Detectability of runs of homozygosity is influenced by analysis parameters as well as 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: Avril M. Harder, College of Forestry, Wildlife and Environment, Auburn University, Auburn, Alabama, USA, (773) 688-8564, avrilharder@gmail.com

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. However, selecting an appropriate program and parameter values for such analyses is often difficult for species of conservation concern, for which little is often known about population demographic histories or few high-quality genomic resources are available. Herein, we analyze simulated and empirical data sets to demonstrate the downstream effects of program selection 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 program-specific tendencies. This is particularly important for the management of endangered species, as some program and parameter combinations consistently underestimate inbreeding signals in the genome, substantially undermining conservation initiatives. Based on our conclusions, 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.

Running title: Testing runs of homozygosity inference tools

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 (Diffenbaugh & Field, 2013; Haddad et al., 2015). 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 (*e.g.*, Crnokrak & Roff, 1999; Robinson et al., 2019). 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 (Gibbs & Grant, 1989; Liberg et al., 2005; Saccheri et al., 1998; Slate, Kruuk, Marshall, Pemberton, & Clutton-Brock, 2000). 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 (Keller & Waller, 2002; O'Grady et al., 2006). However, applying whole genome sequencing strategies to identify runs of homozygosity (ROHs; genomic regions where both inherited haplotypes are identical) opens up lines of inquiry previously not accessible via pedigree- or microsatellite-based studies (Kardos, Taylor, Ellegren, Luikart, & Allendorf, 2016).

One long-standing question important for ongoing conservation efforts is whether inbreeding depression primarily occurs as a result of increasing homozygosity of recessive deleterious alleles or absences of heterozygote advantage (Hedrick & Garcia-Dorado, 2016). Analyses of genome-wide data have addressed parts of this question by quantitatively documenting and illustrating inbreeding depression (Huisman, Kruuk, Ellis, Clutton-Brock, & Pemberton, 2016; Harrisson et al., 2019; Stoffel, Johnston, Pilkington, & Pemberton, 2021) and identifying support for increasing homozygosity of strongly deleterious mutations as a genetic mechanism of inbreeding depression (Robinson et al., 2019; Stoffel et al., 2021). Theoretical predictions regarding the genetic mechanisms of inbreeding depression mitigation have also been empirically tested. For example, coupling genomic and fitness data reveals positive correlations between ROH length and mutational load resulting from genetic purging (Stoffel, Johnston, Pilkington, & Pemberton, 2021; Szpiech et al., 2013), suggesting that ROH length distribution data can provide actionable insight for managers. Analyses of ROHs in ancient samples have even clarified the genomic and demographic changes preceding historic species extinction events (Liu et al., 2021; Palkopoulou et al., 2015). Despite these advances and insights, causative mechanisms of inbreeding depression remain unclear for many taxonomic groups, and this can hinder management efforts that seek to mitigate fitness declines in wild populations.

Estimating ROHs can provide crucial insights into populations' evolutionary histories, but these histories can in turn affect which ROH-calling software and combination of parameter values are most appropriate. For example, the settings best suited for inferring ROHs in a small, long-isolated population experiencing high levels of inbreeding would not be suitable for individuals sampled from a large, genetically diverse population because underlying sources of error in these two scenarios are very different (*e.g.*, differences in ROH length distributions, numbers of variable sites, expected minor allele frequencies; Ceballos, Joshi, Clark, Ramsay, & Wilson, 2018). While some studies include comparisons of results from multiple programs or parameter value combinations (*e.g.*, Saremi et al., 2019; Grossen, Guillaume, Keller, & Croll, 2020; von Seth et al., 2021; Mueller et al., 2022), many more studies rely on default settings and do not explore the effects of varying these parameter values on their results. Without extensive knowledge of a population's demographic history (*e.g.*, prevalence and degree of consanguineous mating or immigration), it can be challenging to determine the most appropriate combination of parameter values, and it is always impossible to know how close the resulting estimates approximate reality.

We address this challenge by leveraging simulated and empirical genomic sequencing data to compare ROH identification programs and test a systematic process for determining software parameter values. We focus on whole-genome sequencing data because although previous studies have examined ROH inference for data sets with lower marker densities (Ceballos, Hazelhurst, & Ramsay, 2018; Duntsch, Whibley, Brekke, Ewen, & Santure, 2021; Meyermans, Gorssen, Buys, & Janssens, 2020), the insights from these previous works do not cover the spectrum of issues encountered when analyzing whole genome data. Specifically, we test a wide array of setting combinations for two programs commonly used in population genomic studies—PLINK and BCFtools/RoH—and, for PLINK, apply a sensitivity analysis to evaluate the effects of parameter values on ROH inference. Based on these results, we outline a set of recommendations for ROH estimation at multiple levels of sequencing effort typical of conservation genomics studies. These guidelines are particularly relevant when population histories are poorly understood or when a reference genome assembly is more fragmented than for a typical model species—two common conditions for species that are targets of conservation action.

Methods

Part I: Simulated data

<u>Data generation and genotype calling</u>

We used SLiM v3.6 and modifications of Recipe 7.3 distributed with SLIMgui to simulate a population (N = 10,000), wherein each individual consisted of a homologous pair of 30-Mb chromosomes (Haller & Messer, 2019a, 2019b). The population was simulated for 10,000 generations, followed by a bottleneck to 250 individuals that was sustained for 5,000 additional generations. Recombination rate (1×10^{-7} per site per generation), base mutation rate (1.75×10^{-7} per site per generation), and population parameters were selected to produce a final population with F_{ROH} values ranging from 0.075 to 0.440 when considering ROHs ≥ 100 kb in length. The VCF file output from SLiM was converted to FASTA sequence files using a custom script in R v4.0.3 and a haploid ancestral sequence produced by SLiM (R Core Team, 2020).

Using the known genotypes for all individuals, we generated two files: (i) a record of all true heterozygous sites and (ii) the start and end coordinates for all true ROHs ≥ 100 kb in length. We imposed this lower limit on ROH length because ROHs less than 100 kb in length likely originated in a single common ancestor approximately 500 generations ago (assuming a recombination rate of 1 cM/1 Mb; Thompson, 2013), and would not be expected to influence contemporary individual fitness as strongly as more recently acquired autozygous segments (Stoffel et al., 2021). This threshold has also gained popularity in population genetics studies of non-model species (Robinson et al., 2019; Hasselgren et al., 2021; Sánchez-Barreiro et al., 2021; Xie et al., 2022), and we follow this convention for all downstream analyses.

For 100 randomly selected individuals, FASTQ read files were generated from each of the two FASTA files representing homologous chromosomes using ART (version MountRainier-2016-06-05) (Huang, Li, Myers, & Marth, 2012). We simulated 150-bp paired-end reads using the HiSeq 2500 error model to a depth of 50X per individual (*i.e.*, 25X per homologous chromosome). Each FASTQ file was quality-checked using FASTQC v0.11.9 (Andrews, 2015). We aligned reads to the ancestral sequence using the BWA-MEM algorithm implemented in BWA v0.7.17 and downsampled the resulting BAM files using SAMtools v1.11 to simulate four additional levels of coverage per individual: 5X, 10X, 15X, and 30X (Li, 2013; Li et al., 2009).

For each sorted BAM file, we called genotypes using the 'HaplotypeCaller' algorithm in Genomic Variant Call Format (GVCF) mode as implemented in GATK v4.1.9.0 (McKenna et al., 2010). For each level of coverage, individual GVCF files were combined using 'CombineGVCFs' and genotyped using 'GenotypeGVCFs'. We applied 'VariantFiltration' to these VCF files in GATK to flag SNPs with low variant confidence (QualByDepth < 2), exhibiting strand bias (FisherStrand > 40), or with low mapping quality (RMSMappingQuality < 20). Finally, SNPs failing these filters and indels were removed using 'SelectVariants.'

ROH calling: hidden Markov model approach (BCFtools)

We applied the same ROH calling approaches to all multisample VCF files produced from the simulated data set using two of the programs most commonly applied to non-model species. First, we tested an extension of the BCFtools software package, BCFtools/RoH v1.11 (Narasimhan et al., 2016). This program uses a hidden Markov model to detect regions of autozygosity, requiring only a VCF file for all samples, population allele frequency information, and an optional recombination map. Because additional genetic information is not likely to be available for many wild populations, we relied on allele frequencies calculated from each of our sample sets. The main decision faced when running BCFtools/RoH is whether to estimate autozygous regions using called genotypes or genotype likelihood values. We tested the effects of this decision on ROH estimation by either including the --GTs-only setting to limit inference based on genotypes (hereafter, BCFtools Genotypes) or omitting it and allowing genotype likelihood values to be considered (hereafter, BCFtools Likelihoods) (Table 1).

ROH calling: sliding window approach (PLINK)

We tested a large number of parameter value combinations in PLINK v1.90b6.26 (Chang et al., 2015; Purcell et al., 2007). Unlike BCFtools/RoH, PLINK employs a sliding window approach to ROH identification: for each window placement, SNPs are examined for conformity to the PLINK parameter values (*e.g.*, fewer than the number of heterozygous or missing calls allowed). It is then determined, for each SNP, whether a sufficient proportion of windows overlapping that SNP are homozygous and thus, whether the SNP is determined to be located within in a ROH. PLINK has multiple parameters that can be set by the user, and we initially tested a total of 486 combinations of six of these parameters for each level of coverage (see Table 1 for list of parameters, initial values, and parameter descriptions). We focus on how changing software parameters affect ROH inference rather than the effects of various SNP-filtering strategies, as these questions have been addressed elsewhere (Howrigan, Simonson, & Keller, 2011; Meyermans et al., 2020).

Before comparing the results from the two BCFtools/RoH approaches and PLINK, we had to select one set of PLINK parameter values. We applied an iterative approach designed by Mathur et al. (2021; non-peer-reviewed preprint) to identify a combination of parameter values that minimizes the effect of value selection on inferred F_{ROH} (*i.e.*, the bias in F_{ROH} inference due to each parameter value). For each iteration and level of coverage, we performed four steps:

1. Run PLINK with all possible combinations of different parameters to be tested, ultimately generating a matrix of parameter values (predictor variables) and inferred F_{ROH} (response variable) for each sample.
2. Create a linear model for each combination of parameter values ($F_{ROH} = a + b_1x_1 + \cdots + b_nx_n + e$; where b_i = weight of parameter x_i), where the values of parameter x_i are standardized to 1.
3. Extract standardized rank regression coefficients (SRC) from the linear regression models using the *sensitivity* package in R and visualize sensitivity indices (SRC_i) to rank weights of each parameter (Iooss, Da Veiga, Janon, & Pujol, 2021).
4. If $\mathrm{SRC}_i \approx 0$ 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 select 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 into R v4.0.3 (R Core Team, 2020). Details of the parameter value selection process for the simulated data are provided in Box 1. Briefly, we applied the four steps outlined above by examining the results from Iteration 1 (486 parameter value combinations) and noting that increasing the value of one parameter (*phwh*) had a positive effect on inferred F_{ROH} whereas increasing the values of two other parameters (*phws* and *phzs*) had negative effects on inferred F_{ROH}. For *phwh*, we allowed one heterozygous site per window to avoid (i) discarding a true homozygous window due to an erroneous heterozygous call and (ii) retaining too many spurious homozygous windows due to inclusion of true heterozygous calls. For *phws* (scanning window length in SNPs) and *phzs* (minimum number of SNPs that can comprise a ROH), we tested two additional sets of parameter values and used these outputs to select the values for *phws* and *phzs* that (Table S2 and Box 1).

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 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 determine whether each method over- or underestimated true $\mathrm{F}_{\mathrm{ROH}}$ at each coverage level, and how the degree of over- or underestimation changed with increasing true F_{ROH} values.

At each coverage level, we compared the mean F_{ROH} values among ROH identification methods to determine whether different methods produce significantly different results. We also compared mean F_{ROH} across coverage levels within each method to test whether coverage significantly affects inferred F_{ROH}. For each method and coverage level combination, we randomly sampled 15 individuals (to mirror the sample size for the empirical data, see below) from the 100 individuals with simulated genotypes, calculated mean F_{ROH}, and repeated this process 1,000 times. We generated 95% CIs around this mean using the 95% quantile of these 1,000 values. We interpreted non-overlapping 95% CIs as indicative of significant differences within and among ROH identification method and coverage levels.

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

We calculated F_{ROH} for ROHs in four different length bins to explore how ROH identification methods may differ in their capabilities to accurately call ROHs of different sizes. We defined length bins as: (i) 100 kb ≤ short ROHs < 250 kb; (ii) 250 kb ≤ intermediate ROHs < 500 kb; (iii) 500 kb ≤ long ROHs < 1 Mb; (iv) 1 Mb ≤ very long ROHs. We examined how F_{ROH}

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

Part II: Empirical data

Data curation and genotype calling

To test the effects of program and parameter value selection on identifying ROHs from empirical data, we analyzed publicly available whole genome sequencing data for a species of conservation concern, the Tasmanian devil (*Sarcophilus harrisii*; BioProject PRJNA549794 in NCBI's Sequence Read Archive; Wright et al., 2020). From the full dataset, we selected the 15 individuals from this data set with the highest number of reads. The accession numbers and relevant metadata for each set of sequences are provided in Table S1. Adapters and low-quality bases were trimmed from raw sequences using Trim Galore v0.6.6 (Krueger, 2019), and cleaned reads were mapped to the mSarHar1.11 *S. harrissii* reference genome (NCBI GenBank accession GCA_902635505.1) using BWA-MEM (Li, 2013).

We used Qualimap v2.2.1 to determine mean coverage per individual from each sorted BAM file (Okonechnikov, Conesa, & García-Alcalde, 2016). These results were used to calculate the downsampling proportions required to approximate 5X, 10X, 15X, and 30X coverage for each individual. Following downsampling, BAM files were processed in the same manner as for the simulated data, with additional SNP filtering criteria applied in VCFtools v0.1.17 (Danecek et al., 2011), including filtering SNPs within 5 bp of indels and requiring minor allele frequencies ≥ 0.05 and < 20% missing data across individuals.

ROH calling and sensitivity analyses

We called ROHs from the final multisample VCF files using the same approaches as for the simulated data. We called ROHs in two ways, (i) using BCFtools/RoH (*i.e.*, relying on genotypes or on genotype likelihood values) and (ii) testing 486 parameter combinations in PLINK at each level of coverage and identifying robust values for each parameter following the same sensitivity analysis process described above. Parameter values for all iterations tested are provided in Table S2 with additional details provided for the empirical data in the Supplementary Material.

Data summarization and statistical analysis

Output files from BCFtools/RoH and the final PLINK runs for the empirical data were read into R for summarization and statistical analyses. Following the approach we used for the simulated data, we filtered all called ROHs to retain ROHs ≥ 100 kb in length and calculated inferred F_{ROH} for each individual, coverage level, and method. We also calculated F_{ROH} for ROHs in four different length bins, where length bins were defined as: (i) 100 kb ≤ short ROHs < 500 kb; (ii) 500 kb ≤ intermediate ROHs < 1 Mb; (iii) 1 Mb ≤ long ROHs < 2 Mb; (iv) 2 Mb ≤ very long ROHs. To compare results across methods, coverage levels, and ROH lengths, we calculated mean F_{ROH} values and compared the 95% CIs around these means among methods and coverage levels.

Results

Part I: Simulated data

Data collection and curation

For the simulated data set, all analyses were based on 100 individuals randomly sampled from the small simulated population (N = 250) that underwent a strong bottleneck 5,000 generations ago. Mean heterozygosity for these 100 individuals was 7.68 x 10^{-5} (SD = 5.88 x 10^{-6}). After retaining only ROHs ≥ 100 kb in length, mean F_{ROH} was 0.151 (SD = 0.045) and ranged from 0.083 to 0.293. Following downsampling and SNP filtering, the final mean coverage was 4.80, 9.70, 14.62, and 28.91 for the 5X, 10X, 15X, and 30X downsampled sets, respectively.

ROH calling results

We used our simulated data set and linear models to determine whether each approach tends to over- or underestimate true F_{ROH}. Both of the BCFtools methods (Genotypes and Likelihoods) underestimated F_{ROH}, with all model intercepts across coverage levels negative and different from zero (*i.e.*, no 95% CIs for intercepts included zero; Fig. 1; Table S3). For BCFtools Genotypes, model slopes were approximately one (*i.e.*, all 95% CIs for slopes included one; Table S3), whereas the slopes of all BCFtools Likelihoods models were significantly less than one, indicating that F_{ROH} estimated using Likelihoods can vary relative to true F_{ROH}. PLINK tended to produce overestimates of F_{ROH}, but estimates at the highest coverage levels were accurate (*i.e.,* the 95% CI for model intercepts included zero at 30X and 50X). The 95% CI for the PLINK 5X model was larger and did not overlap the 95% CIs for the other PLINK coverage level model intercepts, indicating greater overestimation occurred using 5X PLINK compared to using PLINK at other coverages. PLINK model slopes did not differ from one at 5X or 10X, but did differ at 15X, 30X, and 50X, with these slope estimates exceeding one, again indicating that the estimated F_{ROH} varied with true F_{ROH}.

We also compared F_{ROH} values across methods and observed the largest differences for F_{ROH} calculated from 5X coverage data (Fig. 2A-C). We compared the 95% CIs around mean F_{ROH} and found that at 5X, BCFtools Likelihoods produced significantly smaller F_{ROH} estimates than BCFtools Genotypes, and both BCFtools estimates were smaller than PLINK's estimate. At all other coverage levels, mean F_{ROH} did not differ between the two BCFtools methods and PLINK again produced significantly higher F_{ROH} estimates. For both BCFtools approaches, there were no significant differences in mean F_{ROH} across coverage levels, but for PLINK, F_{ROH} estimated at 5X was significantly greater than estimates at higher coverage levels. For all methods and at all coverage levels, inferred mean F_{ROH} differed from the true mean F_{ROH} value (*i.e.*, none of the bootstrapped 95% CIs included the true mean F_{ROH} value). Raw results for all individuals are presented in Fig. S1.

We calculated false negative (*i.e.*, failing to call a ROH present in an individual) and false positive (*i.e.*, calling a ROH that was not present in an individual) rates to further assess each method's accuracy. With respect to false positive rates, PLINK performed poorly relative to the other methods, with median false positive rates of 0.078 for PLINK, 0.018 for BCFtools Genotypes, and 4.09×10^{-8} for BCFtools Likelihoods across all tested coverage levels (Fig. 3A). For all three methods, increasing coverage to 10X corresponded to decreasing false positive rates, but these tended to level off at high coverages. Variation in false positive rates among samples at each coverage level was smallest for BCFtools Likelihoods, followed by BCFtools Genotypes, with PLINK showing the greatest variation across samples (summary statistics provided in Table S4). Generally speaking, the patterns in false negative rates were 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 (overall median = 0.552) than BCFtools Likelihoods (overall median = 0.744; Fig. 3B). PLINK exhibited lower false negative rates than the BCFtools approaches (overall median = 0.165) and less variation among samples at each coverage level. All three methods produced false negative rates that increased with increasing coverage up to 10X. Examples of false negative and false positive scenarios can be seen in Fig. 4, which illustrates a 6-Mb window of true and called ROHs for one exemplar individual (full chromosome-level examples can be seen for three individuals in Fig. S2).

We also examined how true and called values of F_{ROH} varied for ROHs of different lengths. For the simulated data, all three methods almost always underestimated the proportion of the genome located in short ROHs, with the 95% CI less than zero for all tests other than PLINK at 5X coverage (Fig. 5A-C). For ROHs of intermediate, long, and very long lengths, all of the 95% CIs included zero. PLINK produced the highest overestimates of F_{ROH} and the most variation across samples of the three approaches, followed by BCFtools Genotypes. However, 95% CIs for BCFtools Likelihoods included zero for these three length bins, and variation among individuals decreased with both increasing coverage and increasing ROH length, suggesting increased accuracy with increasing depth and ROH length (Fig. 5B). PLINK and BCFtools Genotypes almost exclusively overestimated F_{ROH} for very long ROHs, even though most (94/100) of the simulated individuals had no very long true ROHs (Fig. 5D). Finally, one coverage-related trend emerged across ROH length categories and methods, with F_{ROH} estimates calculated at 5X coverage often exceeding estimates calculated at higher coverage levels. Across all length bins combined, individual estimates of F_{ROH} calculated at 5X were greater than those calculated at 10X for 34%, 56%, and 72% of BCFtools Likelihoods, BCFtools Genotypes, and PLINK estimates, respectively.

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 and at most coverage levels (Fig. 6). 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). 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 all three methods, the mean number of true ROHs combined into a single called ROH increased with increasing ROH length with the exception of BCFtools Likelihoods at coverage levels ≥ 30X (Fig. S4). Examples of this lumping tendency can be seen in Fig. 4 and Fig. S2.

Part II: Empirical data

Genotype and ROH calling results

For the 15 sets of reads we downloaded from NCBI, the mean number of reads per sample was 9.75×10^8. Read mapping rates to the mSarHar1.11 *S. harrissii* reference genome were high, with an average of 95.4% of reads mapped and properly paired. For the final sets of filtered SNPs (n = 1,532,598), average depth across samples was 48.43 for the full coverage set (*i.e.*, not downsampled) and 6.37, 11.84, 16.63, 30.75 for the 5X, 10X, 15X, and 30X downsampled sets, respectively (Table S1).

Across methods, F_{ROH} estimated at 5X coverage was significantly higher than F_{ROH} estimates at all higher levels of coverage (95% CIs did not overlap, Fig. 2 D-F). At 5X coverage, F_{ROH} estimates produced by the two BCFtools approaches significantly differed from one another, with neither approach's estimates differing from PLINK's. For all higher levels of coverage, F_{ROH} estimates produced by BCFtools Genotypes and PLINK did not differ but estimates from both methods differed from those produced by BCFtools Likelihoods.

When comparing how the three methods estimated length-specific F_{ROH} values, patterns varied across ROH length categories. For short ROHs, PLINK produced the highest F_{ROH} estimates, followed by BCFtools Likelihoods and then by BCFtools Genotypes, with differences among the three methods significant (*i.e.*, non-overlapping 95% CIs) at 5X-30X coverage and differences between BCFtools Genotypes and the other two methods significant at 50X (Fig. 7). For longer ROHs, BCFtools Genotypes generally had higher F_{ROH} estimates than the other two approaches, and these differences were significant at all coverage levels for long and very long ROHs. Across all methods and ROH length bins, F_{ROH} estimated at 5X coverage were all significantly different from estimates at all other coverage levels within each method and ROH length bin combination.

Discussion

In this manuscript, we highlight the quantitative differences in ROH detection between multiple programs and effects on downstream interpretations associated with these differences. However, these are dependent on our ability to choose appropriate program parameter values, which is particularly complicated when there are a large number of possible parameter value combinations. Although some studies describe testing multiple sets of PLINK parameter values (*e.g.*, Saremi et al., 2019; Grossen et al., 2020; von Seth et al., 2021; Mueller et al., 2022), many do not and there is no widely used, previously published approach to systematically compare results produced by different parameter value combinations.

In Box 1, we demonstrate the exploratory utility of the sensitivity analysis process we followed to select parameter values for our data (see the Supplementary Material for corresponding information for the empirical data). This process is important because disparate sequencing data characteristics are likely to require different parameter values, meaning that it may not be appropriate to use the values we used herein when analyzing other data. For example, studies that use fewer SNPs (*e.g.*, populations that are less genetically diverse, studies with reduced sequencing efforts) should test the effects of altering the minimum SNP density required on ROH inference results. Interactions between specific parameters should also be visualized, such as between the number of heterozygous calls allowed in a window and window size in SNPs, particularly if a reference genome is not assembled to chromosome-level or if mapping rates are somewhat heterogenous across the genome. Sensitivity analysis provides a quick and convenient way to visualize how different parameter values affect F_{ROH} estimates for an entire data set and the degree of variation in those effects across individuals. For samples where inbreeding is anticipated to be highly variable across individuals or for data sets where coverage varies between 5X and 10X, evaluating inter-individual variation in F_{ROH} inference results is particularly important, especially in light of the length-specific ROH inference issues we describe for our results.

Inferred F$_{ROH}$ *value accuracy varies with method and level of coverage*

The patterns of F_{ROH} we estimated tended to vary with program choice and an individual's inbreeding history, potentially leading to uncertainty when incorporating these inbreeding values into management action plans. Between the two BCFtools methods when considering identified ROHs of all lengths, Genotypes produced more accurate overall F_{ROH} estimates than Likelihoods, with F_{ROH} estimates from Likelihoods also increasingly diverging from the true F_{ROH} value with increasing true F_{ROH} (Fig. 1A,B). For populations expected to have considerable variation in F_{ROH} among individuals (*e.g.*, a population that has remained somewhat small for an extended period of time with evidence of recent immigration), applying the BCFtools Likelihoods approach could result in increasingly skewed values for the individuals with the highest levels of inbreeding. For example, using the linear model parameters estimated for 15X coverage, an individual with a true F_{ROH} of 0.10 would be assigned an inferred F_{ROH} of 0.01 (difference = -0.09), whereas an individual with a true F_{ROH} value of 0.40 would be assigned 0.23 (difference = -0.17). This could be particularly problematic when dealing with species or populations of conservation concern because the individuals with the highest true F_{ROH} also have the largest magnitude of error, meaning that concerning signals of inbreeding could go undetected.

In contrast to the underestimations produced by the BCFtools/RoH methods, the sliding window approach implemented in PLINK overestimated F_{ROH}. This was particularly evident at 5X coverage where F_{ROH} estimates differ more from their true values than any other method and coverage level combination in our study (Fig. 1C). However, at coverages above 5X, PLINK produced better estimates than either BCFtools approach (*i.e.*, in our linear models, intercepts for PLINK at 10X-50X are closer to zero than for either BCFtools method and 95% CIs for these parameter estimates do not overlap with any BCFtools intercept 95% CIs). In the context of endangered species conservation, small overestimations of F_{ROH} may be more desirable than underestimations because these are likely to be more conservative (*i.e.*, indicating more close inbreeding than is present in reality) in many situations. Importantly though, as with BCFtools Likelihoods, F_{ROH} estimates diverged from true F_{ROH} at increasing values of true F_{ROH}. However, these values diverged at a much lower rate in the PLINK estimates compared to BCFtools Likelihoods. Again using our simulated data as a model, an individual with a true F_{ROH} value of 0.40 would be estimated to have an F_{ROH} of 0.46 (difference = 0.06) when estimated at 10X-50X with PLINK.

For the two BCFtools methods, patterns of underestimation were consistent with these approaches' high false negative rates and low false positive rates (Fig. 3). Conversely, PLINK produced higher false positive rates and lower false negative rates than either BCFtools method, consistent with overestimation of F_{ROH}. In terms of absolute difference between true and called F_{ROH} values, PLINK outperformed BCFtools at 10X coverage and above, suggesting that PLINK will often provide the most robust estimate of F_{ROH}. However, at lower coverages (5X-10X), BCFtools Genotypes could be considered, given that this method produces F_{ROH} estimates closer to true F_{ROH} than either PLINK or BCFtools Likelihoods. On the other hand, the underestimates produced by this approach are likely related to the high false negative rates we observed (especially relative to PLINK), and the appearance of convergence on true F_{ROH} may be due to length-specific ROH calling rates by this program (see below) and therefore highly variable across populations. It is important to note that while the trends we describe may be consistent with some empirical results (*e.g.*, Robinson et al., 2019), individual variation in genomic characteristics exerts strong influence over F_{ROH} inference results as suggested by comparisons between our simulated and empirical results.

Coverage ≤ 10X strongly influences called ROH lengths

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

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

In our simulated and empirical data, we observed patterns indicating that underlying ROH length distributions influence the patterns of F_{ROH} under- and overestimation. For example, even though PLINK produced higher F_{ROH} estimates than both BCFtools methods for the simulated data and PLINK and BCFtools Genotypes produced statistically indistinguishable estimates for the empirical data (Fig. 2), length-specific F_{ROH} estimates suggest that differences in underlying true

ROH length distributions between the simulated and empirical data may be responsible for the differences in relative F_{ROH} results we observed. For the simulated data, BCFtools Genotypes increasingly overestimated F_{ROH} as ROH length increased (Fig. 5, Fig. S2), with increasing numbers of true ROHs erroneously combined into single called ROHs (Fig. 6). Although we cannot know the true ROH length distributions for the empirical data, long ROHs were called at higher frequencies in the empirical data (at 15X: 382, 397, and 1,281 total ROHs ≥ 1 Mb in length called by PLINK, BCFtools Likelihoods, and BCFtools Genotypes, respectively, in 15 individuals) relative to the simulated data (31 total true ROHs ≥ 1 Mb in length in 25 individuals). The tendency of BCFtools Genotypes to overestimate F_{ROH} for long ROHs combined with the presence of more called long ROHs in our empirical data set may have minimized differences in overall F_{ROH} estimates between BCFtools Genotypes and PLINK in the empirical results relative to the simulated results (Fig. 2). Increased frequencies of long ROHs in the empirical data may have also led to greater differences in F_{ROH} between 5X and 10X across all three methods for the empirical results compared to the simulated results (Fig. 2). All three methods call significantly more intermediate to very long ROHs from the empirical data at 5X than at 10X (Fig. 7B-D), and this may be related to the increased false positive rates we noted at 5X in the simulated data. These results again illustrate the effects of a population's or individual's actual ROH complement, which is determined by typically unknown demographic and breeding patterns, on the relative reliability and utility of ROH identification programs.

Particularly for endangered species with potentially complicated demographic histories, interpreting ROH patterns in a population may be most accurate when multiple tools are used to create an integrated picture. For example, comparing overall and length-specific F_{ROH} estimates between BCFTools/RoH and PLINK can be used to understand the underlying length distributions; an abundance of shorter ROHs would be indicated by higher overall F_{ROH} estimates in PLINK compared to BCFtools Genotypes but similar length-specific F_{ROH} patterns, whereas a ROH complement comprising many longer ROHs would be indicated by similar overall F_{ROH} estimates between PLINK and BCFtools Genotypes but higher intermediate to very long F_{ROH} estimates from BCFtools Genotypes related to PLINK. These accurate assessments of past and ongoing inbreeding could then be used to inform management options, such as translocations to ameliorate close inbreeding.

Conclusions

Inferring the presence and characteristics of ROHs can shed important light on population demographic histories, detect inbreeding depression when combined with fitness information, and even disentangle the mechanisms underlying or loci contributing to inbreeding depression. However, given the variation in ROH-calling accuracy (overall and length-specific) demonstrated here, we caution against direct comparisons of F_{ROH} values generated from different data types or sources or using different inference parameters. Data from disparate studies could be combined and re-analyzed in a standardized fashion, although special attention should be paid to variation in reference genome assembly quality for interspecific comparisons (Brüniche-Olsen, Kellner, Anderson, & DeWoody, 2018). Regardless of the number of data sets to be analyzed, we strongly recommend that studies relying on ROH inference (i) employ at least two ROH-calling programs and interpret their results with each method's biases in mind and/or (ii) compare multiple parameter value combinations via sensitivity analysis, taking care to vary parameters of particular relevance to a data set.

Acknowledgments

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.

References

Andrews, S. (2015). *FastQC: A quality control tool for high throughput sequence data*. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Brüniche-Olsen, A., Kellner, K. F., Anderson, C. J., & DeWoody, J. A. (2018). Runs of homozygosity have utility in mammalian conservation and evolutionary studies. *Conservation Genetics*, *19*(6), 1295–1307. doi: 10.1007/s10592-018-1099-y

Ceballos, F. C., Hazelhurst, S., & Ramsay, M. (2018). Assessing runs of homozygosity: A comparison of SNP array and whole genome sequence low coverage data. *BMC Genomics*, *19*(1), 106. doi: 10.1186/s12864-018-4489-0

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, *19*(4), 220–234. doi: 10.1038/nrg.2017.109

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. doi: 10.1186/s13742-015-0047-8

Crnokrak, P., & Roff, D. A. (1999). Inbreeding depression in the wild. *Heredity*, *83*(3), 260–270. doi: 10.1038/sj.hdy.6885530

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi: 10.1093/bioinformatics/btr330

Diffenbaugh, N. S., & Field, C. B. (2013). Changes in ecologically critical terrestrial climate conditions. *Science*, *341*(6145), 486–492. doi: 10.1126/science.1237123

Duntsch, L., Whibley, A., Brekke, P., Ewen, J. G., & Santure, A. W. (2021). Genomic data of different resolutions reveal consistent inbreeding estimates but contrasting homozygosity landscapes for the threatened Aotearoa New Zealand hihi. *Molecular Ecology*, mec.16068. doi: 10.1111/mec.16068

Gibbs, H. L., & Grant, P. R. (1989). Inbreeding in Darwin's medium ground finches (*Geospiza fortis*). *Evolution*, *43*, 1273–1284.

Grossen, C., Guillaume, F., Keller, L. F., & Croll, D. (2020). Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nature Communications*, *11*(1), 1001. doi: 10.1038/s41467-020-14803-1

Haddad, N. M., Brudvig, L. A., Clobert, J., Davies, K. F., Gonzalez, A., Holt, R. D., … Townshend, J. R. (2015). Habitat fragmentation and its lasting impact on Earth's ecosystems. *Science Advances*, *1*(2), e1500052. doi: 10.1126/sciadv.1500052

Haller, B. C., & Messer, P. W. (2019a). Evolutionary modeling in SLiM 3 for beginners. *Molecular Biology and Evolution*, *36*(5), 1101–1109. doi: 10.1093/molbev/msy237

Haller, B. C., & Messer, P. W. (2019b). SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637. doi: 10.1093/molbev/msy228

Harrisson, K. A., Magrath, M. J. L., Yen, J. D. L., Pavlova, A., Murray, N., Quin, B., … Sunnucks, P. (2019). Lifetime fitness costs of inbreeding and being inbred in a critically endangered bird. *Current Biology*, *29*(16), 2711-2717.e4. doi: 10.1016/j.cub.2019.06.064

Hasselgren, M., Dussex, N., Seth, J., Angerbjörn, A., Olsen, R., Dalén, L., & Norén, K. (2021). Genomic and fitness consequences of inbreeding in an endangered carnivore. *Molecular Ecology*, *30*(12), 2790–2799. doi: 10.1111/mec.15943

Hedrick, P. W., & Garcia-Dorado, A. (2016). Understanding inbreeding depression, purging, and genetic rescue. *Trends in Ecology & Evolution*, *31*, 940–952.

Howrigan, D. P., Simonson, M. A., & Keller, M. C. (2011). Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics*, *12*(1), 460. doi: 10.1186/1471-2164-12-460

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594. doi: 10.1093/bioinformatics/btr708

Huisman, J., Kruuk, L. E. B., Ellis, P. A., Clutton-Brock, T., & Pemberton, J. M. (2016). Inbreeding depression across the lifespan in a wild mammal population. *Proceedings of the National Academy of Sciences*, *113*(13), 3585–3590. doi: 10.1073/pnas.1518046113

Iooss, B., Da Veiga, S., Janon, A., & Pujol, G. (2021). *sensitivity: Global Sensitivity Analysis of Model Outputs* [R]. Retrieved from https://CRAN.R-project.org/package=sensitivity

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, *9*(10), 1205–1218. doi: 10.1111/eva.12414

Keller, L., & Waller, D. (2002). Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, *17*(5), 230–241. doi: 10.1016/S0169-5347(02)02489-8

Krueger, F. (2019). *Trim Galore*. Cambridge, U.K.: Babraham Institute.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. Retrieved from http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Liberg, O., Andrén, H., Pedersen, H.-C., Sand, H., Sejberg, D., Wabakken, P., … Bensch, S. (2005). Severe inbreeding depression in a wild wolf (*Canis lupus*) population. *Biology Letters*, *1*(1), 17–20. doi: 10.1098/rsbl.2004.0266

Liu, S., Westbury, M. V., Dussex, N., Mitchell, K. J., Sinding, M.-H. S., Heintzman, P. D., … Gilbert, M. T. P. (2021). Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. *Cell*, S0092867421008916. doi: 10.1016/j.cell.2021.07.032

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. doi: 10.1101/gr.107524.110

Meyermans, R., Gorssen, W., Buys, N., & Janssens, S. (2020). How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. *BMC Genomics*, *21*(1), 94. doi: 10.1186/s12864-020-6463-x

Mueller, S. A., Prost, S., Anders, O., Breitenmoser-Würsten, C., Kleven, O., Klinga, P., … Nowak, C. (2022). Genome-wide diversity loss in reintroduced Eurasian lynx populations urges immediate conservation management. *Biological Conservation*, *266*, 109442. doi: 10.1016/j.biocon.2021.109442

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, *32*(11), 1749–1751. doi: 10.1093/bioinformatics/btw044

O'Grady, J. J., Brook, B. W., Reed, D. H., Ballou, J. D., Tonkyn, D. W., & Frankham, R. (2006). Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. *Biological Conservation*, *133*(1), 42–51. doi: 10.1016/j.biocon.2006.05.016

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*, 292–294.

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., … Dalén, L. (2015). Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology*, *25*(10), 1395–1400. doi: 10.1016/j.cub.2015.04.007

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. doi: 10.1086/519795

R Core Team. (2020). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Robinson, J. A., Räikkönen, J., Vucetich, L. M., Vucetich, J. A., Peterson, R. O., Lohmueller, K. E., & Wayne, R. K. (2019). Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances*, *5*(5), eaau0757. doi: 10.1126/sciadv.aau0757

Saccheri, I., Kuussaari, M., Kankare, M., Vikman, P., Fortelius, W., & Hanski, I. (1998). Inbreeding and extinction in a butterfly metapopulation. *Nature*, *392*(6675), 491–494. doi: 10.1038/33136

Sánchez-Barreiro, F., Gopalakrishnan, S., Ramos-Madrigal, J., Westbury, M. V., Manuel, M., Margaryan, A., … Gilbert, M. T. P. (2021). Historical population declines prompted significant genomic erosion in the northern and southern white rhinoceros (*Ceratotherium simum*). *Molecular Ecology*, *30*(23), 6355–6369. doi: 10.1111/mec.16043

Saremi, N. F., Supple, M. A., Byrne, A., Cahill, J. A., Coutinho, L. L., Dalén, L., … Shapiro, B. (2019). Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. *Nature Communications*, *10*(1), 4769. doi: 10.1038/s41467-019-12741-1

Slate, J., Kruuk, L. E. B., Marshall, T. C., Pemberton, J. M., & Clutton-Brock, T. H. (2000). Inbreeding depression influences lifetime breeding success in a wild population of red deer (*Cervus elaphus*). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1453), 1657–1662. doi: 10.1098/rspb.2000.1192

Stoffel, M. A., Johnston, S. E., Pilkington, J. G., & Pemberton, J. M. (2021). Genetic architecture and lifetime dynamics of inbreeding depression in a wild mammal. *Nature Communications*, *12*(1), 2972. doi: 10.1038/s41467-021-23222-9

Stoffel, Martin A., Johnston, S. E., Pilkington, J. G., & Pemberton, J. M. (2021). Mutation load decreases with haplotype age in wild Soay sheep. *Evolution Letters*, *5*(3), 187–195. doi: 10.1002/evl3.229

Szpiech, Z. A., Xu, J., Pemberton, T. J., Peng, W., Zöllner, S., Rosenberg, N. A., & Li, J. Z. (2013). Long runs of homozygosity are enriched for deleterious variation. *The American Journal of Human Genetics*, *93*(1), 90–102. doi: 10.1016/j.ajhg.2013.05.003

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, *194*(2), 301–326. doi: 10.1534/genetics.112.148825 von Seth, J., Dussex, N., Díez-del-Molino, D., van der Valk, T., Kutschera, V. E., Kierczak, M., … Dalén, L. (2021). Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nature Communications*, *12*(1), 2393. doi: 10.1038/s41467-021-22386-8

Wright, B. R., Farquharson, K. A., McLennan, E. A., Belov, K., Hogg, C. J., & Grueber, C. E. (2020). A demonstration of conservation genomics for threatened species management. *Molecular Ecology Resources*, *20*(6), 1526–1541. doi: 10.1111/1755-0998.13211

Xie, H.-X., Liang, X.-X., Chen, Z.-Q., Li, W.-M., Mi, C.-R., Li, M., … Du, W.-G. (2022). Ancient demographics determine the effectiveness of genetic purging in endangered lizards. *Molecular Biology and Evolution*, *39*(1), msab359. doi: 10.1093/molbev/msab359

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 data analyses. AMH wrote the manuscript with input and final approval from all authors.

Figure Captions

Figure 1. Both BCFtools methods underestimate true F_{ROH} whereas PLINK produces overestimates. A-C) True vs. called F_{ROH} for each method and level of coverage. Each regression line represents linear model results for a single level of coverage with the shaded areas representing 95% confidence intervals. Each point represents data for a single simulated individual; dashed line is 1:1 line. For PLINK, increasing coverage increases F_{ROH} estimation accuracy, whereas accuracy decreases for both BCFtools approaches.

Figure 2. Increasing coverage from 5X to 10X can have significant effects on F_{ROH} estimates. A-C) True and inferred F_{ROH} values for simulated data and D-F) inferred F_{ROH} values for empirical data at varying coverage levels for all three methods. True mean F_{ROH} values for simulated data are indicated by horizontal dashed line. For the simulated data, error bars are bootstrapped 95%

CIs and points represent mean values (n=100); lines for 15 randomly subsampled individuals are displayed for simplicity (all individual data presented in Fig. S1). For the empirical results, points represent mean values (n=15) and error bars correspond to 95% CIs. Across methods and data types, mean F_{ROH} decreases from 5X to 10X, with significant differences detected when simulated data are analyzed with PLINK and for all three methods applied to the empirical data.

Figure 3. PLINK outperforms BCFtools with respect to false negative rates, but underperforms with respect to false positive rates. A) False positive (*i.e.*, calling a ROH that was not present in an individual) and B) false negative (*i.e.*, failing to call a ROH present in an individual) rates for simulated data across coverage levels 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.

Figure 4. True and called ROH positions for a ~6-Mb window in one exemplar individual. Evidence of false negative and false positive calls can be seen across all methods and coverage levels, and the lumping issue (*i.e.*, the erroneous combining of multiple true ROHs into a single called ROH) is apparent for BCFtools Genotypes, BCFtools Likelihoods (at 5X coverage), and PLINK. Full chromosome plots are provided for three individuals in Fig. S2.

Figure 5. BCFtools Likelihoods produces more accurate length-specific F_{ROH} estimates than BCFtools Genotypes or PLINK (but see the Discussion for additional context of this result). A-C) Called F_{ROH} – true F_{ROH} across methods, ROH length bins, and coverage levels. Dashed horizontal line is at $y = 0$ and values above this line indicate overestimation of F_{ROH} whereas

values below this line indicate underestimation. Length bins were defined as: (i) 100 kb ≥ short ROHs < 250 kb; (ii) 250 kb ≤ intermediate ROHs < 500 kb; (iii) 500 kb ≥ long ROHs < 1 Mb; (iv) 1 Mb ≥ very long ROHs. D) Histograms for bin-specific true F_{ROH} values (*i.e.*, total frequencies sum to 100 individuals within each plot). Despite very few very long ROHs present in simulated individuals, PLINK and BCFtools Genotypes consistently overestimate F_{ROH} for this bin. All individual data for called F_{ROH} – true F_{ROH} are presented in Fig. S2.

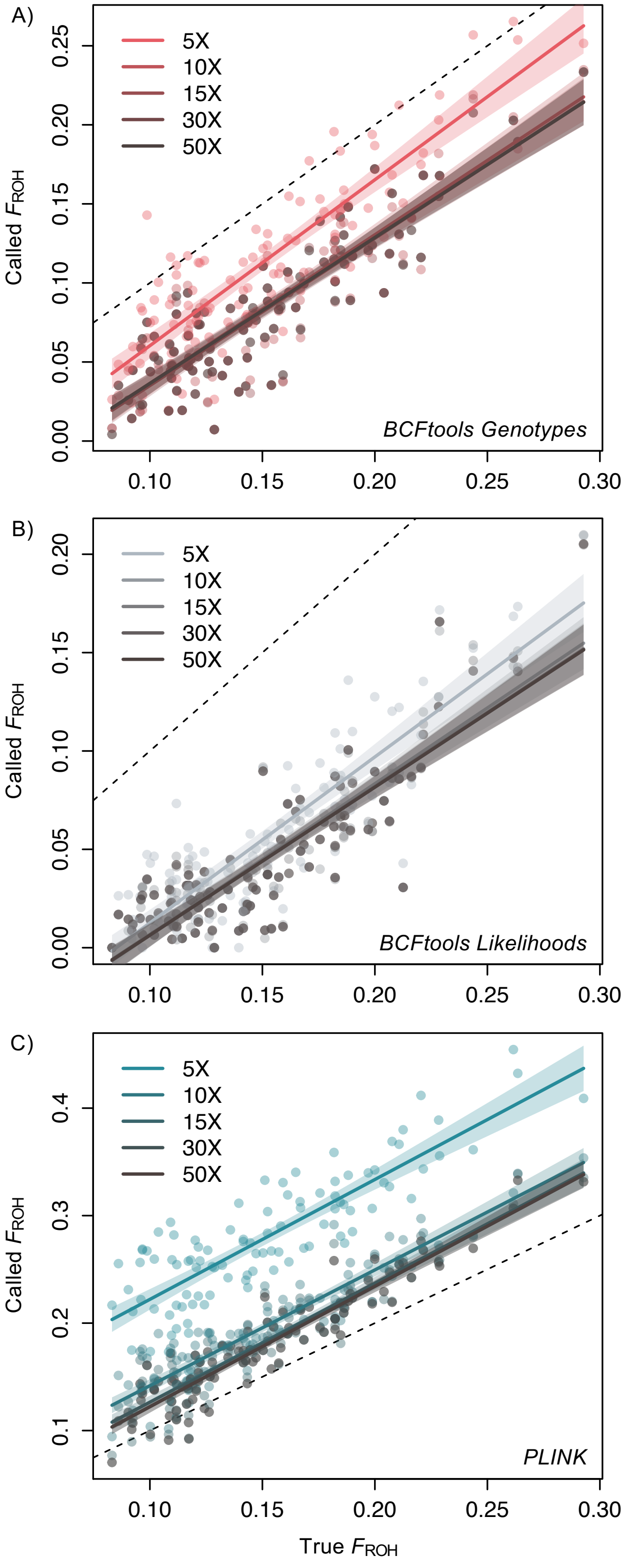
Figure 6. For BCFtools Genotypes and PLINK (and BCFtools Likelihoods at low coverage), multiple true ROHs are increasingly lumped into single called ROHs with increasing true ROH length. A) Illustration of relationships between true ROHs and the called ROHs they are often lumped into. B-E) Number of true ROHs lumped into a single called ROH for each ROH length bin, method, and coverage level. Total number of called ROHs falling into each length bin is provided in the upper right corner of each panel. Degree of circle transparency corresponds to the number of called ROHs matching that particular *y*-value. Transparency levels are normalized to the total number of called ROHs within each panel (all methods and coverage levels combined). Diamonds represent mean values. A simplified version of this figure showing trends in mean values is provided in Fig. S4. Lumping patterns can also be seen in Figs. 4 and S2.

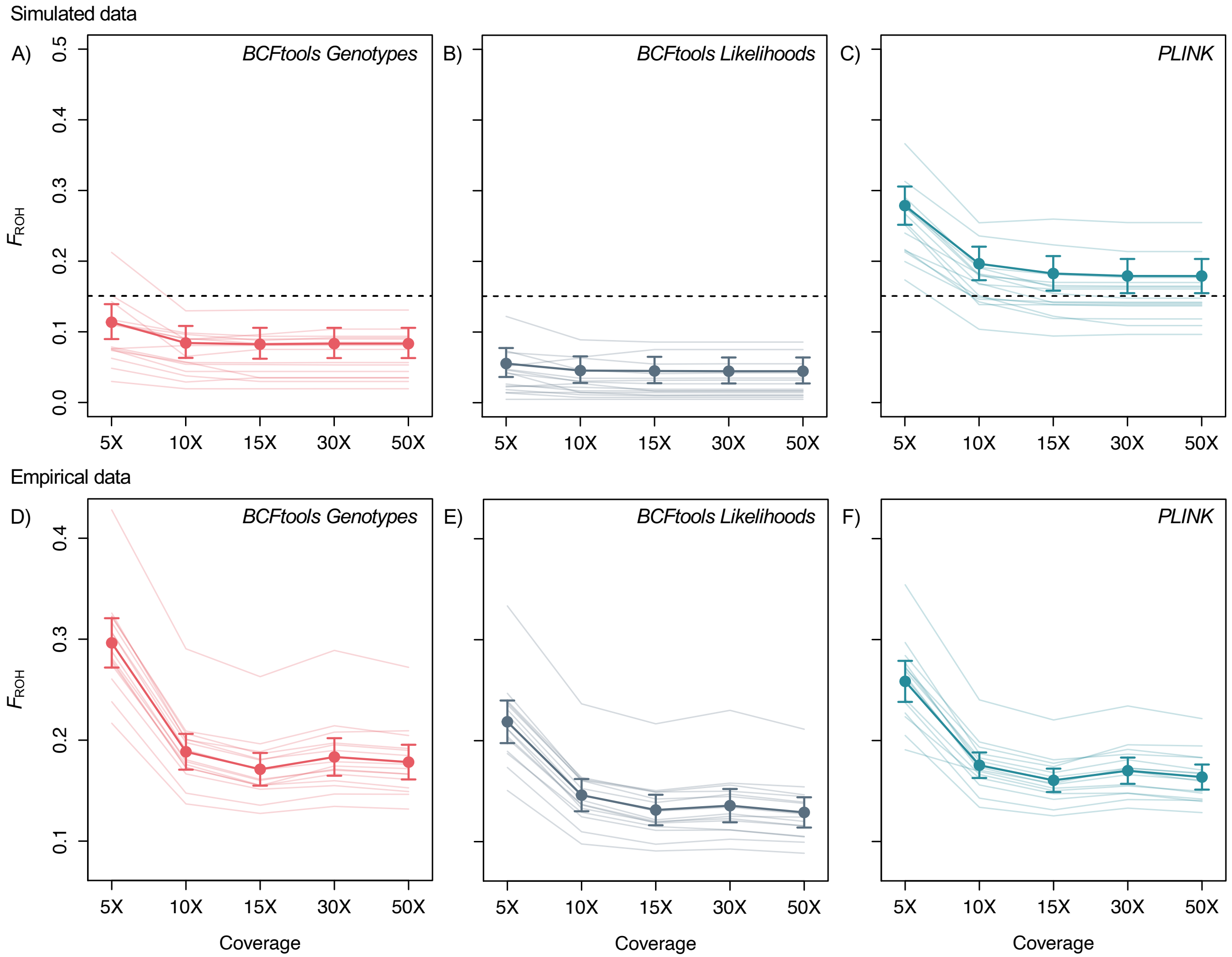
Figure 7. For the empirical data, PLINK tends to call more short ROHs than the BCFtools approaches whereas BCFtools Genotypes tends to call more intermediate to very long ROHs than the other two methods. ROH length-specific F_{ROH} values for A) short, B) intermediate, C) long, and D) very long ROHs. Length bins were defined as: (i) 100 kb ≤ short ROHs < 500 kb; (ii) 500 kb ≤ intermediate ROHs < 1 Mb; (iii) 1 Mb ≥ long ROHs < 2 Mb; (iv) 2 Mb ≥ very long ROHs. Points correspond to mean values and error bars are 95% CIs. Across all methods and

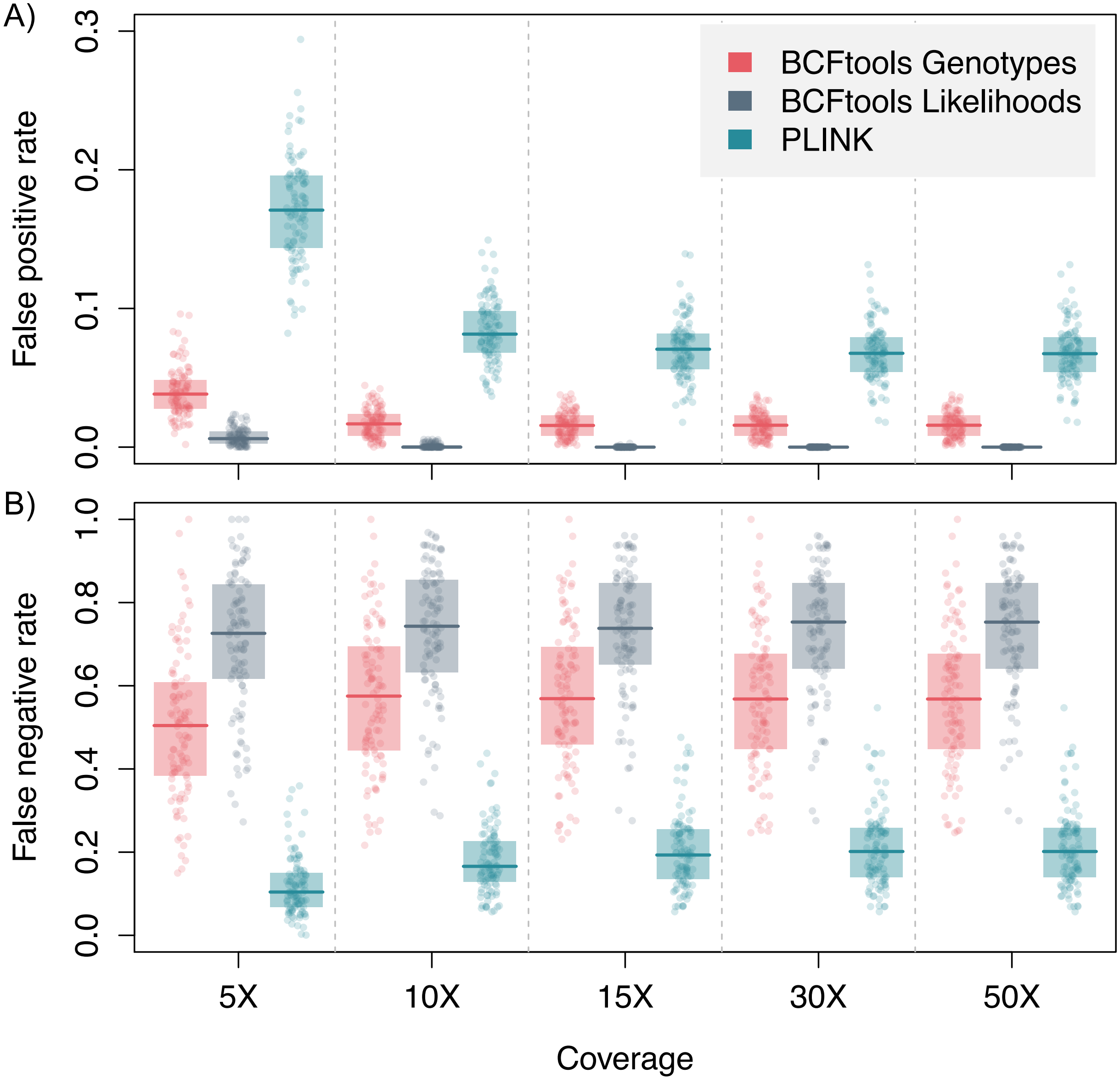
ROH length bins, F_{ROH} estimates at 5X coverage are significantly different from estimates at all other coverage levels within each method and ROH length bin combination.

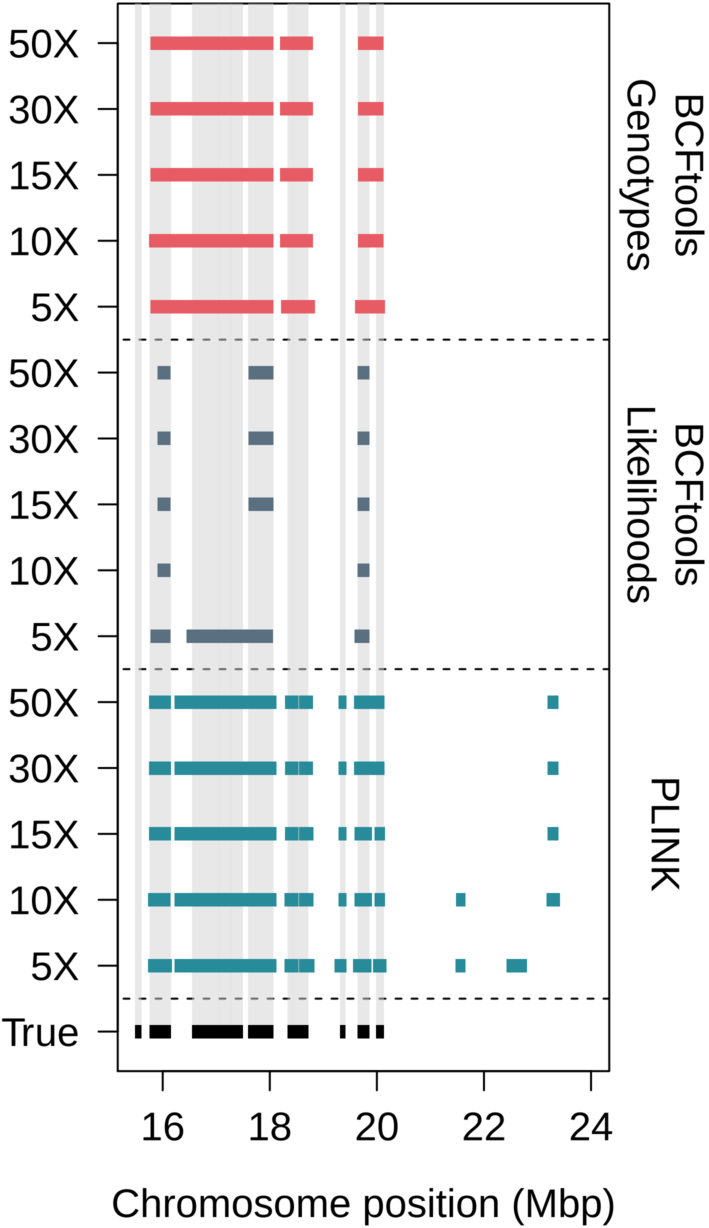
Table 1. Parameter values applied during ROH calling for both simulated and empirical data. For PLINK, a total of 486 combinations were tested. PLINK default values are underlined. ROH = run of homozygosity.

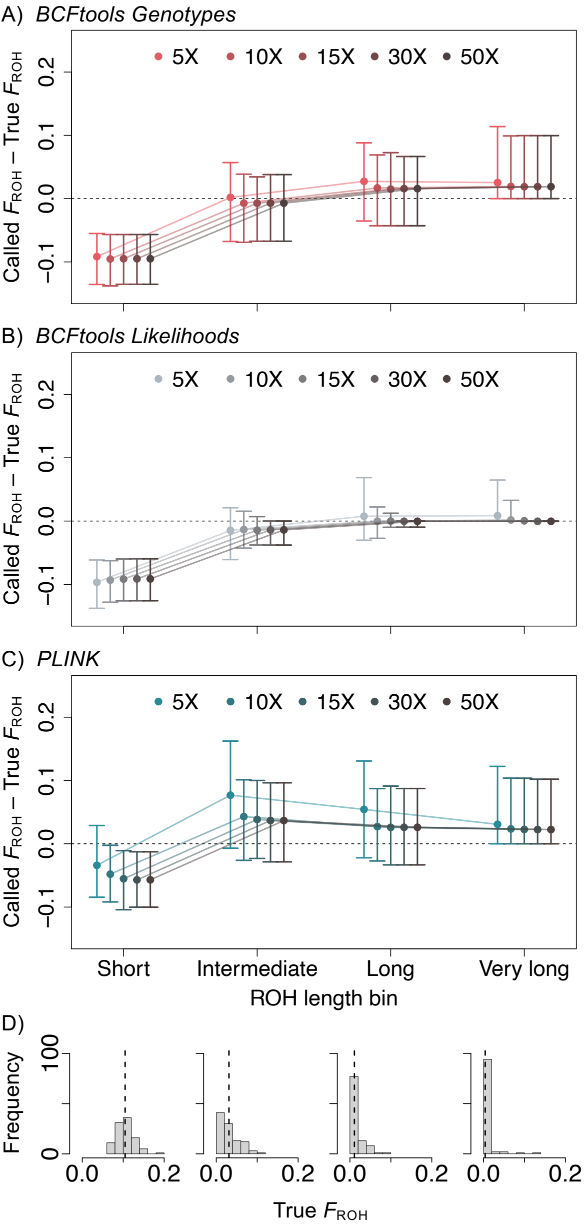
Program	Parameter	Abbreviation	Values	Description
BCFtools/RoH	`--GTs-only`	--	30	If set, uses genotypes only and ignores likelihood values
PLINK	`--homozyg-window-het`	phwh	0, 1, 2	Number of heterozygous sites allowed within a window; default = 1 heterozygous site
	`--homozyg-window-missing`	phwm	2, 5, 50	Number of missing calls allowed in a window; default = 5 missing calls
	`--homozyg-window-snp`	phws	50, 100, 1000	Scanning window length in SNPs; default = 50 SNPs
	`--homozyg-density`	phzd	50	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, 1000	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, 0.05, 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, 100, 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

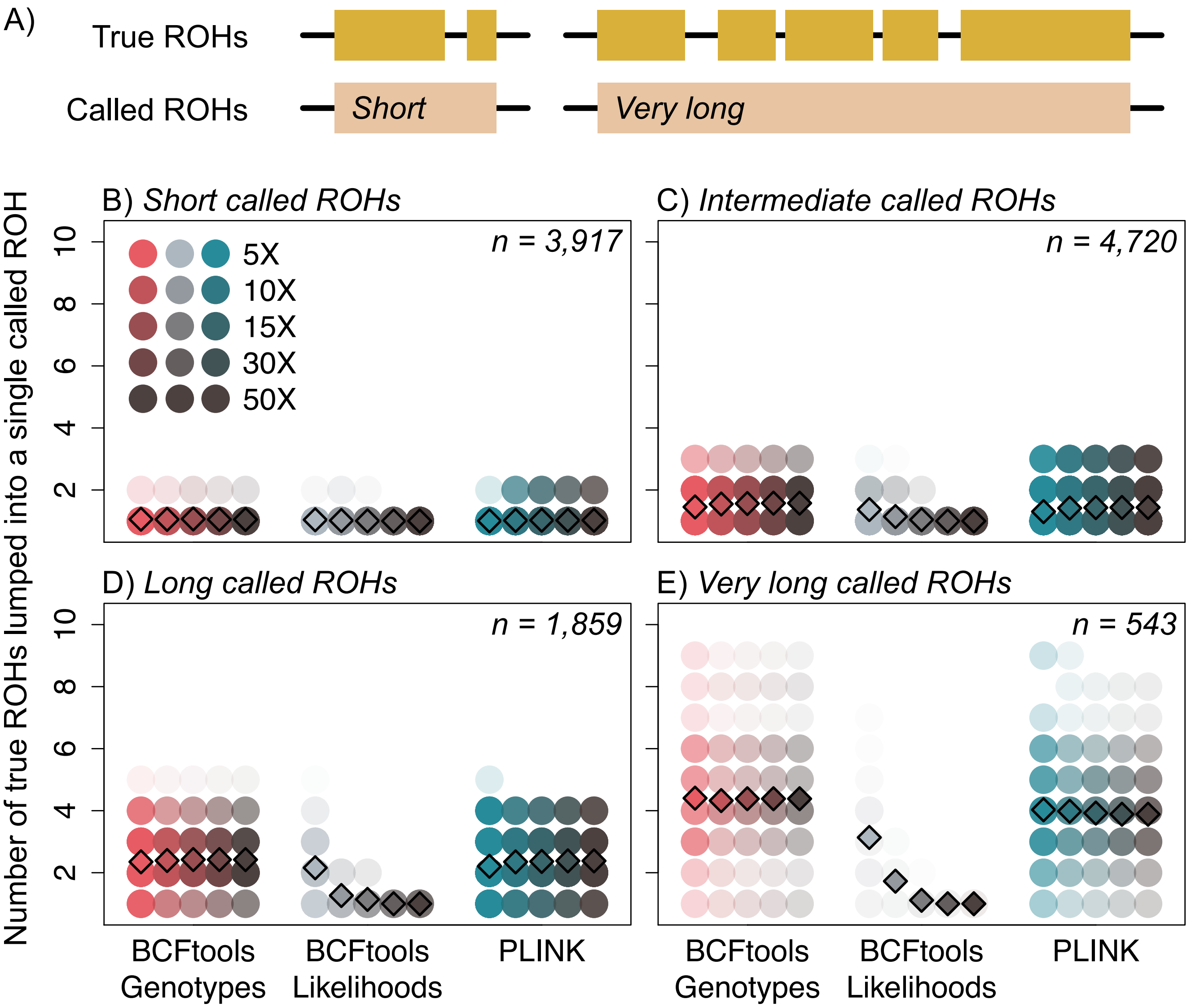

A)
5X
10X
15X
30X
50X
BCFtools Genotypes
Called F_{ROH}
0.00
0.05
0.10
0.15
0.20
0.25
0.10
0.15
0.20
0.25
0.30
B)
5X
10X
15X
30X
50X
BCFtools Likelihoods
Called F_{ROH}
0.00
0.05
0.10
0.15
0.20
0.10
0.15
0.20
0.25
0.30
C)
5X
10X
15X
30X
50X
PLINK
Called F_{ROH}
0.1
0.2
0.3
0.4
0.10
0.15
0.20
0.25
0.30
True F_{ROH}

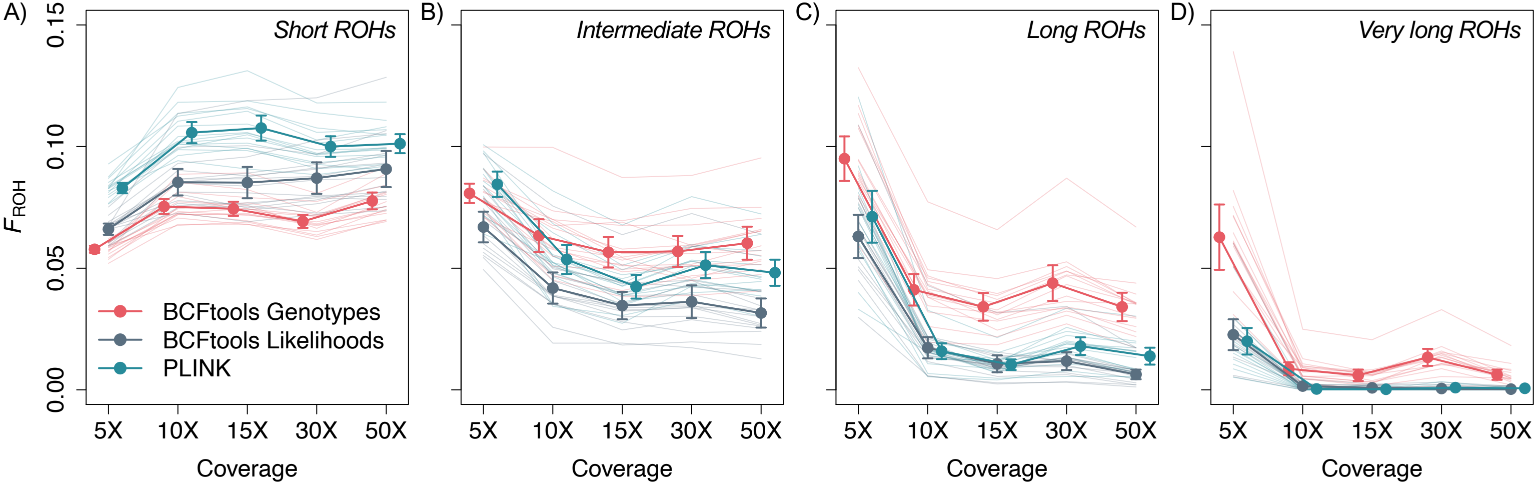

Simulated data
A)
B)
C)
BCFtools Genotypes
BCFtools Likelihoods
PLINK
F_{ROH}
0.0
0.1
0.2
0.3
0.4
0.5
5X
10X
15X
30X
50X
Empirical data
D)
E)
F)
BCFtools Genotypes
BCFtools Likelihoods
PLINK
F_{ROH}
0.1
0.2
0.3
0.4
5X
10X
15X
30X
50X
Coverage
Coverage
Coverage


A)
B)
False positive rate
0.0
0.1
0.2
0.3
False negative rate
0.0
0.2
0.4
0.6
0.8
1.0
BCFtools Genotypes
BCFtools Likelihoods
PLINK
5X
10X
15X
30X
50X
Coverage


50X
30X
15X
10X
5X
50X
30X
15X
10X
5X
50X
30X
15X
10X
5X
True
BCFtools Genotypes
BCFtools Likelihoods
PLINK
16
18
20
22
24
Chromosome position (Mbp)

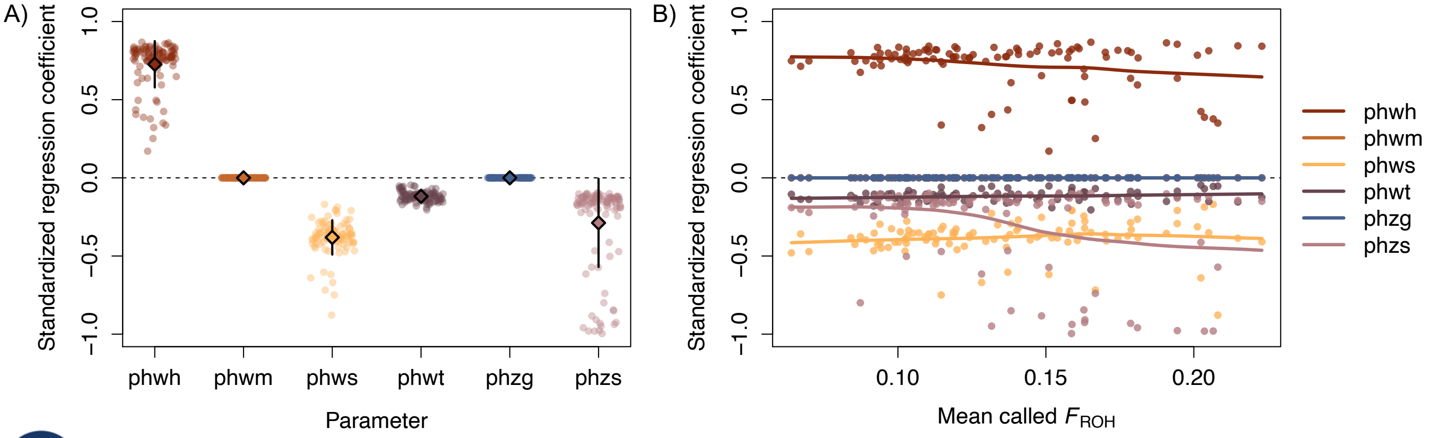

A) BCFtools Genotypes
● 5X ● 10X ● 15X ● 30X ● 50X
Called F_{ROH} − True F_{ROH}
0.2
0.1
0.0
−0.1
B) BCFtools Likelihoods
● 5X ● 10X ● 15X ● 30X ● 50X
Called F_{ROH} − True F_{ROH}
0.2
0.1
0.0
−0.1
C) PLINK
● 5X ● 10X ● 15X ● 30X ● 50X
Called F_{ROH} − True F_{ROH}
0.2
0.1
0.0
−0.1
Short
Intermediate
Long
Very long
ROH length bin
D)
Frequency
100
0
0.0
0.2
0.0
0.2
0.0
0.2
0.0
0.2
True F_{ROH}


A)
True ROHs
Called ROHs
Short
Very long
B) Short called ROHs
C) Intermediate called ROHs
D) Long called ROHs
E) Very long called ROHs
n = 3,917
n = 4,720
n = 1,859
n = 543
5X
10X
15X
30X
50X
Number of true ROHs lumped into a single called ROH
2
4
6
8
10
BCFtools Genotypes
BCFtools Likelihoods
PLINK



A)
B)
C)
D)
Short ROHs
Intermediate ROHs
Long ROHs
Very long ROHs
F_{ROH}
0.00
0.05
0.10
0.15
5X
10X
15X
30X
50X
Coverage
BCFtools Genotypes
BCFtools Likelihoods
PLINK

Box 1. PLINK parameter exploration through sensitivity analysis



1

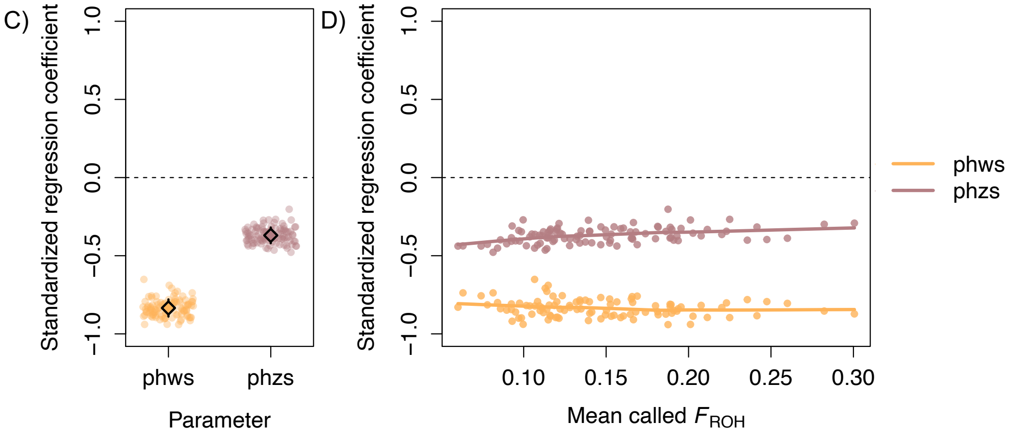
Parameter	Values tested
phwh	0, 1, 2
phwm	2, 5, 50
phws	50, 100, 1000
phzg	500, 1000
phwt	0.01, 0.05, 0.1
phzs	10, 100, 1000

When considering which PLINK parameter values to apply to your data, it is important to determine how individual variation across samples interacts with specific parameter values to influence F_{ROH} estimates. Herein, we demonstrate how we applied sensitivity analysis to select a set of parameter values for our simulated data (downsampled to 15X). In panel A, standardized regression coeffecient (SRC) values are plotted for each parameter, with values SRC values > 0 indicating that increasing the value of a parameter increases mean called F_{ROH} (in the plot, each point corresponds to a single individual). In panel B, SRC values are plotted across mean called F_{ROH} values to show how the relationship between SRC and F_{ROH} changes with F_{ROH}. The parameter values tested are provided in the table at left, with the default PLINK settings underlined. Parameter descriptions are provided in Table 1.

For **phwh** in Iteration 1, SRC values were > 0, indicating a positive effect on F_{ROH} (panel A), with the effect slightly weakening at higher called F_{ROH} values (trendline in panel B). Allowing zero heterozygous calls within a window would discard many windows due to genotyping error, so we conservatively retained the lowest setting > 0 for this parameter, setting it to 1 to avoid inflated F_{ROH} values. Varying **phwm**, **phwt**, and **phzg** had nearly no effect on F_{ROH}, so we retained default values for these parameters. The variable effects of changing **phws** and **phzs** across individuals (i.e., the vertical spread of points in panel A and B) indicate that we should further explore these parameter values, because appropriate values for sliding window length and minimum number of sites per ROH are data-dependent, and thus, value selection will differentially affect ROH estimates due to individual variation in genetic architecture and sequencing errors. We first tested large values (e.g., ≥ 500 for **phws** and ≥ 200 for **phzs**) and found that these settings result in no ROH calls for many individuals (Table S2). We next tested two sets of values near the default value for **phws** and a more narrow range of values near the default value for **phzs** than in Iteration 1.

2

Parameter	Values tested
phws	50, 75, 100, 125
phzs	50, 75, 100, 125

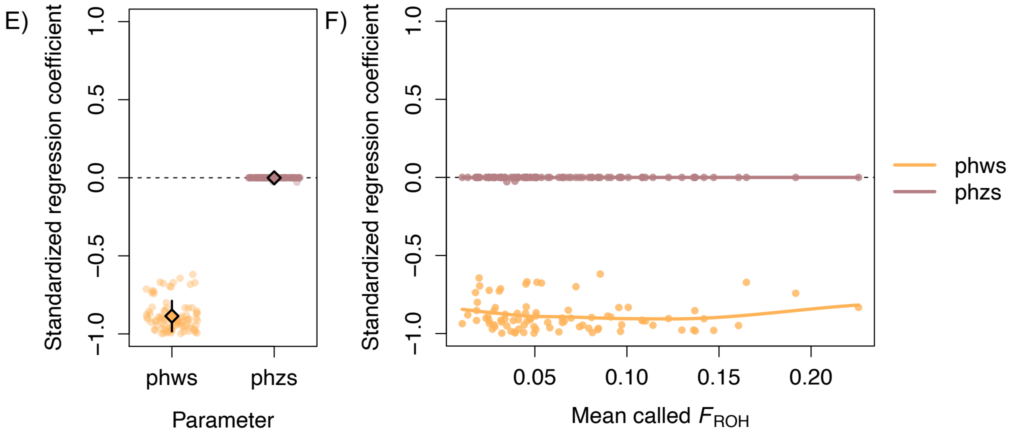


In Iteration 2, we tested values at the smaller end of the ranges tested for **phws** in Iteration 1. For the values tested, increasing **phws** still had a negative effect on F_{ROH}, but the effect was somewhat constant across F_{ROH} values (panels C and D). Increasing **phzs** also corresponds to decreased F_{ROH}, but compared to Iteration 1, the variation in that effect across individuals was much smaller (panels C and D). For both **phws** and **phzs**, these results indicate that selecting a value somewhere within the tested range is unlikely to have substantial sample-specific impacts on called F_{ROH} values.

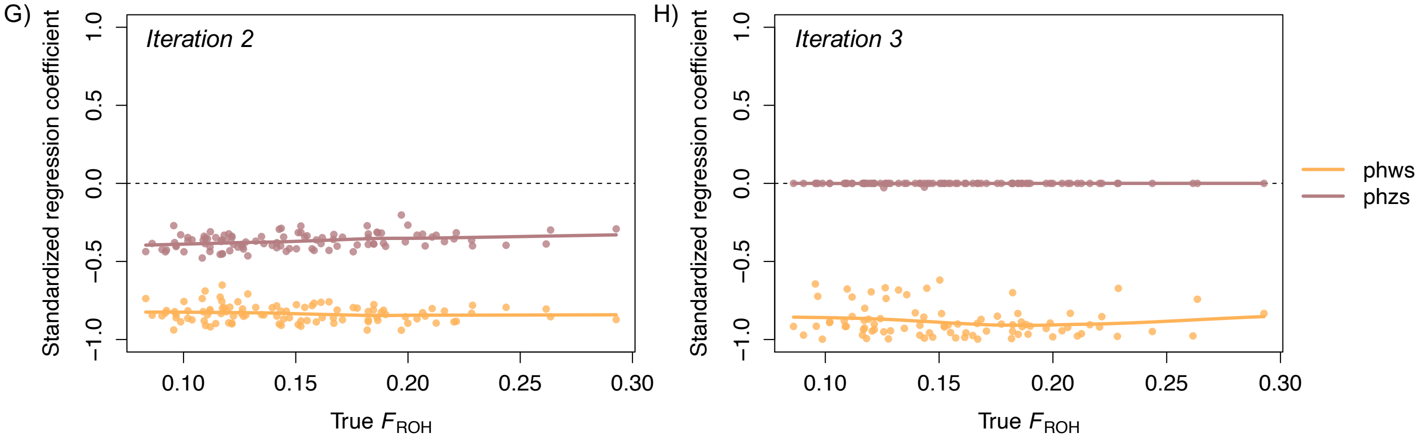
In Iteration 3, we tested slightly higher values for **phws** than in Iteration 2. Although this change minimized the effect that varying the value of **phzs** has on F_{ROH}, variation in the effects of **phws** settings across individuals increased substantially when compared with Iteration 2. In this iteration, mean called F_{ROH} values have also shifted towards zero, indicating that increasing **phws** to the tested values may be leading to some ROHs not being called in some individuals. Based on this comparison with the Iteration 2 results, we opted to retain the default values for both **phws** and **phzs**, which were included in the Iteration 2 tested values.

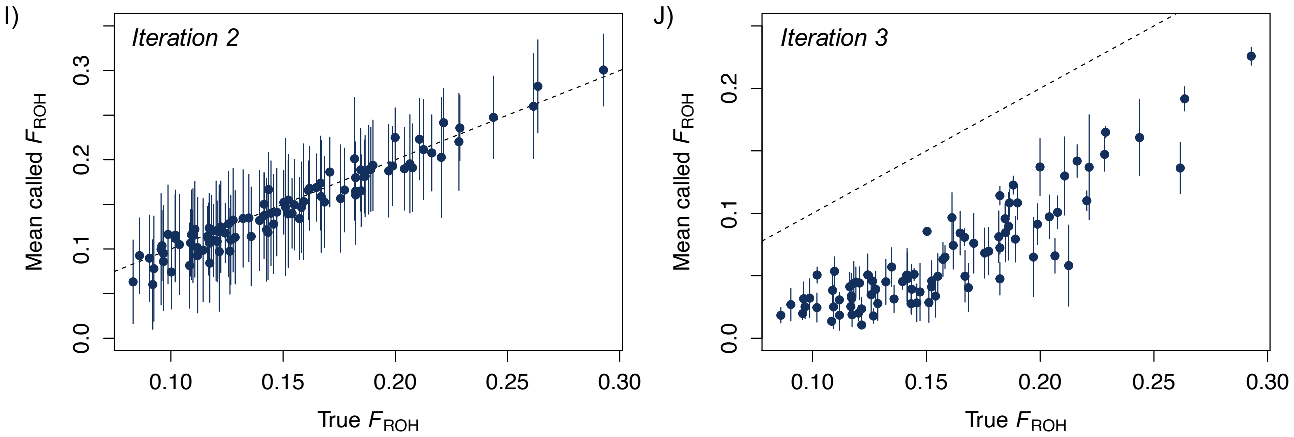
3

Parameter	Values tested
phws	150, 175, 200, 225
phzs	50, 75, 100, 125



Because we tested this approach on simulated data, we can also examine the relationship between true F_{ROH} values and SRCs. For Iteration 2 (the set of values we retained for all downstream analyses), although increasing the values of both **phws** and **phzs** decreased called F_{ROH}, there is no consistent pattern in variation across individuals as true F_{ROH} varies and the variation is small. For Iteration 3, however, there is substantial individual variation in how varying values of **phws** affect called F_{ROH}. This variation is likely due to variation in individuals' true ROH length distributions, with individuals with a greater proportion of short ROHs more strongly affected by increasing **phws**. This is also reflected in lower mean values for called F_{ROH} relative to true F_{ROH} (panel J; vertical lines are ± 1 SD).




I)
Iteration 2
Mean called F_{ROH}
0.0
0.1
0.2
0.3
0.10
0.15
0.20
0.25
0.30
True F_{ROH}
J)
Iteration 3
Mean called F_{ROH}
0.0
0.1
0.2
0.10
0.15
0.20
0.25
0.30
True F_{ROH}