects of various SNP-filtering strategies on ROH inference have been
133 addressed elsewhere [24,39].

134 <u>ROH calling: hidden Markov model approach (BCFtools)</u>

135 We applied the same ROH calling approaches to all multisample VCF files produced from the
136 simulated data set using two of the programs most commonly applied to non-model species.
137 First, we tested an extension of the BCFtools software package, BCFtools/RoH v1.17 [40]. This
138 program uses a hidden Markov model to detect regions of autozygosity, requiring only a VCF
139 file for all samples, population allele frequency information, and an optional recombination map.
140 Because additional genetic information is not likely to be available for many wild populations,
141 we relied on allele frequencies calculated from each of our sample sets. The main decision faced
142 when running BCFtools/RoH is whether to estimate autozygous regions using called genotypes
143 or genotype likelihood values. We tested the effects of this decision on ROH estimation by either
144 including the *--GTs-only* setting to limit inference based on genotypes (hereafter, BCFtools
145 Genotypes) or omitting it and allowing genotype likelihood values to be considered (hereafter,

146 BCFtools Likelihoods) (Table 1). There are also two HMM options that can be set: *--hw-to-az*,
147 which sets the transition probability of Hardy-Weinberg to autozygous state, and *--az-to-hw*,
148 which sets the transition probability of autozygous to Hardy-Weinberg state. The default values
149 for these two probabilities are 6.7×10^{-8} and 5×10^{-9}, respectively. However, BCFtools/RoH can
150 be run in Viterbi training mode (*--viterbi-training* option) to estimate these two probabilities for
151 a specific data set. In total for BCFtools/RoH, we applied both methods, Genotypes and
152 Likelihoods, using default transition probabilities and population-specific probabilities calculated
153 using the *--viterbi-training* option and a custom R script (available in the GitHub repository for
154 this project) to each demographic scenario and coverage level.

155 **Table 1.** Parameter values applied during ROH calling for both simulated and empirical data. For PLINK, a total of 486 combinations
156 were tested. PLINK default values are underlined. ROH = run of homozygosity.
157

Program	Parameter	Abbreviation	Values	Description
BCFtools/RoH	--GTs-only	--	30	If set, uses genotypes only and ignores likelihood values
	--viterbi-training	--	$\underline{1 \times 10^{-10}}$	If set, runs BCFtools/RoH in Viterbi training mode to estimate HMM transition probabilities; value sets convergence threshold
	--hw-to-az	--	$\underline{6.7 \times 10^{-8}}$	Transition probability from Hardy-Weinberg to autozygous state
	--az-to-hw	--	$\underline{5 \times 10^{-9}}$	Transition probability from autozygous to Hardy-Weinberg state
PLINK	--homozyg-window-het	phwh	0, <u>1</u>, 2	Number of heterozygous sites allowed within a window; default = 1 heterozygous site
	--homozyg-window-missing	phwm	2, <u>5</u>, 50	Number of missing calls allowed in a window; default = 5 missing calls
	--homozyg-window-snp	phws	<u>50</u>, 100, 1000	Scanning window length in SNPs; default = 50 SNPs
	--homozyg-density	phzd	<u>50</u>	Minimum density in kb (*i.e.*, maximum inverse density (kb/variant); *e.g.*, to specify minimum 1 SNP per 50 kb, set to 50); default = 50 kb
	--homozyg-gap	phzg	500, <u>1000</u>	Threshold distance in kb at which to split a ROH into two if two SNPs are too far apart; default = 1000 kb
	--homozyg-window-threshold	phwt	0.01, <u>0.05</u>, 0.1	Proportion of overlapping windows that must be called homozygous to assign any SNP to a ROH; default = 0.05
	--homozyg-snp	phzs	10, <u>100</u>, 1000	Minimum number of variants that must be included in a ROH of minimum length --homozyg-kb to report it; default = 100 SNPs
	--homozyg-kb	phzk	100	Required minimum length of sequence (in kb) spanned by number of homozygous sites specified by --homozyg-snp; default = 1000 kb

158 <u>ROH calling: sliding window approach (PLINK)</u>

159 We tested a large number of parameter value combinations in PLINK v1.90b6.26 [41,42].
160 Unlike BCFtools/RoH, PLINK employs a sliding window approach to ROH identification: for
161 each window placement, SNPs are examined for conformity to the PLINK parameter values
162 (*e.g.*, fewer than the number of heterozygous or missing calls allowed). It is then determined, for
163 each SNP, whether a sufficient proportion of windows overlapping that SNP are homozygous
164 and thus, whether the SNP is determined to be located within in a ROH. PLINK has multiple
165 parameters that can be set by the user, and we initially tested a total of 486 combinations of six
166 of these parameters for each level of coverage (see Table 1 for list of parameters, initial values,
167 and parameter descriptions). To explore the effects of these combinations on ROH inference, we
168 applied an iterative approach designed by Mathur et al. [43] to explore through visualization how
169 varying parameter settings influence overall F_{ROH} estimates across samples within each
170 population and coverage level. For each iteration, we performed four steps:

171 1. Run PLINK with all possible combinations of different parameters to be tested,
172 ultimately generating a matrix of parameter values (predictor variables) and inferred
173 F_{ROH} (response variable) for each sample.
174 2. Create a linear model for each combination of parameter values ($F_{ROH} = a + b_1x_1 +$
175 $\cdots + b_nx_n + e$; where b_i = weight of parameter x_i), where the values of parameter x_i are
176 standardized to 1.
177 3. Extract standardized rank regression coefficients (SRC) from the linear regression
178 models using the *sensitivity* package in R and visualize sensitivity indices (SRC_i) to rank
179 weights of each parameter [44].

4. If $SRC_i \approx 0$ with little individual variation, then set the parameter *i* to the default value. If SRC_i is > 0 or < 0, then consider the effect described by SRC_i (*i.e.*, whether increasing the value of the parameter increases or decreases F_{ROH} and how SRC_i varies with called F_{ROH}) and either define a new set of parameter values to test or select a value from the tested set.

We began the first iteration by reading the results from the initial 486 combinations of parameter values for each demographic scenario into R v4.0.3 [29]. After examining the output figures, we selected new sets of parameter values to test for each demographic scenario (S2 Table).

Data summarization and statistical analyses

Output files from BCFtools/RoH and the final PLINK runs were read into R for summarization and statistical analyses. We also read in true ROH data (*i.e.*, start and end coordinates for known ROHs ≥ 100 kb in length) and calculated true F_{ROH} values for each individual. We filtered all called ROHs to retain ROHs ≥ 100 kb in length and calculated inferred F_{ROH} for each demographic scenario, individual, coverage level, and method. To describe relationships between true F_{ROH} and called F_{ROH} values, we constructed a linear model for each method and coverage level with true F_{ROH} as the predictor variable and called F_{ROH} as the response variable. For each model, we calculated the 95% confidence intervals (CIs) for the slope and *y*-intercept parameters using the *confint* function in R. To determine whether true and called F_{ROH} values differed for each model, we tested whether the model's *y*-intercept differed from zero and whether the slope differed from one (*i.e.*, whether the 95% CIs included zero or one, respectively). We also used the *y*-intercept and slope parameters to compare the results obtained from each BCFtools methods when using default vs. Viterbi-trained HMM transition probabilities within each demographic scenario and, for all three approaches, to determine

203 whether each method over- or underestimated true F_{ROH} at each coverage level and how the
204 degree of over- or underestimation changed with increasing true F_{ROH} values and across
205 demographic scenarios.

206 To further evaluate the accuracy of each ROH identification method, we also calculated
207 false negative (*i.e.*, failing to call a ROH present in an individual) and false positive (*i.e.*, calling
208 a ROH that was not present in an individual) rates for called ROHs. We began by identifying
209 overlap between true and called ROHs on a per-position basis by summing the number of bases
210 covered by both the true ROH and called ROH(s). From this information, we calculated (i) the
211 false negative rate: the total chromosomal length covered by true ROHs but not by called ROHs
212 divided by the total length of true ROHs; and (ii) the false positive rate: the total chromosomal
213 length covered by called ROHs but not by true ROHs divided by the total chromosomal length
214 not covered by true ROHs. For each demographic scenario, method, and level of coverage, we
215 calculated median false positive and negative rates and compared these medians and the 50%
216 quantiles between all method and coverage level combinations to provide insight into method-
217 specific differences in ROH calling errors.

218 We calculated F_{ROH} for ROHs in four different length bins to explore how ROH
219 identification methods may differ in their capabilities to accurately call ROHs of different sizes.
220 We defined length bins as:

221 (i) Short ROHs: ≥ 100 kb and < 500 kb in length;
222 (ii) Intermediate ROHs: ≥ 500 kb and < 1 Mb in length;
223 (iii) Long ROHs: ≥ 1 Mb and < 2 Mb in length; and
224 (iv) Very long ROHs: ≥ 2 Mb in length.

225 We examined how F_{ROH} for each bin changed with increasing coverage and also how patterns of
226 over- and underestimation of F_{ROH} varied with increasing coverage by subtracting true F_{ROH}
227 from called F_{ROH} for each individual. For each method, level of coverage, and length bin, we
228 compared mean called F_{ROH} – true F_{ROH} and the 95% CI around these means (estimated using
229 the *quantiles* function in R), with CIs < 0 indicating underestimation of true F_{ROH} and CIs > 0
230 indicating overestimation. We further explored relationships between true and called ROHs by
231 examining how true and called ROHs overlap. We tabulated how many true ROHs each called
232 ROH overlaps (or contains) and vice-versa for each unique combination of demographic
233 scenario, ROH detection method, coverage level, and ROH length bin.

234 *Part II: Empirical data*

235 Data curation and genotype calling

236 To test the effects of program and parameter value selection on identifying ROHs from empirical
237 data, we analyzed publicly available whole genome sequencing data for a species of
238 conservation concern, the Tasmanian devil (BioProject PRJNA549794 in NCBI's Sequence
239 Read Archive; [45]). From the full data set, we selected the 15 individuals from this data set with
240 the highest number of reads. The accession numbers and relevant metadata for each set of
241 sequences are provided in S2 Table. Adapters and low-quality bases were trimmed from raw
242 sequences using Trim Galore v0.6.6 [46], and cleaned reads were mapped to the chromosome-
243 level mSarHar1.11 *S. harrissii* reference genome (NCBI GenBank accession
244 GCA_902635505.1) using BWA-MEM [36].

245 We used Qualimap v2.2.1 to determine mean coverage per individual from each sorted
246 BAM file [47]. These results were used to calculate the downsampling proportions required to
247 approximate 5X, 10X, 15X, and 30X coverage for each individual. Following downsampling,

248 BAM files were processed in the same manner as for the simulated data, with additional SNP
249 filtering criteria applied in VCFtools v0.1.17 [48], including filtering SNPs within 5 bp of indels
250 and requiring minor allele frequencies ≥ 0.05 and < 20% missing data across individuals.

251 ROH calling and sensitivity analyses

252 We called ROHs from the final multisample VCF files using the same approaches as for the
253 simulated data. We called ROHs in two ways, (i) using BCFtools/RoH (*i.e.*, relying on
254 genotypes or on genotype likelihood values and using default or Viterbi-trained HMM transition
255 probabilities) and (ii) testing 486 parameter combinations in PLINK at each level of coverage
256 and visualizing the effect of altering each parameter following the same sensitivity analysis
257 process described above. Parameter values for all iterations tested are provided in S1 Table with
258 additional details provided for the empirical data in the Supporting Information.

259 Data summarization and statistical analysis

260 Output files from BCFtools/RoH and the final PLINK runs for the empirical data were read into
261 R for summarization and statistical analyses. Following the approach we used for the simulated
262 data, we filtered all called ROHs to retain ROHs ≥ 100 kb in length and calculated inferred F_{ROH}
263 for each individual in each scenario, coverage level, and method. We also calculated F_{ROH} for
264 ROHs in four different length bins, where length bins were defined as described above for the
265 simulated data. To compare results across methods, coverage levels, and ROH lengths, we
266 calculated mean length-specific F_{ROH} values and compared the 95% CIs around these means
267 among methods and coverage levels.

Results

Part I: Simulated data

Data collection and curation

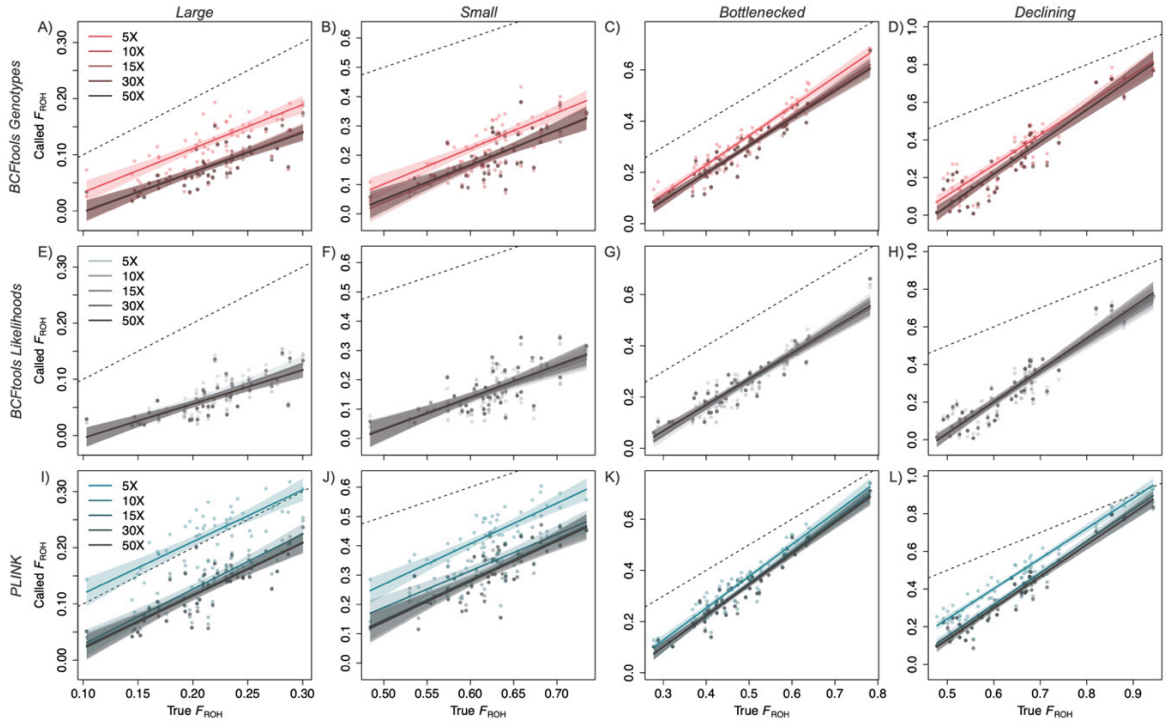
For each simulated population, all analyses were based on 50 randomly sampled individuals. For the declining population, extremely low heterozygosity in two individuals precluded estimation of HMM transition probabilities via Viterbi training, and all downstream analyses for this demographic scenario were based on the remaining 48 individuals. Mean heterozygosity (± SD) varied across scenarios: large population mean = $3.17 \times 10^{-5} \pm 1.26 \times 10^{-5}$; bottlenecked population mean = $3.05 \times 10^{-5} \pm 5.81 \times 10^{-6}$; small population mean = $2.87 \times 10^{-5} \pm 1.22 \times 10^{-5}$; declining population mean = $2.44 \times 10^{-5} \pm 8.81 \times 10^{-6}$. Mean F_{ROH} generally increased with decreasing mean heterozygosity: large population mean = 0.21; bottlenecked population mean = 0.48; small population mean = 0.62; declining population mean = 0.65 (Fig. 1B,D,F,H). For all demographic scenarios, length-specific F_{ROH} mostly decreased with increasing ROH length with some notable exceptions in the bottlenecked and declining population scenarios (Fig. 1C,E,G,I). For these two final sets of samples, a few individuals had high proportions of their genomes located in very long ROHs (Fig. 1G,I), consistent with very recent inbreeding events in each population. Following downsampling and SNP filtering, the final mean coverage across all demographic scenarios was 4.87, 9.77, 14.69, and 29.20, and 44.99 for the 5X, 10X, 15X, 30X, and 50X VCF files, respectively.

ROH calling

Before comparing results from the three main methods—BCFtools Genotypes, BCFtools Likelihoods, and PLINK—we examined the effect of using either default or Viterbi-trained HMM transition probabilities on the results produced by each BCFtools method. Within each

291 method, we compared overall F_{ROH} values obtained when using default HMM transition
292 probabilities (default values: *--hw-to-az* = 6.7 x 10^{-8}, *--az-to-hw* = 5 x 10^{-9}) against those
293 calculated when using probabilities estimated via Viterbi training (mean values across all
294 groups: *--hw-to-az* = 9.2 x 10^{-7}, *--az-to-hw* = 3.8 x 10^{-6}; S3 Table). (In practice, mean
295 probabilities were calculated and applied for each demographic scenario and BCFtools method
296 combination, but note the small magnitude of variation in Viterbi-trained probabilities across all
297 groups relative to the difference between these probabilities and the program defaults.) The slope
298 and intercept parameters never significantly differed between the two sets of results across all
299 demographic scenarios and coverage levels, but for most comparisons, the slope and intercept
300 parameters of the model built using Viterbi-trained results were closer to 1 and 0, respectively,
301 than for the default values model (S4 Table; S1-4 Figs.). Using Viterbi-trained transition
302 probabilities slightly increased false positive rates but also decreased false negative rates (S5-8
303 Figs.). Because of these slight, if not significant, improvements in overall F_{ROH} estimation
304 accuracy and false negative rates, we chose to proceed with Viterbi-trained HMM transition
305 probabilities for both BCFtools methods and present only these results below (we also describe
306 one additional benefit of Viterbi-trained probabilities below).
307 We used our simulated data set and linear models to determine whether each approach tends to
308 over- or underestimate true F_{ROH}. All three methods underestimated F_{ROH}, with all model
309 intercepts across coverage levels negative and different from zero, except for PLINK when
310 calling ROHs in the large population at 5X coverage (*i.e.*, no other 95% CIs for intercepts
311 included zero; Fig. 2; S5 Table). Within each demographic scenario, the three methods produced
312 similar results with respect to overall F_{ROH} accuracy. For example, within all four scenarios,
313 intercepts did not differ among methods or coverage levels, with two exceptions at 5X coverage

when using PLINK: intercepts differed between this group and others within the large, declining, and majority of coverage levels for small populations. Perhaps the most visualizing striking points of comparison across demographic scenarios are the slope estimates specific to each. Within each method, demographic scenarios were consistently ordered by slope estimates, with the declining population having the largest slope estimate (for BCFtools, all slopes > 1 with no 95% CIs including one), followed by the small, bottlenecked, and large populations (for the large population, all BCFtools slopes < 1, all PLINK slope 95% CIs include one; Fig. 2; S5 Table). Thus, accuracy of overall F_{ROH} estimates calculated for the declining population varied the most with respect to true F_{ROH} (*i.e.*, of all demographic scenarios, slope parameter estimates for this population most greatly differed from one), followed by the large population. With respect to true F_{ROH}, accuracy went up with increasing true F_{ROH} for the declining population and went down with increasing true F_{ROH} for the large population.



328 **Fig. 2.** The relationship between true and inferred F_{ROH} values depends on inference method and
329 population demographic scenario. Each regression line represents linear model results for a
330 single level of coverage with the shaded areas representing 95% confidence intervals. Each point
331 represents data for a single simulated individual. Dashed line is 1:1 line and *x*- and *y*-axes are
332 consistent within each demographic scenario. Note the differing slopes across demographic
333 scenarios (*e.g.*, among panels A-D) and differing overall accuracies across methods (*e.g.*,
334 differing distances between regression lines and 1:1 line among panels D, H, and L).

335 We calculated false negative (*i.e.*, failing to call a ROH present in an individual) and
336 false positive (*i.e.*, calling a ROH that was not present in an individual) rates to further assess
337 each method's accuracy. With respect to false positive rates, PLINK performed poorly relative to
338 the other two methods, with false positive rates consistently higher than BCFtools Genotypes
339 and BCFtools Likelihoods (Fig. 3A). Across all demographic scenarios and coverage levels,
340 median false positive rates ranged from 6.14×10^{-7} to 0.031 for BCFtools Genotypes, from 4.54
341 $\times 10^{-8}$ to 0.007 for BCFtools Likelihoods, and from 0.008 to 0.112 for PLINK. For all three
342 methods, increasing coverage to 10X corresponded to decreasing false positive rates, but rates
343 did not continue to improve at coverages above 10X (*i.e.*, 50% quantiles around median values
344 overlap at 10-50X within each method and demographic scenario; results for all coverage levels
345 provided in S9-10 Figs.). Variation in false positive rates among samples at each coverage level
346 was smallest for BCFtools Likelihoods, followed by BCFtools Genotypes, with PLINK showing
347 the greatest variation across samples (summary statistics provided in S6 Table).

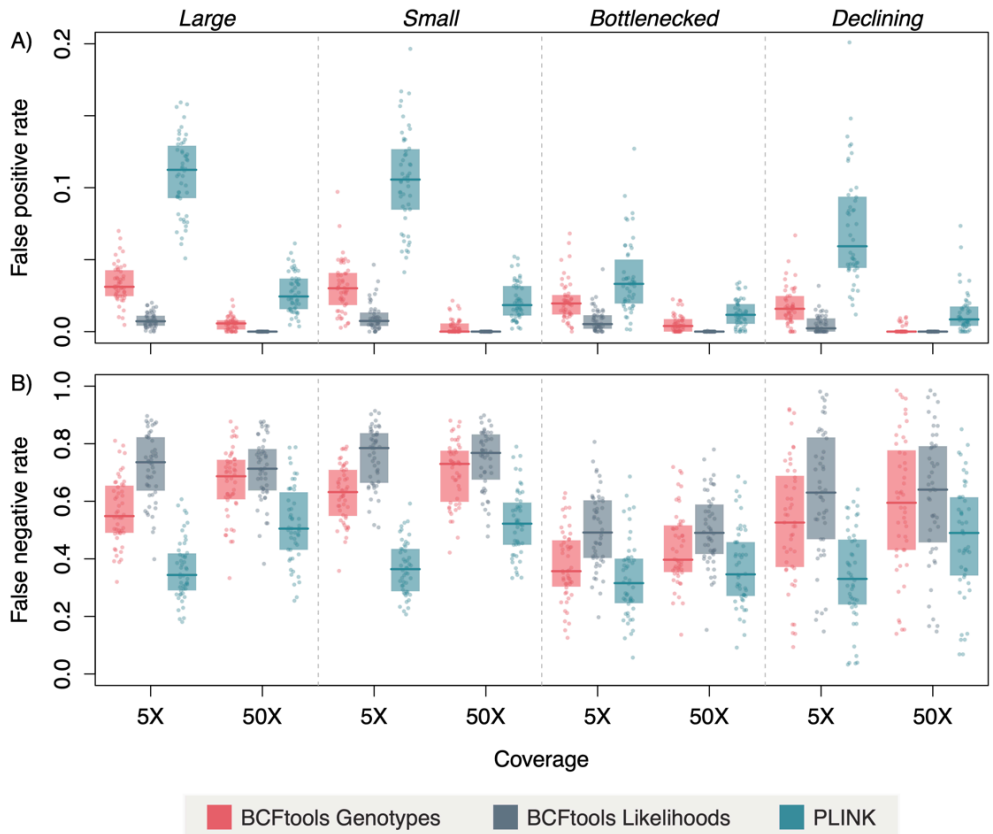


Fig. 3. PLINK outperforms BCFtools with respect to false negative rates, but underperforms with respect to false positive rates. A) False positive (*i.e.*, incorrectly calling a base position as being located in a ROH) and B) false negative (*i.e.*, failing to identify a base position as being located in a ROH) rates across demographic scenarios and methods. Horizontal lines indicate median values and shaded boxes are 50% quantiles. Note difference in scale of *y*-axis between panels A and B. Both BCFtools approaches outperform PLINK with respect to false positive rates but the reverse is true for false negative rates. Increasing coverage corresponds to decreasing false positive rates and to increasing false negative rates. Values displayed for 5X and 50X coverages; data for all coverage levels presented in Fig. S9-10.

Patterns in false negative rates were generally in the opposite direction and magnitude to those we observed with false positives: both BCFtools methods performed poorly relative to PLINK, with BCFtools Genotypes producing slightly lower rates (median range across all demographic scenarios and coverage levels = 0.36-0.74) than BCFtools Likelihoods (median range = 0.49-0.78; Fig. 3B). PLINK exhibited lower false negative rates than the BCFtools approaches (median range = 0.32-0.53) and less variation among samples at each coverage level.

364 All three methods produced false negative rates that increased slightly at 10X coverage relative
365 to 5X, with rates leveling off at 10X and higher coverage levels. Examples of false negative and
366 false positive scenarios can be seen in S11 Fig., which illustrates a full chromosome of true and
367 called ROHs for one exemplar individual from each demographic scenario.

368 We also examined how true and called values of F_{ROH} varied for ROHs of different
369 lengths. For the simulated data, all three methods almost always underestimated the proportion
370 of the genome located in short ROHs across demographic scenarios, with the 95% CI less than
371 zero for all tests other than PLINK at 5X coverage (Fig. 4; all levels of coverage presented in
372 S12-15 Figs.). For longer ROHs, results were variable and dependent on demographic scenario.
373 For example, all three methods appeared to approach accuracy for longer ROHs in the large
374 population, but this is likely due to the relative paucity of longer ROHs in this demographic
375 scenario (Fig. 1). For the bottlenecked and declining populations (wherein some individuals had
376 very long ROHs), both PLINK and BCFtools Genotypes tended to overestimate F_{ROH} for very
377 long ROHs. PLINK generally produced the highest overestimates of F_{ROH} and the most variation
378 across samples of the three approaches, followed by BCFtools Genotypes. Finally, one coverage-
379 related trend emerged across ROH length categories, methods, and demographic scenarios, with
380 F_{ROH} estimates calculated at 5X coverage often exceeding estimates calculated at higher
381 coverage levels. Across all length bins and demographic scenarios, individual estimates of
382 length-specific F_{ROH} calculated at 5X were greater than those calculated at 10X for 37%, 53%,
383 and 53% of BCFtools Likelihoods, BCFtools Genotypes, and PLINK estimates, respectively.

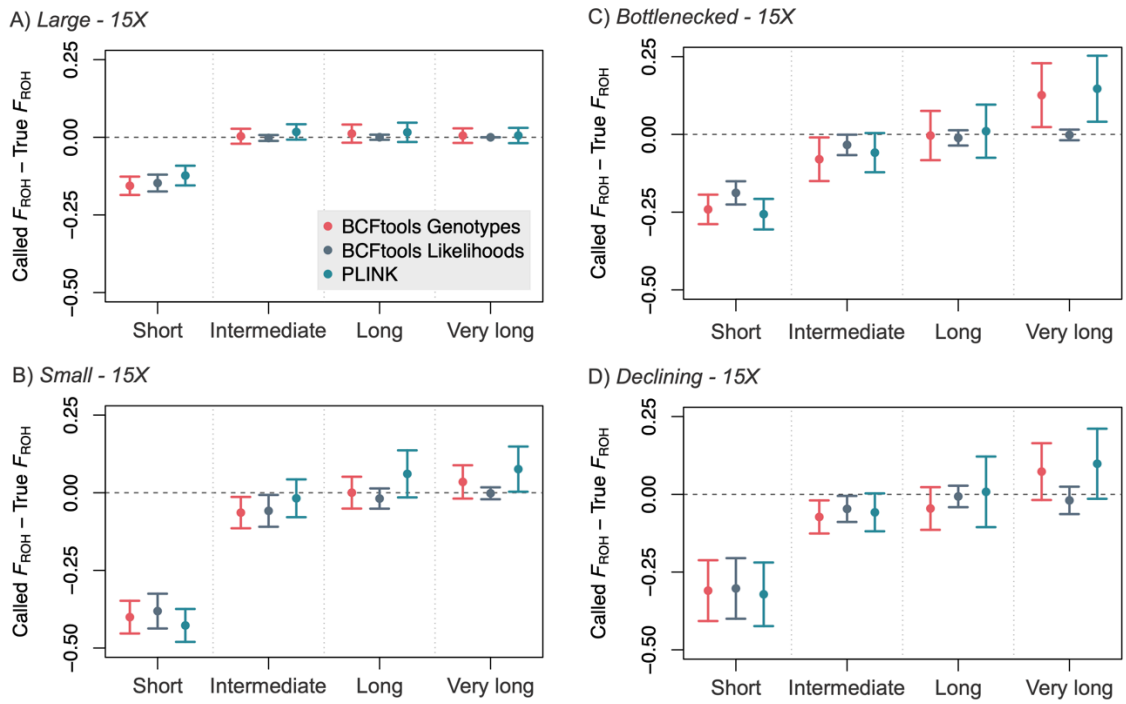


Fig. 4. Increasing true ROH length corresponds to increasing detection. Called F_{ROH} – True F_{ROH} by length bin and demographic scenario at 15X (results for all coverage levels presented in Fig. S12-15). For BCFtools Genotypes and PLINK, F_{ROH} for short ROHs is consistently underestimated whereas F_{ROH} for very long ROHs is overestimated when these ROHs are present. BCFtools Likelihoods does not overestimate ROHs in any length bin.

To further investigate how called ROHs correspond to true ROHs, we identified regions of overlap between true and called ROHs within each individual and at each coverage level using a unique identifier for each true and called ROH. We found no instances of true ROHs being split into multiple called ROHs, but multiple true ROHs were often lumped together into a single called ROH. This pattern held true for all three methods at most coverage levels (Fig. 5; S16 Fig.). For BCFtools Genotypes and PLINK, increasing coverage did not appear to ameliorate this problem (*i.e.*, the mean number of true ROHs lumped into a single called ROH changed very little with increasing coverage; Fig. 5C,D). However, for BCFtools Likelihoods, the number of true ROHs contained in a single called ROH decreased with increasing coverage, reaching a 1:1 ratio at 30X. Across methods, the mean number of true ROHs combined into a

single called ROH increased with increasing ROH length, except for BCFtools Likelihoods at coverage levels ≥ 30X (S16 Fig.). This lumping can be seen in Fig. 5B.

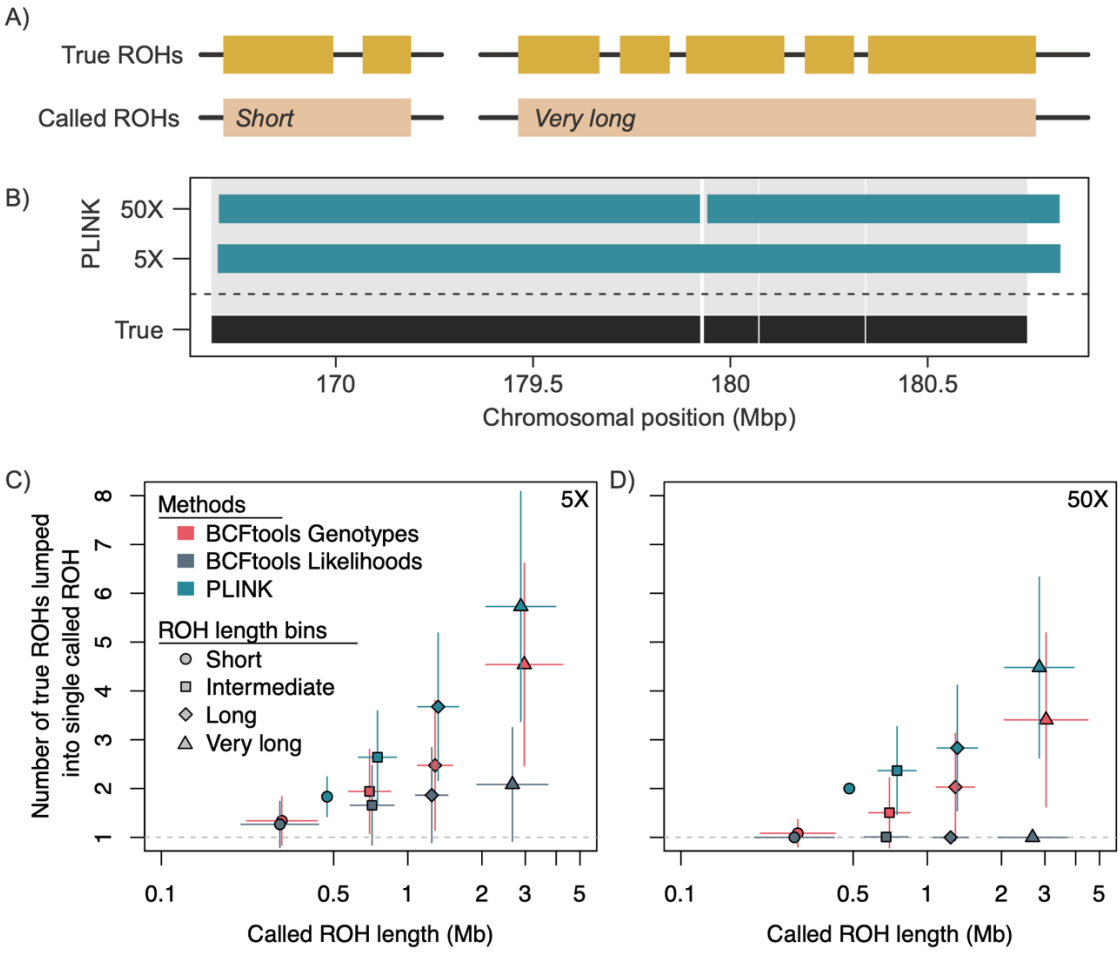


Fig. 5. All three methods tested combine multiple true ROHs into single called ROHs, with increasing coverage only providing improvements for BCFtools Likelihoods. A) Diagram illustrating this lumping issue. B) Examples of this issue at 5X and 50X in a single simulated individual drawn from the small population demographic scenario. C) Number of true ROHs combined into a single called ROH for ROHs of varying lengths when called by all three methods at 5X and (D) at 50X in the small population (results for all coverage levels and demographic scenarios provided in Fig. S16). Points correspond to mean values and vertical and horizontal error lines indicate 95% confidence intervals. Dashed horizontal line corresponds to $y = 1$ (a 1:1 relationship between numbers of true and called ROHs).

412 *Part II: Empirical data*

413 Genotype and ROH calling results

414 For the 15 sets of reads we downloaded from NCBI, the mean number of reads per sample was
415 9.75×10^8. Read mapping rates to the mSarHar1.11 *S. harrissii* reference genome were high,
416 with an average of 95.4% of reads mapped and properly paired. For the final sets of filtered
417 SNPs (n = 1,532,598), average depth across samples was 48.43 for the full coverage set (*i.e.*, not
418 downsampled) and 6.37, 11.84, 16.63, 30.75 for the 5X, 10X, 15X, and 30X downsampled sets,
419 respectively (S2 Table). Given the modest improvements in F_{ROH} accuracy conferred by using
420 Viterbi-trained HMM transition probabilities rather than the default values for the simulated
421 data, we calculated Viterbi-trained values for both BCFtools methods applied to the empirical
422 data.

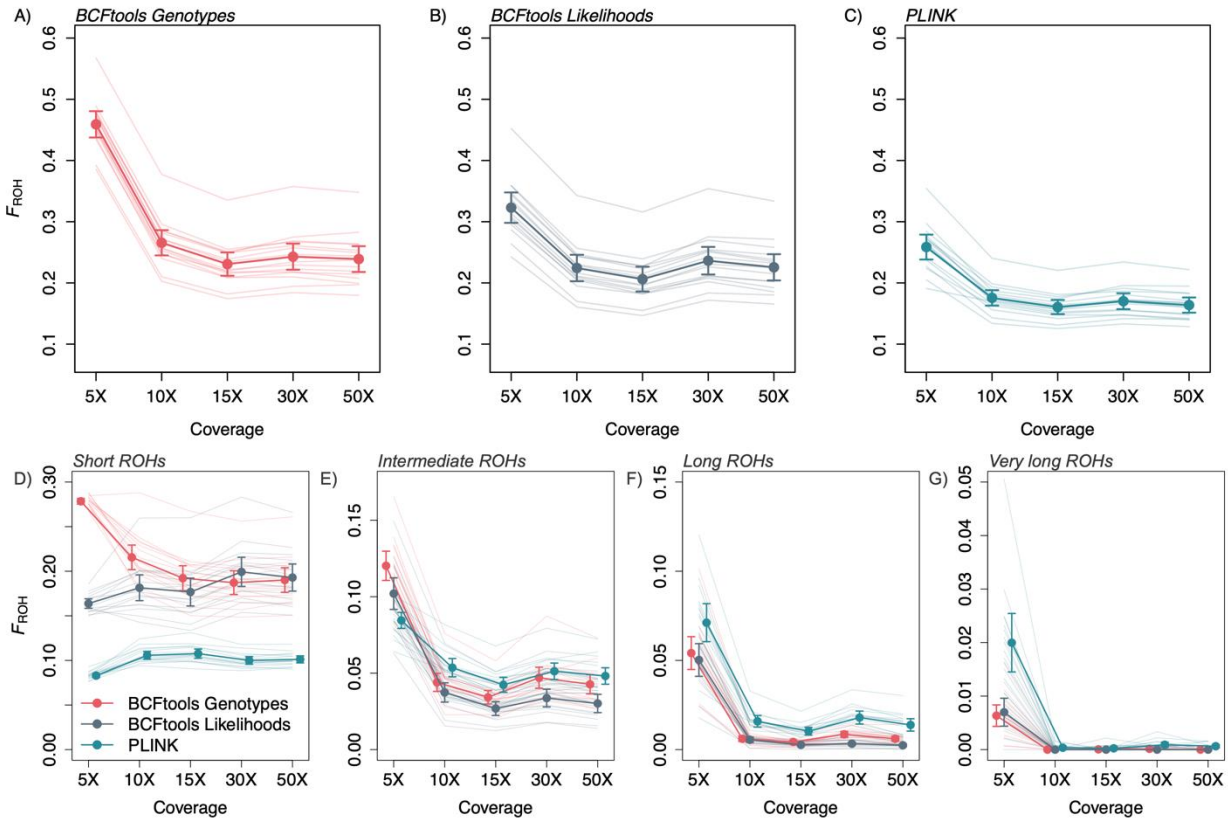
423 Across methods, F_{ROH} estimated at 5X coverage was significantly higher than F_{ROH}
424 estimates at all higher levels of coverage (95% CIs did not overlap, Fig. 6A-C). At 5X coverage,
425 F_{ROH} estimates differed significantly among all three methods, with the lowest estimate
426 produced by PLINK and the highest by BCFtools Genotypes. For all higher levels of coverage,
427 F_{ROH} estimates produced by the two BCFtools methods did not differ from one another but
428 estimates from both methods differed from those produced by PLINK, with PLINK again
429 providing lower estimates.

430 When comparing how the three methods estimated length-specific F_{ROH} values, patterns
431 varied across ROH length categories. For short ROHs, PLINK produced much lower F_{ROH}
432 estimates then the two BCFtools methods across all coverage levels, with differences among the
433 three methods significant (*i.e.*, non-overlapping 95% CIs) at 5X coverage (Fig. 6D). For longer
434 ROHs, results of the three methods were generally concordant, with all methods producing

435 higher estimates of F_{ROH} at 5X than at higher coverages (Fig. 6D-G). However, PLINK

436 produced much higher estimates of F_{ROH} for very long ROHs at 5X than the other two methods,

437 with a mean value more than double those inferred by the BCFtools methods (Fig. 6G).



438

439 **Fig. 6.** When applied to the empirical data set, the three ROH calling methods differ greatly in
440 their inferences, particularly at 5X coverage. A-C) Overall F_{ROH} and (D-G) length bin F_{ROH}
441 results for each method and level of coverage, with means and 95% confidence intervals
442 indicated by points and vertical lines, respectively. Lighter background lines indicate results for
443 individual samples.

444 **Discussion**

445 In this study, we systematically analyzed the discrepancies that arise in ROH detection when

446 populations with varied demographic histories are analyzed and highlight effects on downstream

447 interpretations associated with these differences. This is important because the demographic

448 scenarios affected the accuracy, with declining populations showing the highest variance in F_{ROH}
449 estimation relative to true F_{ROH} values. Additionally, ROH detection accuracy varied with
450 sequencing coverage; improvements were observed up to a 10X coverage threshold beyond
451 which the benefits plateaued. Our findings emphasize the impact of choosing appropriate
452 computational parameters and underscore the challenges in standardizing ROH detection
453 methods values (*e.g.*, Saremi et al., 2019; Grossen et al., 2020; von Seth et al., 2021; Mueller et
454 al., 2022), highlighting the necessity for robust, systematic comparisons to enhance
455 reproducibility and reliability in genomic studies.

456 We demonstrated the exploratory utility of the sensitivity analysis process we followed to
457 select parameter values for our data (see the Supporting Information for corresponding
458 information for the empirical data). This process is important because disparate sequencing data
459 characteristics are likely to require different parameter values, meaning that it may not be
460 appropriate to use the values we used herein when analyzing other data. For example, studies
461 that use fewer SNPs (*e.g.*, populations that are less genetically diverse, studies with reduced
462 sequencing efforts) should test the effects of altering the minimum SNP density required on
463 ROH inference results. Sensitivity analysis provides a relatively quick and convenient way to
464 visualize how different parameter values affect F_{ROH} estimates for an entire data set and the
465 degree of variation in those effects across individuals.

466 *Inference methodology biases* F_{ROH} *values*

467 Between the two BCFtools methods when considering identified ROHs of all lengths, Genotypes
468 produced more accurate overall F_{ROH} estimates than Likelihoods, with F_{ROH} estimates from
469 Likelihoods also increasingly diverging from the true F_{ROH} value with increasing true F_{ROH} (Fig.
470 1A,B). For populations expected to have considerable variation in F_{ROH} among individuals (*e.g.*,

471 a population that has remained somewhat small for an extended period of time with evidence of
472 recent immigration), applying the BCFtools Likelihoods approach could result in increasingly
473 skewed values for the individuals with the highest levels of inbreeding. For example, using the
474 linear model parameters estimated for 15X coverage, an individual with a true F_{ROH} of 0.10
475 would be assigned an inferred F_{ROH} of 0.01 (difference = -0.09), whereas an individual with a
476 true F_{ROH} value of 0.40 would be assigned 0.23 (difference = -0.17). This could be particularly
477 problematic when dealing with species or populations of conservation concern because the
478 individuals with the highest true F_{ROH} also have the largest magnitude of error, meaning that
479 concerning signals of inbreeding could go undetected.

480 In contrast to the underestimations produced by the BCFtools/RoH methods, the sliding
481 window approach implemented in PLINK overestimated F_{ROH}. This was particularly evident at
482 5X coverage where F_{ROH} estimates differ more from their true values than any other method and
483 coverage level combination in our study (Fig. 1C). However, at coverages above 5X, PLINK
484 produced better estimates than either BCFtools approach (*i.e.*, in our linear models, intercepts for
485 PLINK at 10X-50X are closer to zero than for either BCFtools method and 95% CIs for these
486 parameter estimates do not overlap with any BCFtools intercept 95% CIs). In the context of
487 endangered species conservation, small overestimations of F_{ROH} may be more desirable than
488 underestimations because these are likely to be more conservative (*i.e.*, indicating more close
489 inbreeding than is present in reality) in many situations. Importantly though, as with BCFtools
490 Likelihoods, F_{ROH} estimates diverged from true F_{ROH} at increasing values of true F_{ROH}.
491 However, these values diverged at a much lower rate in the PLINK estimates compared to
492 BCFtools Likelihoods. Again using our simulated data as a model, an individual with a true

493 F_{ROH} value of 0.40 would be estimated to have an F_{ROH} of 0.46 (difference = 0.06) when
494 estimated at 10X-50X with PLINK.

495 For the two BCFtools methods, patterns of underestimation were consistent with these
496 approaches' high false negative rates and low false positive rates (Fig. 3). Conversely, PLINK
497 produced higher false positive rates and lower false negative rates than either BCFtools method,
498 consistent with overestimation of F_{ROH}. In terms of absolute difference between true and called
499 F_{ROH} values, PLINK outperformed BCFtools at 10X coverage and above, suggesting that PLINK
500 will often provide the most robust estimate of F_{ROH}. However, at lower coverages (5X-10X),
501 BCFtools Genotypes could be considered, given that this method produces F_{ROH} estimates closer
502 to true F_{ROH} than either PLINK or BCFtools Likelihoods. On the other hand, the underestimates
503 produced by this approach are likely related to the high false negative rates we observed
504 (especially relative to PLINK), and the appearance of convergence on true F_{ROH} may be due to
505 length-specific ROH calling rates by this program (see below) and therefore highly variable
506 across populations. It is important to note that while the trends we describe may be consistent
507 with some empirical results (*e.g.*, [5]), individual variation in genomic characteristics exerts
508 strong influence over F_{ROH} inference results as suggested by comparisons between our simulated
509 and empirical results.

510 *Demographic history influences* F_{ROH} *value accuracy*

511 The relationship between true and inferred F_{ROH} values varied with demographic scenario,
512 demonstrating the potential for uncertainty when incorporating these inbreeding values into
513 management action plans. For example, in the large population where true F_{ROH} values are small
514 (< 0.3), inferred F_{ROH} accuracy decreased with increasing values of true F_{ROH} for the BCFtools
515 methods, with both tending to underestimate F_{ROH} with increasing true F_{ROH} (Fig. 2A,E). The

516 opposite trend occurred for the declining population, where true F_{ROH} values exceed 0.9, with all
517 three methods approaching more accurate F_{ROH} estimates with increasing values of true F_{ROH}
518 (Fig. 2D,H,L). These contrasting patterns are likely due to different ROH length distributions in
519 these two demographic scenarios. In the declining population, ROHs in the individuals with the
520 highest overall F_{ROH} values tend to comprise very long ROHs (Fig. 1I). Consistent
521 overestimation of F_{ROH} for these very long ROHs by BCFtools Genotypes and PLINK directly
522 leads to higher inferred F_{ROH} estimates in these individuals. However, individuals with lower
523 overall F_{ROH} values have a greater proportion of their ROHs in shorter length categories, which
524 are underestimated by all three methods (Fig. 4D), leading to inferred overall F_{ROH} estimates that
525 diverge from true F_{ROH} as true F_{ROH} decreases (Fig. 2D,H,L). In the large population, although
526 some individuals have higher true F_{ROH} values, most ROHs are still short (Fig. 1C), and likely to
527 be underestimated by all three methods. Therefore, as true F_{ROH} increases, so do the number of
528 true short ROHs and the cumulative effects of underestimation across methods, leading to
529 overall F_{ROH} estimates that diverge from true F_{ROH} as true F_{ROH} increases.
530 Overall, we found that the accuracy of ROH detection significantly varies across
531 populations with different demographic histories. Our analysis revealed that declining
532 populations exhibited the highest variance in F_{ROH} estimation relative to true F_{ROH} values. In
533 contrast, large populations tended to show more stable estimates but with a general tendency for
534 underestimation of true F_{ROH}, particularly as true F_{ROH} values increased. These findings
535 highlight the need for an analysis approach that considers demographic history of the population.
536 For conservationists and researchers working with genomic data from populations known or
537 suspected to be in decline, it is vital to interpret ROH analysis outputs with the breadth of
538 demographic scenarios in mind. When demographic history is unknown and cannot be

539 estimated, the distribution of F_{ROH} present in the dataset could be compared to the distributions
540 simulated here (Fig. 1B,D,F,H), or to simulated distributions of F_{ROH} with more similar natural
541 history parameters. Although tedious, these approaches will minimize the effects of
542 misestimation of F_{ROH}, reducing the risk of exacerbating genetic diversity loses in populations
543 and species that are most at risk to extinction.

544 *Coverage ≤ 10X strongly influences called ROH lengths*

545 For the empirical data at 5X coverage, relative to higher coverage levels, all methods
546 consistently produced lower F_{ROH} estimates for short ROHs and higher F_{ROH} estimates for longer
547 ROHs (Fig. 7). The overcalling of intermediate to very long ROHs at 5X could be related to the
548 ROH-lumping issue noted in the simulated results, wherein multiple true ROHs are erroneously
549 called as a single ROH (Fig. 6). While we cannot confirm the accuracy of F_{ROH} inference for the
550 empirical data, comparisons between results generated at 5X and higher levels of coverage are
551 consistent with the simulated results, suggesting that these patterns are accurate (Fig. 2). For the
552 Tasmanian devil samples we analyzed, the results from 5X coverage suggest much more
553 frequent, recent inbreeding than the results from ≥ 10X coverage, painting a much more dire
554 demographic scenario than is presented when more coverage is obtained. Therefore, if one of the
555 goals of a whole-genome sequencing project is to assess recent or historical patterns of
556 inbreeding from ROH lengths, ~10X coverage appears to be a minimum requirement for
557 generating robust ROH inferences.

558 *Patterns of under- or overestimation may vary with ROH length distributions*

559 In our simulated and empirical data, we observed patterns indicating that underlying ROH length
560 distributions influence the patterns of F_{ROH} under- and overestimation. For example, even though
561 PLINK produced higher F_{ROH} estimates than both BCFtools methods for the simulated data and

562 PLINK and BCFtools Genotypes produced statistically indistinguishable estimates for the
563 empirical data (Fig. 2), length-specific F_{ROH} estimates suggest that differences in underlying true
564 ROH length distributions between the simulated and empirical data may be responsible for the
565 differences in relative F_{ROH} results we observed. For the simulated data, BCFtools Genotypes
566 increasingly overestimated F_{ROH} as ROH length increased (Fig. 5, Fig. S2), with increasing
567 numbers of true ROHs erroneously combined into single called ROHs (Fig. 6). Although we
568 cannot know the true ROH length distributions for the empirical data, long ROHs were called at
569 higher frequencies in the empirical data relative to the simulated data. The tendency of BCFtools
570 Genotypes to overestimate F_{ROH} for long ROHs combined with the presence of more called long
571 ROHs in our empirical data set may have minimized differences in overall F_{ROH} estimates
572 between BCFtools Genotypes and PLINK in the empirical results relative to the simulated
573 results (Fig. 2). Increased frequencies of long ROHs in the empirical data may have also led to
574 greater differences in F_{ROH} between 5X and 10X across all three methods for the empirical
575 results compared to the simulated results (Fig. 2). All three methods call significantly more
576 intermediate to very long ROHs from the empirical data at 5X than at 10X (Fig. 7B-D), and this
577 may be related to the increased false positive rates we noted at 5X in the simulated data. These
578 results again illustrate the effects of a population's or individual's actual ROH complement,
579 which is determined by typically unknown demographic and breeding patterns, on the relative
580 reliability and utility of ROH identification programs.

581 Particularly for endangered species with potentially complicated demographic histories,
582 interpreting ROH patterns in a population may be most accurate when multiple tools are used to
583 create an integrated picture. For example, comparing overall and length-specific F_{ROH} estimates
584 between BCFTools/RoH and PLINK can be used to understand the underlying length

585 distributions; an abundance of shorter ROHs would be indicated by higher overall F_{ROH}
586 estimates in PLINK compared to BCFtools Genotypes but similar length-specific F_{ROH} patterns,
587 whereas a ROH complement comprising many longer ROHs would be indicated by similar
588 overall F_{ROH} estimates between PLINK and BCFtools Genotypes but higher intermediate to very
589 long F_{ROH} estimates from BCFtools Genotypes related to PLINK. These accurate assessments of
590 past and ongoing inbreeding could then be used to inform management options, such as
591 translocations to ameliorate close inbreeding.

592 *Conclusions*

593 Inferring the presence and characteristics of ROHs can shed important light on population
594 demographic histories, detect inbreeding depression when combined with fitness information,
595 and even disentangle the mechanisms underlying or loci contributing to inbreeding depression.
596 However, given the variation in ROH-calling accuracy demonstrated here, we caution against
597 direct comparisons of F_{ROH} values generated from different data types or sources or using
598 different inference parameters. Data from disparate studies could be combined and re-analyzed
599 in a standardized fashion, although special attention should be paid to variation in reference
600 genome assembly quality for interspecific comparisons [49]. Regardless of the number of data
601 sets to be analyzed, we strongly recommend that studies relying on ROH inference (i) employ at
602 least two ROH-calling programs and interpret their results with each method's biases in mind
603 and (ii) compare multiple parameter value combinations via sensitivity analysis, taking care to
604 vary parameters of particular relevance to a data set.

605 **Acknowledgments**

606 We thank members of the Willoughby lab for constructive comments on earlier versions of this
607 manuscript as well as M Kardos and 2 anonymous reviewers.

Funding

This material is based on work supported by the National Science Foundation Postdoctoral Research Fellowships in Biology Program under Grant No. 2010251 to AMH. This work was also supported by the U.S. Department of Agriculture, National Institute of Food and Agriculture, Hatch project 1025651 to JRW.

Data Availability

Code for all bioinformatic analyses available at https://github.com/avril-m-harder/roh_inference_testing and https://github.com/kennethb22/roh_parameter_project_kk. All FASTA files for simulated individuals and final VCF files (for simulated and empirical data and all coverage levels) will be uploaded to a public repository upon manuscript acceptance.

Author Contributions

AMH and JRW conceived the study. AMH, KBK, and SM performed formal data analyses. AMH wrote the manuscript with input and final approval from KBK, SM, and JRW.

References

1. Diffenbaugh NS, Field CB. Changes in ecologically critical terrestrial climate conditions. Science. 2013;341: 486–492. doi:10.1126/science.1237123

2. Haddad NM, Brudvig LA, Clobert J, Davies KF, Gonzalez A, Holt RD, et al. Habitat fragmentation and its lasting impact on Earth's ecosystems. Sci Adv. 2015;1: e1500052. doi:10.1126/sciadv.1500052

3. Crnokrak P, Roff DA. Inbreeding depression in the wild. Heredity. 1999;83: 260–270. doi:10.1038/sj.hdy.6885530

4. Keller L, Waller D. Inbreeding effects in wild populations. Trends in Ecology & Evolution. 2002;17: 230–241. doi:10.1016/S0169-5347(02)02489-8

5. Robinson JA, Räikkönen J, Vucetich LM, Vucetich JA, Peterson RO, Lohmueller KE, et al. Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. Sci Adv. 2019;5: eaau0757. doi:10.1126/sciadv.aau0757

634 6. Stoffel MA, Johnston SE, Pilkington JG, Pemberton JM. Genetic architecture and lifetime
635 dynamics of inbreeding depression in a wild mammal. Nat Commun. 2021;12: 2972.
636 doi:10.1038/s41467-021-23222-9

637 7. Gibbs HL, Grant PR. Inbreeding in Darwin's medium ground finches (*Geospiza fortis*).
638 Evolution. 1989;43: 1273–1284.

639 8. Liberg O, Andrén H, Pedersen H-C, Sand H, Sejberg D, Wabakken P, et al. Severe
640 inbreeding depression in a wild wolf (*Canis lupus*) population. Biol Lett. 2005;1: 17–20.
641 doi:10.1098/rsbl.2004.0266

642 9. Saccheri I, Kuussaari M, Kankare M, Vikman P, Fortelius W, Hanski I. Inbreeding and
643 extinction in a butterfly metapopulation. Nature. 1998;392: 491–494. doi:10.1038/33136

644 10. Slate J, Kruuk LEB, Marshall TC, Pemberton JM, Clutton-Brock TH. Inbreeding depression
645 influences lifetime breeding success in a wild population of red deer (*Cervus elaphus*). Proc
646 R Soc Lond B. 2000;267: 1657–1662. doi:10.1098/rspb.2000.1192

647 11. O'Grady JJ, Brook BW, Reed DH, Ballou JD, Tonkyn DW, Frankham R. Realistic levels of
648 inbreeding depression strongly affect extinction risk in wild populations. Biological
649 Conservation. 2006;133: 42–51. doi:10.1016/j.biocon.2006.05.016

650 12. Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW. Genomics advances the study
651 of inbreeding depression in the wild. Evol Appl. 2016;9: 1205–1218.
652 doi:10.1111/eva.12414

653 13. Stoffel MA, Johnston SE, Pilkington JG, Pemberton JM. Mutation load decreases with
654 haplotype age in wild Soay sheep. Evolution Letters. 2021;5: 187–195.
655 doi:10.1002/evl3.229

656 14. Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zöllner S, Rosenberg NA, et al. Long runs of
657 homozygosity are enriched for deleterious variation. The American Journal of Human
658 Genetics. 2013;93: 90–102. doi:10.1016/j.ajhg.2013.05.003

659 15. Liu S, Westbury MV, Dussex N, Mitchell KJ, Sinding M-HS, Heintzman PD, et al. Ancient
660 and modern genomes unravel the evolutionary history of the rhinoceros family. Cell. 2021;
661 S0092867421008916. doi:10.1016/j.cell.2021.07.032

662 16. Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, et al. Complete genomes
663 reveal signatures of demographic and genetic declines in the woolly mammoth. Current
664 Biology. 2015;25: 1395–1400. doi:10.1016/j.cub.2015.04.007

665 17. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows
666 into population history and trait architecture. Nat Rev Genet. 2018;19: 220–234.
667 doi:10.1038/nrg.2017.109

668 18. Saremi NF, Supple MA, Byrne A, Cahill JA, Coutinho LL, Dalén L, et al. Puma genomes
669 from North and South America provide insights into the genomic consequences of
670 inbreeding. Nat Commun. 2019;10: 4769. doi:10.1038/s41467-019-12741-1

671 19. Grossen C, Guillaume F, Keller LF, Croll D. Purging of highly deleterious mutations
672 through severe bottlenecks in Alpine ibex. Nat Commun. 2020;11: 1001.
673 doi:10.1038/s41467-020-14803-1

674 20. von Seth J, Dussex N, Díez-del-Molino D, van der Valk T, Kutschera VE, Kierczak M, et al.
675 Genomic insights into the conservation status of the world's last remaining Sumatran
676 rhinoceros populations. Nat Commun. 2021;12: 2393. doi:10.1038/s41467-021-22386-8

677 21. Mueller SA, Prost S, Anders O, Breitenmoser-Würsten C, Kleven O, Klinga P, et al.
678 Genome-wide diversity loss in reintroduced Eurasian lynx populations urges immediate
679 conservation management. Biological Conservation. 2022;266: 109442.
680 doi:10.1016/j.biocon.2021.109442

681 22. Ceballos FC, Hazelhurst S, Ramsay M. Assessing runs of homozygosity: a comparison of
682 SNP array and whole genome sequence low coverage data. BMC Genomics. 2018;19: 106.
683 doi:10.1186/s12864-018-4489-0

684 23. Duntsch L, Whibley A, Brekke P, Ewen JG, Santure AW. Genomic data of different
685 resolutions reveal consistent inbreeding estimates but contrasting homozygosity landscapes
686 for the threatened Aotearoa New Zealand hihi. Mol Ecol. 2021; mec.16068.
687 doi:10.1111/mec.16068

688 24. Meyermans R, Gorssen W, Buys N, Janssens S. How to study runs of homozygosity using
689 PLINK? A guide for analyzing medium density SNP data in livestock and pet species.
690 BMC Genomics. 2020;21: 94. doi:10.1186/s12864-020-6463-x

691 25. Haller BC, Messer PW. Evolutionary modeling in SLiM 3 for beginners. Hernandez R,
692 editor. Molecular Biology and Evolution. 2019;36: 1101–1109.
693 doi:10.1093/molbev/msy237

694 26. Haller BC, Messer PW. SLiM 3: Forward genetic simulations beyond the Wright-Fisher
695 model. Hernandez R, editor. Molecular Biology and Evolution. 2019;36: 632–637.
696 doi:10.1093/molbev/msy228

697 27. Feigin CY, Newton AH, Doronina L, Schmitz J, Hipsley CA, Mitchell KJ, et al. Genome of
698 the Tasmanian tiger provides insights into the evolution and demography of an extinct
699 marsupial carnivore. Nat Ecol Evol. 2018;2: 182–192. doi:10.1038/s41559-017-0417-y

700 28. Patton AH, Margres MJ, Stahlke AR, Hendricks S, Lewallen K, Hamede RK, et al.
701 Contemporary Demographic Reconstruction Methods Are Robust to Genome Assembly
702 Quality: A Case Study in Tasmanian Devils. Molecular Biology and Evolution. 2019;36:
703 2906–2921. doi:10.1093/molbev/msz191

29. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available: https://www.R-project.org/

30. Thompson EA. Identity by descent: Variation in meiosis, across genomes, and in populations. Genetics. 2013;194: 301–326. doi:10.1534/genetics.112.148825

31. Hasselgren M, Dussex N, Seth J, Angerbjörn A, Olsen R, Dalén L, et al. Genomic and fitness consequences of inbreeding in an endangered carnivore. Mol Ecol. 2021;30: 2790–2799. doi:10.1111/mec.15943

32. Sánchez-Barreiro F, Gopalakrishnan S, Ramos-Madrigal J, Westbury MV, Manuel M, Margaryan A, et al. Historical population declines prompted significant genomic erosion in the northern and southern white rhinoceros (*Ceratotherium simum*). Mol Ecol. 2021;30: 6355–6369. doi:10.1111/mec.16043

33. Xie H-X, Liang X-X, Chen Z-Q, Li W-M, Mi C-R, Li M, et al. Ancient demographics determine the effectiveness of genetic purging in endangered lizards. Harris K, editor. Molecular Biology and Evolution. 2022;39: msab359. doi:10.1093/molbev/msab359

34. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28: 593–594. doi:10.1093/bioinformatics/btr708

35. Andrews S. FastQC: A quality control tool for high throughput sequence data. 2015. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

36. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. 2013 [cited 8 May 2017]. Available: http://arxiv.org/abs/1303.3997

37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

38. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

39. Howrigan DP, Simonson MA, Keller MC. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. BMC Genomics. 2011;12: 460. doi:10.1186/1471-2164-12-460

40. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32: 1749–1751. doi:10.1093/bioinformatics/btw044

41. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci. 2015;4: 7. doi:10.1186/s13742-015-0047-8

42. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007;81: 559–575. doi:10.1086/519795

43. Mathur S, Tomeček JM, Tarango-Arámbula LA, Perez RM, DeWoody JA. An evolutionary perspective on genetic load in small, isolated populations as informed by whole genome resequencing and forward-time simulations. Evolution. 2023;77: 690–704. doi:10.1093/evolut/qpac061

44. Iooss B, Da Veiga S, Janon A, Pujol G. sensitivity: Global Sensitivity Analysis of Model Outputs. 2021. Available: https://CRAN.R-project.org/package=sensitivity

45. Wright BR, Farquharson KA, McLennan EA, Belov K, Hogg CJ, Grueber CE. A demonstration of conservation genomics for threatened species management. Mol Ecol Resour. 2020;20: 1526–1541. doi:10.1111/1755-0998.13211

46. Krueger F. Trim Galore. Cambridge, U.K.: Babraham Institute; 2019.

47. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32: 292–294.

48. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

49. Brüniche-Olsen A, Kellner KF, Anderson CJ, DeWoody JA. Runs of homozygosity have utility in mammalian conservation and evolutionary studies. Conserv Genet. 2018;19: 1295–1307. doi:10.1007/s10592-018-1099-y