Received: 24 May 2020 | Revised: 21 December 2020 | Accepted: 27 January 2021

DOI: 10.1111/1755-0998.13343

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data

Sara Guirao-Rico | Josefa González

Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona, Spain

Correspondence
Sara Guirao-Rico, Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona, Spain.
Email: sara.guirao@ibe.upf-csic.es

Funding information
H2020 European Research Council, Grant/Award Number: 2014-CoG-647900

Abstract

Population genomics is a fast-developing discipline with promising applications in a growing number of life sciences fields. Advances in sequencing technologies and bioinformatics tools allow population genomics to exploit genome-wide information to identify the molecular variants underlying traits of interest and the evolutionary forces that modulate these variants through space and time. However, the cost of genomic analyses of multiple populations is still too high to address them through individual genome sequencing. Pooling individuals for sequencing can be a more effective strategy in Single Nucleotide Polymorphism (SNP) detection and allele frequency estimation because of a higher total coverage. However, compared to individual sequencing, SNP calling from pools has the additional difficulty of distinguishing rare variants from sequencing errors, which is often avoided by establishing a minimum threshold allele frequency for the analysis. Finding an optimal balance between minimizing information loss and reducing sequencing costs is essential to ensure the success of population genomics studies. Here, we have benchmarked the performance of SNP callers for Pool-seq data, based on different approaches, under different conditions, and using computer simulations and real data. We found that SNP callers performance varied for allele frequencies up to 0.35. We also found that SNP callers based on Bayesian (SNAPE-pooled) or maximum likelihood (MAPGD) approaches outperform the two heuristic callers tested (VarScan and PoolSNP), in terms of the balance between sensitivity and FDR both in simulated and sequencing data. Our results will help inform the selection of the most appropriate SNP caller not only for large-scale population studies but also in cases where the Pool-seq strategy is the only option, such as in metagenomic or polyploid studies.

KEYWORDS
Bayesian, high throughput sequencing, low-frequency variants, maximum likelihood, population genomics, site frequency spectrum

1 | INTRODUCTION

Population genomic studies compare whole-genome sequences from several individuals to address evolutionary long-standing questions such as the relative contribution of the different evolutionary forces (i.e., natural selection, genetic drift, mutation and gene flow) that drive the changes in the frequencies of alleles and phenotypes in populations over space and time. However, whole-genome resequencing of many individuals to coverages that allow us to control the sequencing error is still prohibited especially for species with medium-size to large genomes. In the last years, Pool-seq data is increasingly being used for population genomic studies since among other benefits, it bypasses the most expensive step in the sequencing process: library preparation for each individual in the sample (Healy & Burton, 2020; Jónás et al., 2016; Micheletti & Narum, 2018; Neethiraj et al., 2017; Schlötterer et al., 2014; Tilk et al., 2019). Consequently, at equal sequencing effort and cost, Pool-seq allows us to sequence larger samples of individuals and hence, to decrease the variance in the estimates of population allele frequencies (Ferretti et al., 2013; Futschik & Schlötterer, 2010). One of the main drawbacks of Pool-seq data is that sequencing errors are difficult to discern from low-frequency variants, especially when pool sizes and sequencing depths are low (Anderson et al., 2014; Cutler & Jensen, 2010; Futschik & Schlötterer, 2010). Some studies have proposed to remove the low-frequency mutations to avoid having a high rate of false positive variants (Anand et al., 2016). However, the choice of the frequency cutoff value is not trivial and besides, removing low-frequency mutations may results in a biased site frequency spectrum (SFS) since low-frequency variants are underrepresented.

The SFS is one of the most relevant summary statistics describing sequence variation. Obtaining a robust estimate of the SFS is crucial in population genomics because from its shape, both past demographic processes and selection events can be inferred, among others (Bustamante et al., 2001; Excoffier et al., 2013; Gutenkunst et al., 2009; Koropoulis et al., 2020; Ronen et al., 2013). In addition, low-frequency variants are abundant in populations and have been widely used not only to infer demography (Gravel et al., 2011; Keinan & Clark, 2012; Linck & Battey, 2019; Lohmueller, 2014) or to detect the hallmarks of selection (Kim & Stephan, 2002; Martin et al., 2016; Peischl et al., 2018; Tennessen et al., 2010; Vy & Kim, 2015) but also to pinpoint which are the loci associated to traits of interest in QTL and GWAS analysis (Bloom et al., 2019; Bomba et al., 2017; Fischer et al., 2013; Fournier et al., 2019; Wojcik et al., 2019). For instance, GWAS analysis that have tried to decipher which are the genetic basis of human disease have shown that common variants associated with complex traits only explain a small portion of heritability. In some studies, it has been found that ~70% of all nonsynonymous singletons are sufficiently deleterious that they will never reach frequencies >5% suggesting that low-frequency variants are enriched for damaging variants (Nelson et al., 2012). To date, none of the available variant callers for Pool-seq data can completely solve the problem of the correct identification of low-frequency variants (Bansal, 2010; Druley et al., 2009; Koboldt et al., 2009; Lynch et al., 2014; Raineri et al., 2012; Wei et al., 2011; Wilm et al., 2012). However, Single Nucleotide Polymorphism (SNP) callers based on different approaches might differ in their performance in low-frequency variants.

In this work, we have carried out a benchmarking study comparing the performance of four SNP callers for Pool-seq data based on different approaches, Bayesian (SNAPE-pooled), Frequentist (MAPGD), and Heuristic (PoolSNP), and combination of Frequentist and Heuristic approaches (VarScan; Table 1). These callers have been used in previous studies in a wide range of species, such as nematodes, fruit flies, ladybirds, pigs, humans and plants (Adams et al., 2019; Bansal, 2010; Esteve-Codina et al., 2013; Frachon et al., 2018; Gautier et al., 2018; Kapun et al., 2020). We used both computer simulations (under the standard neutral model and under a complex demographic history) and *Drosophila melanogaster* sequencing data (the same samples sequenced separately for each individual and as a pool) to compare their performance using different caller conditions,

TABLE 1 Pool-seq SNP callers and conditions benchmarked in this study

Caller approach	Caller	Condition	Parameters	References
Bayesian	SNAPE-pooled	Flat prior distribution of SFS	sn1	Raineri et al. (2012)
		Informative (SNM) prior distribution of SFS	sn2	
Frequentist: Maximum likelihood	MAPGD	*p*-value for the distribution of the log-likelihood ratio test of polymorphisms	m22: 0.00001 m11: 0.0001 m6: 0.01	Lynch et al. (2014)
Frequentist: Fisher's exact test	VarScan	Cross-sample *p*-value for calling variants	v1: 0.00001 v2: 0.0001 v3: 0.01 v4: 0.05	Koboldt et al. (2009)
Heuristic	PoolSNP	Maximum allowed fraction of samples not fulfilling all parameters	psnp1: 0.1 psnp2: 0.8	Kapun et al. (2020)

Abbreviations: SFS, site frequency spectrum; SNM, standard neutral model.

coverage and sample size. Overall, similar trends in the performance of the four callers were observed in computer simulations and sequencing data. As expected, the main difference in performance between callers was found at low frequencies. However, for some specific sample sizes, coverages and underlying demographic scenarios, our analysis uncovered important differences between methods in allele frequencies up to 0.35, which could be very relevant when planning to design population genomics studies based on pools, metagenomic data, or polyploids. Moreover, we found that SNP callers also differ in terms of the balance between sensitivity and FDR both in simulated and in sequencing data.

2 | MATERIALS AND METHODS

2.1 | Coalescent simulations

We performed coalescent simulations under the standard neutral model (SNM) and under a demographic scenario based on Duchen et al. (2013). The demographic scenario corresponds to the most accepted joint demographic history estimated for *Drosophila melanogaster* in three continents: Africa, North America and Europe (Figure 1). However, we have increased the severity of the bottleneck in Africa to obtain an increasing degree of bias towards low-frequency variants in the three populations (hereafter ΨAfrica, ΨNorth America and ΨEurope) in order to assess the performance of the different callers in increasingly challenging conditions.

For each population scenario, we used the ms program (Hudson, 2002) to simulate 100 replicates of a 100-kb long DNA sequence and setting the per site population-scaled mutation rate and population-scaled recombination parameters to 0.01. For each replicate and population, we run the simulations with different sample sizes: 100, 50 and 20 individuals.

To obtain the ancestral sequence of each simulated population that will also be used as the reference genome for the mapping process, we simulated 100-kb long DNA sequence by randomly sampling nucleotides for each position, using the PIPELINER version 0.2.0 (Nevado & Perez-Enciso, 2015). PIPELINER was additionally used to convert the ms output files to fasta format. Each simulated replica of each model with different number of individuals was used to simulate Pool-seq data. Pool-seq data was generated using the ART _ ILLUMINA version 2.5.8 (Huang et al., 2012) that simulates next generation sequencing (NGS) pair-end reads with the built-in profile for Illumina paired-end technology of 100 bp-long reads and mean size of DNA fragments of 350 bp. For each population and sample size, the average read depth (coverage) per chromosome was set to 1.2×, 0.50×, and 0.20×. Reads were mapped against the simulated reference sequence using BWA 0.7.16a-r1181 (Li & Durbin, 2009), removing reads with mapping quality below 20. Finally, we used PICARD-TOOLS version 2.8.3 (http://broadinstitute.github.io/picard) and SAMTOOLS version

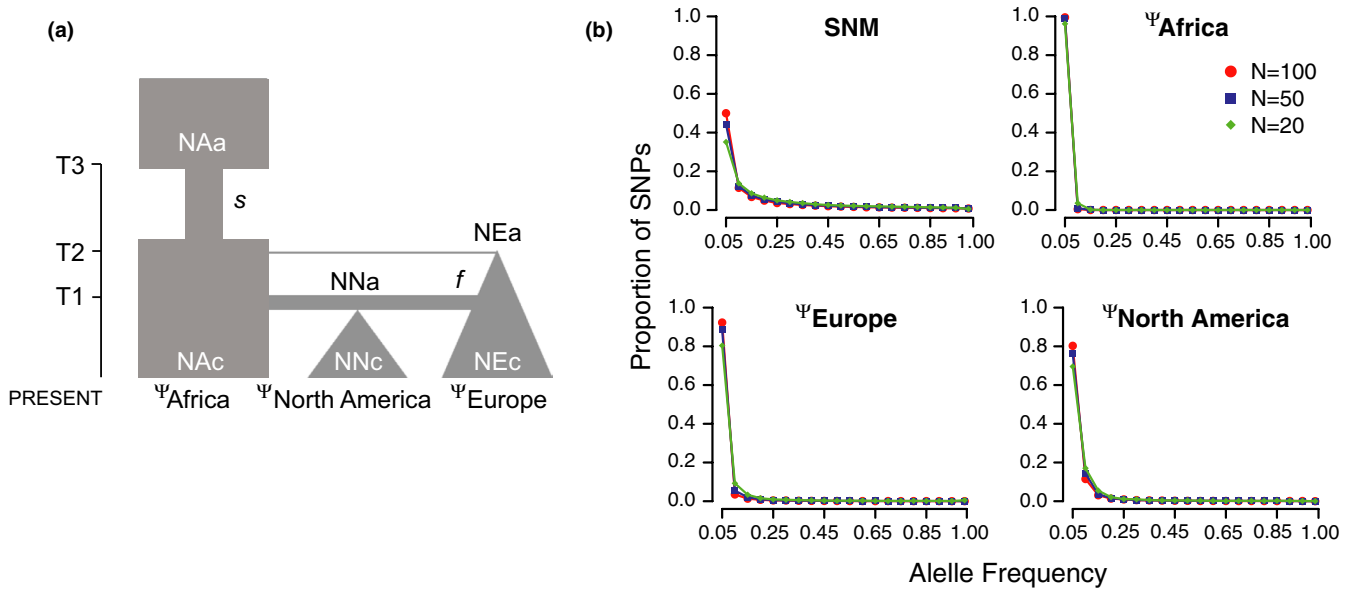

FIGURE 1 Site frequency spectrum of the different simulated samples. (a) Demographic model modified from Duchen et al. (2013) used to perform coalescent simulations. The duration of the bottleneck in African population was modified in other to simulate populations with an increasing degree of bias of the SFS toward low-frequency variants. ΨAfrica, simulated pseudo African population; ΨNorth America, simulated pseudo North American populations, ΨEurope, simulated pseudo European population. NAa, ancient population size of ΨAfrica (4,975,360 individuals); NAc, current population size of ΨAfrica (5,224,100 individuals); NNa, starting population size of ΨNorth America (2,500 individuals); NNc, current population size of ΨNorth America (15,984,500 individuals); NEa, ancient population size of ΨEurope (17,000 individuals); NEc, current population size of ΨEurope (3,122,470 individuals); T1, time of admixture between ΨAfrica and ΨEurope (144 years ago); T2, time of split between ΨAfrica and ΨEurope (19,000 years ago); T3, time of bottleneck in ΨAfrica (237,227 years ago); *s*, Severity of the bottleneck (~3,400); *f*, proportion of admixture (0.85). (b) Proportion of SNPs across the SFS simulated under the SNM and the joint demographic model depicted in (a)

1.6 (Li, 2011) to assign the reads to each read-group and to generate pileup or mpileup outputs from BAM files, respectively.

The program MSTATSPOP (https://github.com/cragenomica/mstatspop) was used to compute the number of SNPs and the SNP frequency for simulated sequences from each of the 12 simulated scenarios (four populations with three different number of individuals each). Note that this software has been previously used in several organisms and has also been included in the PIPELINER tool (Álvarez-Presas et al., 2014; Bianco et al., 2015; Guirao-Rico et al., 2018; Nevado & Perez-Enciso, 2015).

2.2 | *Drosophila melanogaster* sequenced data

We downloaded the FASTQ files corresponding to 30 inbred strains from the *Drosophila Genetic* Reference Panel (DGRP; Mackay et al., 2012) that were individually sequenced at 21× (Table S1). To test how the different callers perform, we need to compare the same data sequenced individually and pooled. Thus, we generated two whole-genome sequencing replicates of a pool of individuals made by using the same 30 DGRP strains described above and with 10 females per strain. The two replicates of the same resequencing library were used to evaluate the performance of VarScan and PoolSNP with the option of joint variant calling for multiple samples. DNA was extracted using the Blood and Cell Culture DNA Mini Kit (Qiagen) according to the manufacturer's protocol and sequenced on a single lane of Illumina NovaSeq with a mean depth around 2.4× (1.2× for noninbred positions) with 151 bp paired-end reads. Library preparation and sequencing was carried out at Macrogen Inc. Raw sequenced data is publicly available under NCBI Bioproject accession PRJNA632498.

The mapping and filtering process were the same for the individual and the Pool-seq data. For each strain and pool, we filtered the raw FASTQ reads to remove low-quality bases (minimum base PHRED quality = 18 as in Kapun et al., 2020) using CUTADAPT version 1.8.3 (Martin, 2011). Trimmed reads were mapped against the *D. melanogaster* reference genome version 6.12 using BWA version 0.7.16a-r1181 (Li & Durbin, 2009) and reads with mapping quality below 20 and secondary alignments were removed using SAMTOOLS version 1.6 (Li et al., 2009). Then, we used PICARD-TOOLS version 2.8.3 (http://broadinstitute.github.io/picard) to remove duplicate reads, and GATK version 3.7-0-gcfedb67 (McKenna et al., 2010) to realigned sequences flanking insertions-deletions (indels). SAMTOOLS version 1.6 (Li et al., 2009) was used to retrieve only autosomal information and to generate pileup or mpileup outputs from BAM files.

2.3 | SNP calling

2.3.1 | Individual sequencing data

We used the program ANGSD version 0.925 (Korneliussen et al., 2014) to estimate the genotype likelihoods for each site of each DGRP strain in order to calculate the inbreeding coefficient of each strain. The program was run twice with the following common settings for each run: -nInd 30 -GL 2 -doGlf 3 -doMajorMinor 4 -doMaf, 1 and with two different *p*-values: 1e−4 and 1e−6. For each site, the major and minor alleles were specified according to the *D. melanogaster v. 6.12* reference allele.

The program NGSF version 1.2.0 (Vieira et al., 2013) was used to estimate the inbreeding coefficient of each DGRP strain under a probabilistic framework using as input the genotype likelihood files resulting from running the ANGSD program. The program was run with the following settings: --n_ind 30 --glf - --min_epsilon 1e−9.

Then, we called biallelic SNPs taking into account the inbreeding coefficients of each strain with the following parameter settings: -nInd 30 -GL 2 -doGeno 2 -doMajorMinor 4 -doMaf 1 -C 50 -baq 1 -doPost 1 -indF -doVcf 1 and with two different *p*-values: 1e−4 and 1e−6. As the number of SNPs called using the two different *p*-values did not vary substantially (2,303,817 and 2,249,247, respectively), we only report the results obtained when using the *p*-value 1e−4.

We filtered out those SNPs that were in interspersed repeats and low-complexity regions using RepeatMasker version 4.0.7 (Smit et al., 2013–2015) and GATK version 3.7-0-gcfedb67 (McKenna et al., 2010).

2.3.2 | Pool-seq data

SNPs were called using four different software based on different approaches: one based on a Bayesian approach, SNAPE-pooled; (Raineri et al., 2012), two frequentists: one based on a maximum likelihood approach, MAPGD (Lynch et al., 2014) and the other combining some heuristic steps with the Fisher's exact test, VarScan (Koboldt et al., 2009); and one relaying in heuristic method, PoolSNP (Kapun et al., 2020). Each caller was run with different parameter settings (hereafter conditions; Table 1).

SNAPE-pooled calls SNPs based on the posterior probability for each site of being polymorphic. This software takes into account sequencing errors, and allows the user to specify two different priors for the expected frequency spectrum (-priortype). SNAPE-pooled was run specifying two different priors for the site frequency spectrum: flat and informative (the site frequency spectrum under the SNM): sn1 and sn2, respectively. The other parameters were: -nchr P -theta 0.01 -D 0.10 -fold unfolded, where P was the number of chromosomes for each condition (p = 30 or p = 60, for full or no inbreeding Pool-seq data sets, respectively). We then filtered out those SNPs with a posterior probability <.90, as suggested by the authors (Raineri et al., 2012).

MAPGD is a series of related programs that among others, estimate allele frequency from population genomic pooled data using a statistically rigorous maximum likelihood approach that takes into account errors associated with sequencing and also a likelihood-ratio test statistic (-a) for evaluating the null hypothesis of monomorphism. MAPGD version 0.4.26 was run using three different thresholds for the log-likelihood ratio (LLR) test of polymorphism

versus no polymorphism (-a flag) 22, 11, and 6, which roughly corresponds to a significant level for the distribution of the log-likelihood ratio test of polymorphisms of ~0.00001, ~0.0001 and ~0.01 (hereafter m22, m11 and m6, respectively).

VarScan combines a heuristic algorithm with a *p*-value computed using a Fisher's exact test on the read counts supporting each type of allele to call SNPs. The reads can be count on single or multiple samples simultaneously based on user-defined parameters such as minimum thresholds for coverage (--min-coverage), variant allele frequency (--min-var-freq), and statistical significance (--*p*-value). The Fisher's exact test is computed comparing the read counts supporting each allele (reference and alternative) with the expected distribution based on sequencing error alone. VARSCAN version 2.4.2 was run using mpileup2snp command and the following user-defined parameters: --min-coverage 8 --min-reads2 2 --min-var-freq 0.50/N; where N corresponds to 100, 50 and 20 for the simulated Pool-seq, and to 30 or 60 (full or no inbreeding) for the DGRP Pool-seq data. We filtered out those SNPs with the cross-sample *p*-values for calling variants >0.00001, >0.0001 and >0.01 and >0.05 (v1, v2, v3 and v4, respectively).

PoolSNP is also based on a heuristic approach, and like VarScan, can call SNPs on single or multiple samples simultaneously. Several user-defined parameters such as minimum and maximum coverage, minimum allele count, allele frequency of a minor allele and maximum percentage of sample that are allowed to not fullfill the above criteria, can be specified. PoolSNP was run specifying two different miss-fraction parameters (i.e., the maximum allowed fraction of samples not fulfilling all parameters): 0.1 and 0.8 (psnp1 and psnp2, respectively). The common user-defined parameters for the two conditions were: minimum coverage = 8, maximum coverage = 0.95 (i.e., the maximum coverage percentile to be computed); minimum count = 2, minimum frequency = 0.50/N, where N corresponds to 100, 50, and 20 for the simulated Pool-seq, and to 30 or 60 (full or no inbreeding) for the DGRP Pool-seq, base quality = 15.

For *D. melanogaster* Pool-seq data, the called SNPs that were in interspersed repeats and low-complexity regions were filtered out as in individual sequencing data (see above).

2.4 | Benchmarking statistics

In order to evaluate the performance of each caller, we computed four different statistics for each caller and condition, population, sample size, and coverage. Because we were interested in comparing only segregating variants, we did not consider fixed variants in any of the callers (frequency = 1). As nearly all the SNPs called by SNAPE-pooled in the range of frequency between 0.95 and 1 have posterior probabilities of being fixed >0.95, we also filtered them out. When comparing the performance of the callers using *D. melanogaster* sequencing data and for those callers that use only one sample (MAPGD and SNAPE-pooled), we only discussed the results for one of the two experimental sequencing replicates.

The four statistics computed were: sensitivity or true positive rate (TPR) and false discovery rate (FDR) were computed as the proportion of true SNPs in the population that were correctly recovered, and as the proportion of SNP calls that were incorrect, respectively.

Lin's concordance correlation coefficient (CCC) was computed using EPIR package version 1.0-4 (Stevenson et al., 2019). This statistic compares the agreement between the SNP frequencies estimated by a specific caller and those of the simulated data/individual sequencing data (true SNP frequencies). This statistic accounts for both precision (correlation between the estimated and true SNP frequencies) and accuracy (bias in the SNP frequency estimation). The guidelines for interpreting CCC suggested by McBride (2005) are: <0.90: poor; 0.90 to 0.95: moderate; 0.95 to 0.99: substantial; and >0.99 almost perfect.

Mean of the percentage of the relative error (%RE): mean of the ratio of the absolute error (difference between the estimated and the true SNP frequencies) to the true SNP frequency expressed as a percentage.

To study the performance of the four callers in recovering the SFS, we divided the estimated and the true SFS in 20 bins and computed the same descriptive statistics for each of the 20 frequency bins. We do not discuss the values of CCC across the SFS since they were well below 0.90 for all bins of frequency probably caused by the low number of SNPs in each frequency category and the high variance in the estimate of the frequencies relative to the range of the frequency of each bin. Note that this is not a problem with the rest of statistics since they are not based in correlations.

Statistical analysis and plots were performed using R version 3.5.1.

3 | RESULTS

We carried out a comparative analysis of the performance of four callers for Pool-seq data based on different conceptual approaches and with different parameter settings (conditions; Table 1). We explored the callers' performance on simulated data generated under the standard neutral model (SNM) or under a joint demographic model for Africa, North America, and Europe (Figure 1). We also explored the performance on *Drosophila melanogaster* sequencing data.

3.1 | SNAPE-pooled and MAPGD outperform heuristic methods under the standard neutral model

Under the standard neutral model (SNM), the average sensitivity of the different callers ranged from 0 to 0.72, with PoolSNP being the most sensitive for most of the sample sizes and coverage (Table S2). However, a good SNP caller should have a good compromise between sensitivity and FDR, since very frequently, an increase in sensitivity is also accompanied by an increase in FDR. Figure 2 shows this relationship for intermediate sample size and

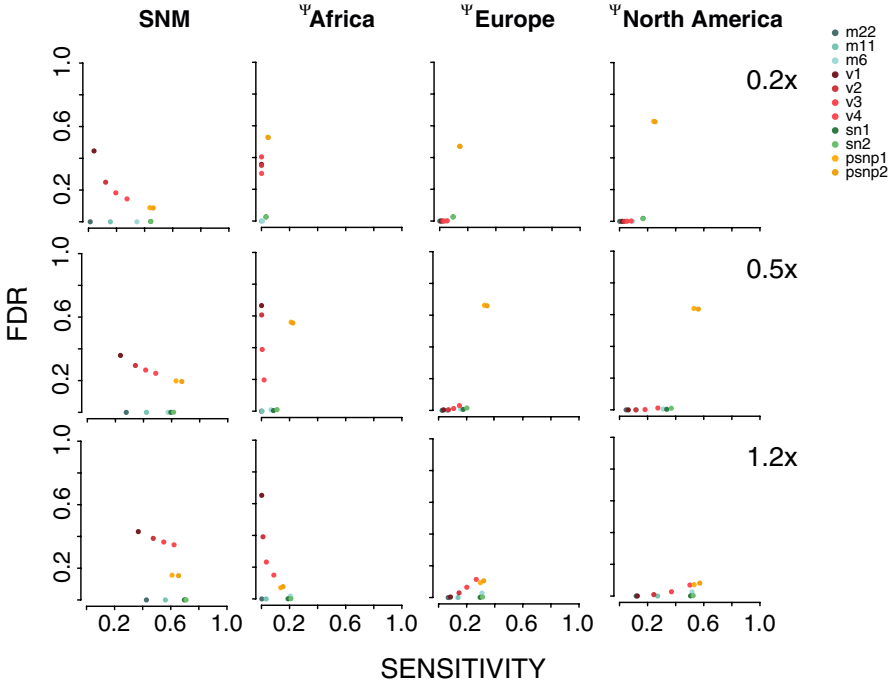

FIGURE 2 Sensitivity versus false discovery rate (FDR) of simulated pools of 50 individuals for the different simulated populations and coverages per chromosome (0.2×, 0.5× and 1.2×) for each condition of the four callers analysed. Each point depicts the mean value of the two statistics for 100 simulated replicates. See Figure S2 for results with simulated pools of 100 and 20 individuals

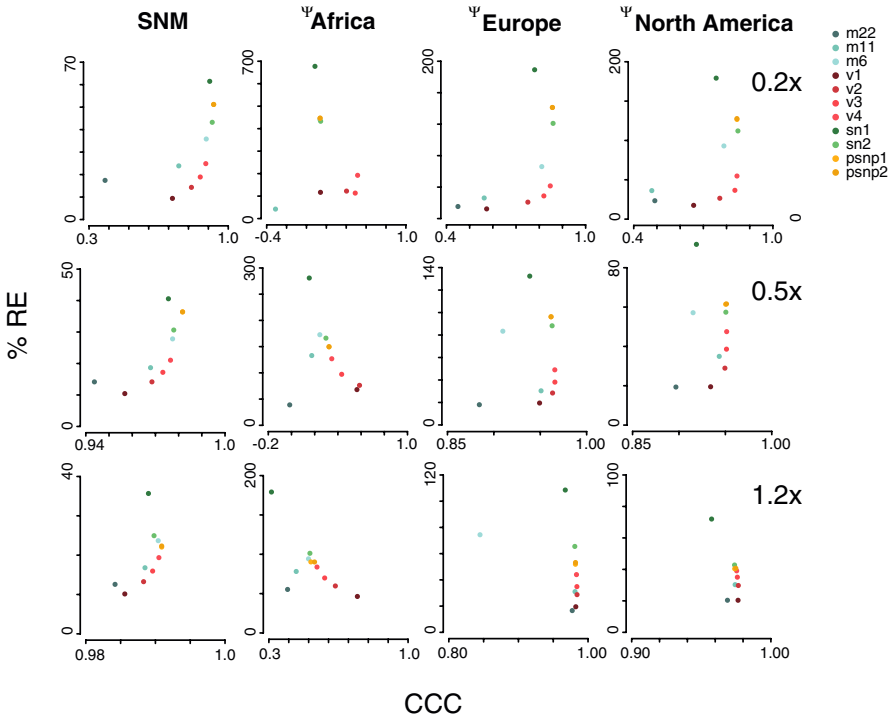

FIGURE 3 Concordance correlation coefficient (CCC) versus the mean percentage of relative error (%RE) of simulated pools of 50 individuals for different simulated populations and coverages per chromosome (0.2×, 0.5× and 1.2×) for each condition of the four callers analysed. Each point depicts the mean value of the two statistics for 100 simulated replicates. See Figure S3 for results with simulated pools of 100 and 20 individuals

three different coverages for the four callers and conditions under the SNM (see Figure S1 for high and low sample sizes). SNAPE-pooled (sn2) and MAPGD (m6) showed the best relationship between these two performance statistics whereas PoolSNP showed excellent sensitivities but higher FDR rates. VarScan showed the best performance in terms of the relationship between sensitivity and FDR with the less restrictive conditions (v4), although the FDR was far above the one exhibited by SNAPE-pooled and MAPGD (Figure 2). Similar results were obtained for low and high sample sizes (Figures S1, S2A,B; Table S2).

In addition to evaluating general performance in terms of SNP detection (sensitivity and FDR), we were also interested in comparing allele frequency estimates across callers and conditions. As expected, the average Concordance Correlation Coefficient (CCC) values, which measures both the accuracy and precision of allele frequency estimates, were maximum with higher sample size and coverage (Figure S2C; Table S2). For intermediate to high sample sizes and coverages, the top callers in terms of SNP detection, MAPGD (m6) and SNAPE-pooled (sn2), showed also excellent performances in estimating allele frequencies (the average CCC values were >0.95).

For most of the sample sizes and coverages, the most restrictive condition of VarScan (v1) and SNAPE-pooled (sn1) were the ones with least and most percentage of relative error (%RE), respectively (Figure S2D). Interestingly, VarScan (v1) clearly outperformed the two top callers for SNP detection although these showed values around or below 30%RE in high and intermediate sample size and coverages (Figures 3 and S3A). PoolSNP also showed a slightly less %RE compared to the two top callers but only when the coverage was high.

In general, the callers that showed higher values of CCC also tend to show higher values of %RE (Figures 3 and S3). The best balance between these two measures was obtained with the less restrictive condition of VarScan (v4).

3.2 | SNAPE-pooled (sn2) showed the best performance in terms of SNP detection in SFSs with an increasing degree of bias towards low-frequency variants

We simulated a join demographic scenario with three populations (hereafter $^{\Psi}$Africa, $^{\Psi}$North America, and $^{\Psi}$Europe; see Section 2), each with a different degree of bias towards low-frequency variants compared to the SNM (Figure 1). This allowed us to assess the performance of the different callers in increasingly challenging conditions, i.e., higher numbers of SNPs in the low-frequency variants category.

3.2.1 | $^{\Psi}$Africa

The average sensitivity of all callers in $^{\Psi}$Africa, the population with the most skewed SFS, was very low, ranged from 0 to 0.3 (Table S3A). In general, SNAPE-pooled and PoolSNP were the most sensitive callers, being SNAPE-pooled the more sensitive when both the number of individuals and the coverage decreased (Figure S4A). However, these two callers showed very different behaviour with respect to FDR (Figure S4B; Table S3A). Again, heuristic callers (Pool-SNP and VarScan) yielded a higher number of false positives compared to nonheuristic callers (SNAPE-pooled and MAPGD), although the low FDR values observed in conditions m22 and m11 of MAPGD were probably due to the low detection power of this caller under these conditions (Figure S4B; Table S3A). The best callers regarding SNP detection were SNAPE-pooled and MAPGD (Figures 2 and S1).

In general, we observed poorer estimates of allele frequencies compared to those for the SNM, with low values of CCC (<0.90; Table S3; Figure S4). Again, despite being one of the callers with higher FDR, VarScan outperformed the other callers in terms of the relationship between CCC and %RE (Figures 3 and S3). MAPGD (m6) and SNAPE-pooled (sn2) showed similar performance in the estimates of the allele frequencies with the later yielding slightly higher CCC values (Table S3A; Figure S4) and less %RE as we moved towards low number of individuals and coverages (Table S3A; Figures S3 and S4C–D).

3.2.2 | $^{\Psi}$Europe

In $^{\Psi}$Europe, where the bias towards low frequency variants is less accentuated, the increase in sensitivity is minimal compared to $^{\Psi}$Africa (ranging from 0 to 0.39; Table S3B). In most cases, the caller with higher sensitivity was PoolSNP (psnp2; Table S3B; Figure S5A), and again the top callers in terms of SNP detection were MAPGD (m6) and SNAPE-pooled (sn2), and both of them detected a similar number of true SNPs (Table S3B; Figure S5A). Instead, as the number of individuals and coverage decreased, SNAPE-pooled was the most sensitive and even outperformed PoolSNP when the number of individuals and coverage was low (Table S3B; Figure S5A). SNAPE-pooled and PoolSNP showed an opposite behaviour regarding the FDR, being SNAPE-pooled the only that maintains a relatively good balance between sensitivity and FDR (Table S3B; Figures 2 and S5B). In this sense, MAPGD (m6) also yielded acceptable FDR values (below 0.05), except for the case with the highest number of individuals and coverage, where FDR increased to 0.1 (Table S3B; Figure S5B).

Regarding CCC values, the behaviour of the different callers was slightly improved with respect to the $^{\Psi}$African population (Table S3B; Figure S5C). SNAPE-pooled (sn2) outperformed MAPGD regarding allele frequency estimation (Table S3B; Figure S5C). When the number of individuals and/or the coverage was intermediate to high, SNAPE-pooled (sn2) performed better than MAPGD (m6), with CCC values classified as substantial and moderate, indicating a reliable estimation of the SFS (see Materials and Methods). SNAPE-pooled (sn2) performed better than SNAPE-pooled (sn1) in terms of %RE in all cases (Table S3B; Figure S5D). Again, VarScan showed the best balance between CCC and %RE (Figure 3).

3.2.3 | $^{\Psi}$North America

Finally, in $^{\Psi}$North America, the less skewed population, the average sensitivity reaches its maximum among the three simulated populations, ranging from 0 to 0.71 (Table S3C). PoolSNP (psnp2) was the caller that detected a higher number of true SNPs (Table S3C; Figure S6A). Among the other callers, MAPGD (m6), VarScan (v4) and SNAPE-pooled (both conditions), showed similar sensitivity when the sample size and the coverage was high or intermediate (Figure S6A). Instead, as the number of individuals and coverage decreased, the best choice to detect true SNPs were PoolSNP and SNAPE-pooled (Table S3C; Figure S6A). However, the high FDR associated to PoolSNP, especially for intermediate to small sample sizes and low coverages (Table S3C; Figure S6B), makes the two conditions of SNAPE-pooled the best options for this task (Table S3C; Figures 2, S1, and S6B). For intermediate and high coverages, SNAPE-pooled

and MAPGD showed the best balance between sensitivity and FDR (Figure 2).

As in the other two simulated populations, the estimates of the allele frequencies with VarScan are more accurate than those of the other callers. However, as it was observed in ΨEurope, the CCC values obtained with SNAPE-pooled (sn2), which had an acceptable FDR values compared to VarScan, indicated that the SNP frequency estimation were also reliable (Table S3C; Figure S6C). SNAPE-pooled (sn2) performed better than SNAPE-pooled (sn1) in terms of %RE in all cases and yielded also slightly better results when the sample size was high or intermediate and the coverage was high. Both SNAPE-pooled (sn2) and MAPG (m6) yielded similar values for the case of intermediate number of individuals and coverage (Table S3C; Figure S6D). For low coverage, the values of %RE were better for MAPG (m6) than for SNAPE-pooled (sn2) (Table S3C; Figure S6D). Finally, the top callers and conditions in terms of SNP detection, SNAPE and MAPGD, showed intermediate performance compared with the other callers (Figures 3 and S3).

3.3 | Callers performed differently for SNP frequencies up to 35%

Since it is well known that the detection of SNPs in Pool-seq data is biased to common frequency variants (Fracassetti et al., 2015; Raineri et al., 2012), as well as analysing all the SNPs together, we also studied the performance of the four callers and their conditions in different parts of the SFS (20 frequency categories). We observed that the most pronounced differences between callers were in the category of rare (<1%), or very low-frequency variants (1%–5%). However, the frequency range where these differences can be observed depends on the sequencing effort, being this range larger as the sample size and the coverage per chromosome decreases. Moreover, the frequency range where these callers perform differently also depended on the shape of the SFS of the simulated scenario. For instance, in the case of the population simulated under the SNM, the differences in sensitivity could be observed in allele frequencies from 0 to 0.35 and from 0.80 to 1 (Figure S7A5). Moreover, the differences in the performance of the different callers were more pronounced in populations with SFSs more skewed toward low-frequency variants (Figures S8A and S9A).

For the majority of frequency categories, MAPGD and VarScan were the callers with sensitivities more affected by the sample size and coverage (Figures S7A–S10A). In addition, MAPGD showed a decrease in sensitivity that was symmetrical, with low- and high-frequency variants more affected than intermediate ones, since this caller estimates a folded SFS and hence only reports the major allele frequency. For very low-frequency variants (<0.05), SNAPE-pooled (sn2) and MAPGD (m6) showed higher sensitivity for the SNM and ΨAfrican population (Figures S7A1 and S8A1) whereas for ΨNorth American and ΨEuropean populations, PoolSNP and MAPGD (m6) were the ones with highest sensitivity (Figures S9A1 and S10A1). In general, the best option in terms of sensitivity was SNAPE-pooled (sn2), specially as the sample size and the coverage decreased.

Regarding the FDR, for common variants in populations with high and intermediate sample size and coverage, most of the callers and conditions performed well, yielding FDR values very low or around 0 (Figures S7B–S10B). Indeed, most of the callers gave values of FDR <0.05 for the frequency range of 0.25–0.90. For rare and low-frequency variants, only SNAPE-pooled (sn2) and MAPGD (m6) yielded FDR values around or below 0.05 in all scenarios, except in some cases where the sample size or the coverage was low (Figures S7B–S10B), showing the best balance between sensitivity and FDR. For those frequencies, PoolSNP yielded relatively good estimates of FDR but only when the sample size and/or coverage was high (Figure S7B1 and S8B1).

Allele frequency estimation across the SFS was evaluated using the %RE (see Section 2). We found that the poorer allele frequency estimates also accumulated in rare and low-frequency variants, although some differences were also found in high-frequency variants (for frequencies from 0 to 0.20 and from 0.85 to 1 in the simulated population under the SNM; Figures S7C–S10C). However, given that relative errors are calculated in percentage units, a high %RE applied to very low-frequencies has very little effect on the estimated SFS. This higher %RE in low-frequency variants progressively decreases as the SFSs are less skewed toward low-frequency variants (is higher in ΨAfrica and ΨEurope than in ΨNorth America and SNM), which is a consequence of the different frequencies in the same frequency category generated by a different underlying demography. For low-frequency variants, among the best callers in terms of sensitivity and FDR, MAPGD (m6) performed slightly better than PoolSNP and SNAPE-pooled but only in some scenarios and in some cases when the sample size and/or the coverage was high (see for example Figures S7C1–C2 and S8C2). For the rest of sample size and conditions, PoolSNP and SNAPE-pooled (sn2) outperform MAPGD (m6), being SNAPE-pooled (sn2) slightly better in low sample size and coverage in the SNM and the ΨAfrican population (Figures S7C6–C8 and S8C6–C8) and vice versa (Figures S10C6–C8 and S9C6–C8). Analogously to the analysis with the total number of SNPs, *VarScan* showed lower %RE for rare and very low-frequency variants (<0.05) than other callers for all simulated scenarios except SNM, where we can observe more similar %RE values. These differences are probably caused by the fact that, for the same frequency bin, both the number of called SNPs (SNPs passing probability/heuristic thresholds) and the SFS of these SNPs can vary importantly across scenarios (especially between simulated demographic scenarios and the SNM).

3.4 | SNAPE-pooled (sn2) and PoolSNP were the callers with the best performance in terms of sensitivity and FDR in Drosophila sequencing data

In addition to assessing the performance of the different callers and settings using coalescent simulations, we further evaluated them

using *D. melanogaster* sequencing data from 30 DGRP strains that were individually and pooled sequenced (Mackay et al., 2012; see Section 2). SNPs called in individual strains were used as the gold standard.

The average sensitivity was higher than that obtained from the simulated SFSs, ranging from 0.28 to 0.84 (Table S4). The highest values (>0.70) were obtained with both PoolSNP (psnp2) and SNAPE-pooled (sn2). As expected, the sensitivity increased when the conditions within each caller were more permissive and with the option of joint for VarScan and PoolSNP where the SNPs could be called in multiple samples (as having multiple samples allows to rescue SNPs that did not pass the calling criteria in one sample). However, it was especially interesting how this increase in sensitivity affected the FDR of each caller and settings. The FDR ranged from 0.02 to 0.08 (Table S4), and as in the simulated data, the highest values were observed for VarScan (v4j) and PoolSNP (*psnp1* and psnp2). Interestingly, the best balance between sensitivity and FDR was obtained with SNAPE-pooled (sn2) and followed very closely by PoolSNP without the joint variant calling option (Figure 4a). For the two callers with the joint variant calling options (VarScan and PoolSNP), the sensitivity increased at the expense of an increase in the FDR compared to when this option was not used (Figure 4a).

On the other hand, CCC values were in general good for all callers and conditions, ranging from 0.89 to 0.96. The callers that performed better in terms of sensitivity and FDR (PoolSNP and SNAPE-pooled [sn2]) yielded similar CCC values (Table S4). The values for these callers were above 0.95, indicating substantial concordance between the SNP frequencies estimated in individual and Pool-seq data. The %RE ranged from 0.16 to 0.46, being SNAPE-pooled (sn1) the caller with the highest %RE (Table S4). As with the sensitivity and FDR, both, the CCC and the %RE increased when the conditions within each caller were more permissive and with the joint variant calling option for those callers where the SNPs can be called in multiple samples (VarScan and PoolSNP). The balance between CCC and %RE was very similar for those callers and conditions that performed well in terms of the balance between sensitivity and FDR –PoolSNP and SNAPE-pooled (sn2)– (Figure 4).

We also analysed the performance of the different callers across the SFS. For frequencies ranging from 0.3 to 0.8 most of the callers and settings showed sensitivities >0.90 (Figure S11). The sensitivity for MAPGD and VarScan showed a decrease for frequencies below 0.3, being this decrease less pronounced for the options less restrictive (m6 and v4) and for the joint variant calling option in the case of VarScan. The callers that performs better for low-frequency variants, PoolSNP and SNAPE-pooled (sn2), showed sensitivities above 0.70 for most of the frequencies <0.10.

In general, the highest FDR was obtained with VarScan, with values ≤0.06 for frequencies between 0.3 and 0.95 and with values up to 0.1 for the rest of the frequencies (Figure S12). The FDR for MAPGD were acceptable, with most of FDR values ≤0.05 for all conditions. Instead, SNAPE-pooled (sn2) and PoolSNP showed low FDR (≤ 0.05) for frequencies above 0.20, but the highest FDR values for rare and low frequencies (from 0 to 0.10).

Regarding the estimate of the frequencies across the SFS, the worst estimates were observed in frequencies up to 0.15 (Figure S13). The values of the %RE above these frequencies were similar or lower than those averaged for all frequencies, with a gradual decrease of the %RE as the frequencies increased (Figure S13). Again, for most of the SFS, SNAPE-pooled (sn2) yielded slightly lower values of the %RE compared with PoolSNP (Figure S13). For VarScan, and for rare and very low-frequency variants (<0.05), we observed the same trend as in the simulated population under the SNM: higher %RE compared with other callers and lower %RE for the less stringent conditions.

Finally, the values of sensitivity, FDR, CCC and %RE obtained with VarScan, SNAPE-pooled and PoolSNP when we considered complete inbreeding were very similar to those obtained considering no inbreeding except for SNAPE-pooled (sn1), where the %RE was higher in the case of inbreeding (Figure S14).

Overall, both SNAPE-pooled (sn2) and PoolSNP were the callers with the best performance in terms of sensitivity and FDR, and with an intermediate performance compared with other callers in frequency estimation. This behaviour was observed both considering all SNPs and across the different frequency categories of the SFS. As in the simulated population under the SNM,

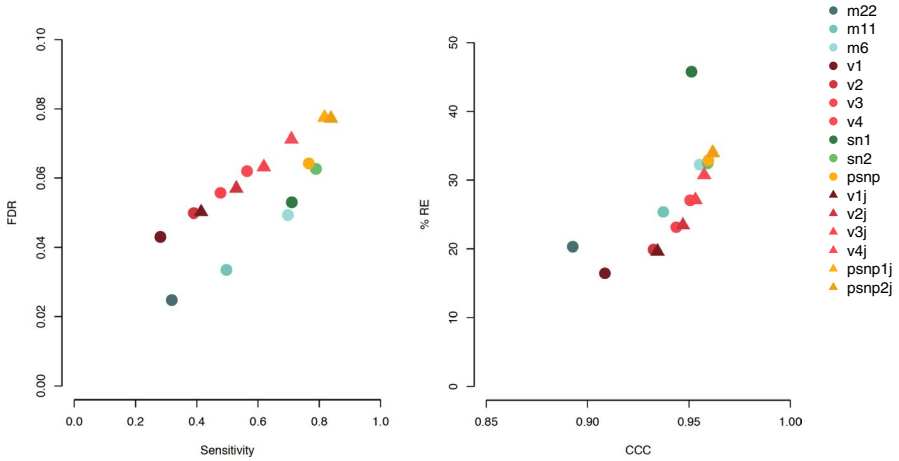

FIGURE 4 Sensitivity versus false discovery rate (FDR) and concordance correlation coefficient (CCC) versus the mean percentage of relative error (%RE) of *Drosophila melanogaster* sequencing data. j, indicates join variant calling

the differences in sensitivity, FDR and %RE among callers can be observed up to frequencies of 0.35.

4 | DISCUSSION

The cost-efficient Pool-seq data strategy is being widely applied to different fields of genomics such as the identification of the genetic basis of a trait (Gautier et al., 2018; Kaiser et al., 2016; Lopes et al., 2016; Olazcuaga et al., 2020), genome-environment association (GEA) analyses (Fischer et al., 2013; Frachon et al., 2018), seasonal adaptation (Bergland et al., 2014; Lamichhaney et al., 2017; Machado et al., 2019), population differentiation (Fischer et al., 2017; Hivert et al., 2018; Martinez Barrio et al., 2016), and domestication genomics (Ayllon et al., 2015; Carneiro et al., 2014; Fleming et al., 2016; Henkel et al., 2019; Vignal et al., 2019), among many others. However, variant calling and allele frequency estimation in Pool-seq data are still challenging. Previous studies have assessed the performance of some SNP callers on Pool-seq data based on different methodologies and using computer simulations and/or sequencing data from model or non-model organisms (Bansal, 2010; Fracassetti et al., 2015; Gautier et al., 2013; Huang et al., 2015; Zhu et al., 2012). Nevertheless, to the best of our knowledge, this is the first work comparing SNP callers that are representatives of a broad range of methodological strategies, *i.e.* Heuristic, Frequentist (maximum likelihood or Fisher's exact test), and Bayesian approaches, and that uses both computer simulations under different demographic scenarios and sequencing data.

Despite the fact that none of the SNP callers evaluated here outperformed the others in all the aspects considered (sensitivity, false discovery rate [FDR], concordance correlation coefficient [CCC], and percentage of relative error [%RE]), our results revealed that full likelihood based methods (SNAPE-pooled and MAPGD) approaches showed the best balance in terms of sensitivity and FDR in both, simulated and sequencing data, although we also observed a good balance for PoolSNP when using sequencing data. Moreover, setting an informative prior for the expected SFS (condition sn2 in the SNAPE-pooled) improves the calling process in populations with a SFS more skewed towards low-frequency variants. Interestingly, the less restrictive conditions for the callers based on frequentist approaches (MAPGD and VarScan) showed a better performance in terms of SNP detection compared with the more restrictive ones. The main contributor to this effect was, in the case of the more restrictive ones, the substantial decrease in sensitivity associated with accepting much smaller probabilities of error. In addition, the top performance callers in terms of SNP detection, SNAPE-pooled (sn2) and MAPGD (m6), also showed a good agreement in terms of correlation and bias between the simulated and the estimated frequencies, with CCC values around or above 0.90–0.95 for high and intermediate samples size and coverages. Similarly, this good agreement was also observed for the DGRP Pool-seq data. Conversely, none of these two callers and conditions were among those that performs better according to the percentage of relative error in frequency estimation (%RE), indicating that the correlation of the frequencies across the SFS but not the magnitude of this estimate was the main contributor to the observed high CCC values. Paradoxically, the callers that performs worse in terms of the balance between sensitivity and FDR (VarScan –with the higher rates of FDR– and MAPGD with stringent criteria of likelihood ratio test –with a decrease in sensitivity–) were in most of the cases the ones that performs better in terms of %RE. In general, the mean absolute differences between the estimated and simulated frequencies ranges between 0.02 and 0.05 for high and intermediate sample size and coverages reaching values of 0.13 for cases with less sequencing effort (Figure S7 C1–C9). These values are in line with those obtained previously in other works (Fracassetti et al., 2015; Rellstab et al., 2013).

One of the most remarkable results of this study is that the differences in the performance among the different callers and settings could mostly be observed up to frequencies of 0.30–0.35. Poor recovery of SNPs at low-frequencies or false-positive SNPs detection can distort the SFS. A biased estimate of the SFS can affect considerably demographic inference, especially for those scenarios where the SFS is skewed toward low-frequency variants such as a recent and rapid population growth, severe bottlenecks or low-migration rates. In addition, a biased estimate of the SFS may influence the detection of positive selection or the identification of traits of interest in association studies (Bloom et al., 2019; Frère et al., 2011). Moreover, an accurate estimate of the SFS if of crucial importance in those works where many individuals or many populations are studied (Frachon et al., 2018; Olazcuaga et al., 2020; Ryu et al., 2018) or in collaborative actions such as DrosEU and DrosRTEC, where different research groups exploit Pool-seq data to answer different questions about the eco-evolutionary dynamics of different species of Drosophila (Kapun et al., 2020; Machado et al., 2019).

In this study, we have explored the performance of four callers based on different approaches on a relatively wide range of conditions and type of data and covering many of the scenarios that are commonly addressed in population genomics studies: different settings for each caller, sample size and coverage using simulations under different demographic scenarios and true pooled samples. Although more complex demographic scenarios or different values for demographic parameters, and caller conditions could be used, our aim was to inform about the selection of the most appropriate SNP caller approach with caller parameters most commonly used or recommended by their authors and in low-frequency enriched SFS rather than a comprehensive exploration of the parametric space of the SNP callers. For instance, for heuristic approaches such as PoolSNP, an exhaustive simulation-based inference for choosing the SNP calling parameters that maximizes the performance of the calling process for specific genomic data can be performed (Kapun et al., 2020). The conclusions drawn from this study could be relevant not only for large-scale population genomics studies but also in cases where the Pool-seq strategy is the only option such as in metagenomic studies and studies with polyploids, among others (Clevenger et al., 2015; Inbar et al., 2020; Pespeni et al., 2013; Shockey et al., 2019).

ACKNOWLEDGEMENTS

We thank Llewellyn Green and Laura Aguilera for technical help with the DNA extractions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (H2020-ERC-2014-CoG-647900).

AUTHOR CONTRIBUTIONS

Sara Guirao-Rico and Josefa González designed the research. Sara Guirao-Rico performed the research and analysed the data. Sara Guirao-Rico and Josefa González wrote the manuscript.

DATA AVAILABILITY STATEMENT

Raw sequenced data is publicly available under NCBI Bioproject accession PRJNA632498. Scripts are available at https://github.com/GonzalezLab/SNP_caller_benchmarking

ORCID

Sara Guirao-Rico https://orcid.org/0000-0001-9896-4665
Josefa González https://orcid.org/0000-0001-9824-027X

REFERENCES

Adams, P. E., Crist, A. L., Young, E. M., Willis, J. H., Phillips, P. C., & Fierst, J. L. (2019). Slow recovery from inbreeding depression generated by the complex genetic architecture of segregating deleterious mutations. *bioRxiv*. https://doi.org/10.1101/862631

Álvarez-Presas, M., Sánchez-Gracia, A., Carbayo, F., Rozas, J., & Riutort, M. (2014). Insights into the origin and distribution of biodiversity in the Brazilian Atlantic forest hot spot: A statistical phylogeographic study using a low-dispersal organism. *Heredity*, *112*, 656–665. https://doi.org/10.1038/hdy.2014.3

Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., Corrado, L., Martinelli Boneschi, F., D'Alfonso, S., & De Bellis, G. (2016). Next generation sequencing of pooled samples: Guideline for variants' filtering. *Scientific Reports*, *6*, 33735. https://doi.org/10.1038/srep33735

Anderson, E. C., Skaug, H. J., & Barshis, D. J. (2014). Next-generation sequencing for molecular ecology: A caveat regarding pooled samples. *Molecular Ecology*, *23*(3), 502–512. https://doi.org/10.1111/mec.12609

Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M. F., Dahle, G., Taranger, G. L., Glover, K. A., Almén, M. S., Rubin, C. J., Edvardsen, R. B., & Wargelius, A. (2015). The vgll3 locus controls age at maturity in wild and domesticated Atlantic Salmon (*Salmo salar* L.) Males. *PLOS Genetics*, *11*(11), e1005628. https://doi.org/10.1371/journal.pgen.1005628

Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, *26*(12), i318–i324. https://doi.org/10.1093/bioinformatics/btq214

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., & Petrov, D. A. (2014). Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLOS Genetics*, *10*(11), e1004775. https://doi.org/10.1371/journal.pgen.1004775

Bianco, E., Nevado, B., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2015). A deep catalog of autosomal single nucleotide variation in the pig. *PLoS One*, *10*(3), e0118867. https://doi.org/10.1371/journal.pone.0118867

Bloom, J. S., Boocock, J., Treusch, S., Sadhu, M. J., Day, L., Oates-Barker, H., & Kruglyak, L. (2019). Rare variants contribute disproportionately to quantitative trait variation in yeast. *Elife*, *8*, e49212. https://doi.org/10.7554/eLife.49212

Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, *18*(1), 77. https://doi.org/10.1186/s13059-017-1212-4

Bustamante, C. D., Wakeley, J., Sawyer, S., & Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, *159*(4), 1779–1788.

Carneiro, M., Rubin, C.-J., Di Palma, F., Albert, F. W., Alfoldi, J., Barrio, A. M., Pielberg, G., Rafati, N., Sayyab, S., Turner-Maier, J., Younis, S., Afonso, S., Aken, B., Alves, J. M., Barrell, D., Bolet, G., Boucher, S., Burbano, H. A., Campos, R., … Andersson, L. (2014). Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science*, *345*(6200), 1074–1079. https://doi.org/10.1126/science.1253714

Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P., & Jackson, S. A. (2015). Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. *Molecular Plant*, *8*(6), 831–846. https://doi.org/10.1016/j.molp.2015.02.002

Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, *186*(1), 41–43. https://doi.org/10.1534/genetics.110.121012

Druley, T. E., Vallania, F. L. M., Wegner, D. J., Varley, K. E., Knowles, O. L., Bonds, J. A., Robison, S. W., Doniger, S. W., Hamvas, A., Cole, F. S., Fay, J. C., & Mitra, R. D. (2009). Quantification of rare allelic variants from pooled genomic DNA. *Nature Methods*, *6*(4), 263–265. https://doi.org/10.1038/nmeth.1307

Duchen, P., Zivkovic, D., Hutter, S., Stephan, W., & Laurent, S. (2013). Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. *Genetics*, *193*(1), 291–301. https://doi.org/10.1534/genetics.112.145912

Esteve-Codina, A., Paudel, Y., Ferretti, L., Raineri, E., Megens, H.-J., Silió, L., Rodríguez, M. C., Groenen, M. A. M., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. *BMC Genomics*, *14*, 148. https://doi.org/10.1186/1471-2164-14-148

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLOS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576. https://doi.org/10.1111/mec.12522

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., Holderegger, R., & Widmer, A. (2017). Estimating genomic diversity and population differentiation - An empirical comparison of microsatellite and SNP variation in Arabidopsis halleri. *BMC Genomics*, *18*(1), 69. https://doi.org/10.1186/s12864-016-3459-7

Fischer, M. C., Rellstab, C., Tedder, A., Zoller, S., Gugerli, F., Shimizu, K. K., & Widmer, A. (2013). Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis halleri from the Alps. *Molecular Ecology*, *22*(22), 5594–5607. https://doi.org/10.1111/mec.12521

Fleming, D. S., Koltes, J. E., Fritz-Waters, E. R., Rothschild, M. F., Schmidt, C. J., Ashwell, C. M., Persia, M. E., Reecy, J. M., & Lamont, S. J. (2016). Single nucleotide variant discovery of highly inbred Leghorn and Fayoumi chicken breeds using pooled whole genome resequencing data reveals insights into phenotype differences. *BMC Genomics*, *17*(1), 812. https://doi.org/10.1186/s12864-016-3147-7

Fournier, T., Abou Saada, O., Hou, J., Peter, J., Caudal, E., & Schacherer, J. (2019). Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *Elife*, *8*. https://doi.org/10.7554/eLife.49258

Fracassetti, M., Griffin, P. C., & Willi, Y. (2015). Validation of pooled whole-genome re-sequencing in Arabidopsis lyrata. *PLoS One*, *10*(10), e0140462. https://doi.org/10.1371/journal.pone.0140462

Frachon, L., Bartoli, C., Carrère, S., Bouchez, O., Chaubet, A., Gautier, M., Roby, D., & Roux, F. (2018). A genomic map of climate adaptation in *Arabidopsis thaliana* at a Micro-Geographic Scale. *Frontiers in Plant Science*, *9*, 967. https://doi.org/10.3389/fpls.2018.00967

Frère, C. H., Prentis, P. J., Gilding, E. K., Mudge, A. M., Cruickshank, A., & Godwin, I. D. (2011). Lack of low frequency variants masks patterns of non-neutral evolution following domestication. *PLoS One*, *6*(8), e23041. https://doi.org/10.1371/journal.pone.0023041

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, *186*(1), 207–218. https://doi.org/10.1534/genetics.110.114397

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., & Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), 3766–3779. https://doi.org/10.1111/mec.12360

Gautier, M., Yamaguchi, J., Foucaud, J., Loiseau, A., Ausset, A., Facon, B., Gschloessl, B., Lagnel, J., Loire, E., Parrinello, H., Severac, D., Lopez-Roques, C., Donnadieu, C., Manno, M., Berges, H., Gharbi, K., Lawson-Handley, L., Zang, L.-S., Vogel, H., … Prud'homme, B. (2018). The genomic basis of color pattern polymorphism in the harlequin ladybird. *Current Biology*, *28*(20), 3296–3302.e3297. https://doi.org/10.1016/j.cub.2018.08.023

Gravel, S., Henn, B., Gutenkunst, R., Indap, A., Marth, G., Clark, A., Yu, F., Gibbs, R., 1000 Genomes Project, & Bustamante, C. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences USA*, *108*(29), 11983–11988. https://doi.org/10.1073/pnas.1019276108

Guirao-Rico, S., Ramirez, O., Ojeda, A., Amills, M., & Ramos-Onsins, S. E. (2018). Porcine Y-chromosome variation is consistent with the occurrence of paternal gene flow from non-Asian to Asian populations. *Heredity*, *120*, 63–76. https://doi.org/10.1038/s41437-017-0002-9

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Healy, T. M., & Burton, R. S. (2020). Strong selective effects of mitochondrial DNA on the nuclear genome. *Proceedings of the National Academy of Sciences USA*, *117*(12), 6616–6621. https://doi.org/10.1073/pnas.1910141117

Henkel, J., Saif, R., Jagannathan, V., Schmocker, C., Zeindler, F., Bangerter, E., Herren, U., Posantzis, D., Bulut, Z., Ammann, P., Drögemüller, C., Flury, C., & Leeb, T. (2019). Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. *PLoS Genetics*, *15*(12), e1008536. https://doi.org/10.1371/journal.pgen.1008536

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics*, *210*(1), 315–330. https://doi.org/10.1534/genetics.118.300900

Huang, H. W., Mullikin, J. C., Hansen, N. F., & Program, N. C. S. (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics*, *16*, 235. https://doi.org/10.1186/s12859-015-0624-y

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

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Inbar, S., Cohen, P., Yahav, T., & Privman, E. (2020). Comparative study of population genomic approaches for mapping colony-level traits. *PLoS Computational Biology*, *16*(3), e1007653. https://doi.org/10.1371/journal.pcbi.1007653

Jónás, Á., Taus, T., Kosiol, C., Schlötterer, C., & Futschik, A. (2016). Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics*, *204*(2), 723–735. https://doi.org/10.1534/genetics.116.191197

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., Neumann, T., Nguyen, L.-T., Betancourt, A. J., Hummel, T., Vogel, H., Dorner, S., Heyd, F., von Haeseler, A., & Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, *540*, 60–73. https://doi.org/10.1038/nature20151

Kapun, M., Barrón, M. G., Staubach, F., Obbard, D. J., Wiberg, R. A. W., Vieira, J., & González, J. (2020). Genomic analysis of European *Drosophila melanogaster* populations reveals longitudinal structure. *Continent-Wide Selection, and Previously Unknown DNA Viruses*, *Molecular Biology and Evolution*, *37*(9), 2661–2678. https://doi.org/10.1093/molbev/msaa120

Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, *336*(6082), 740–743. https://doi.org/10.1126/science.1217283

Kim, Y., & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, *160*(2), 765–777.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., & Ding, L. I. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, *25*(17), 2283–2285. https://doi.org/10.1093/bioinformatics/btp373

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356. https://doi.org/10.1186/s12859-014-0356-4

Koropoulis, A., Alachiotis, N., & Pavlidis, P. (2020). Detecting positive selection in populations using genetic data. *Methods in Molecular Biology*, *2090*, 87–123. https://doi.org/10.1007/978-1-0716-0199-0_5

Lamichhaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., Singh, R., Ruzzante, D. E., & Andersson, L. (2017). Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proceedings of the National Academy of Sciences USA*, *114*(17), E3452–E3461. https://doi.org/10.1073/pnas.1617728114

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), 639–647. https://doi.org/10.1111/1755-0998.12995

Lohmueller, K. E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genetics*, *10*(5), e1004379. https://doi.org/10.1371/journal.pgen.1004379

Lopes, R. J., Johnson, J. D., Toomey, M. B., Ferreira, M. S., Araujo, P. M., Melo-Ferreira, J., Andersson, L., Hill, G. E., Corbo, J. C., & Carneiro, M. (2016). Genetic basis for red coloration in birds. *Current Biology*, *26*(11), 1427–1434. https://doi.org/10.1016/j.cub.2016.03.076

Lynch, M., Bost, D., Wilson, S., Maruki, T., & Harrison, S. (2014). Population-genetic inference from pooled-sequencing data. *Genome Biology and Evolution*, *6*(5), 1210–1218. https://doi.org/10.1093/gbe/evu085

Machado, H. E., Bergland, A. O., Taylor, R., Tilk, S., Behrman, E., Dyer, K., … Petrov, D. A. (2019). Broad geographic sampling reveals predictable,

pervasive, and strong seasonal adaptation in Drosophila. *bioRxiv*. https://doi.org/10.1101/337543

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y. I., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., … Gibbs, R. A. (2012). The *Drosophila melanogaster* genetic reference panel. *Nature*, *482*(7384), 173–178. https://doi.org/10.1038/nature10811

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*, 10–12.

Martin, S. H., Möst, M., Palmer, W. J., Salazar, C., McMillan, W. O., Jiggins, F. M., & Jiggins, C. D. (2016). Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. *Genetics*, *203*(1), 525–541. https://doi.org/10.1534/genetics.115.183285

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H. E., Dainat, J., Ekman, D., Höppner, M., Jern, P., Martin, M., Nystedt, B., Liu, X., Chen, W., Liang, X., Shi, C., Fu, Y., Ma, K., Zhan, X., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *Elife*, *5*, e12081. https://doi.org/10.7554/eLife.12081

McBride, G. B. (2005). A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. In *NIWA Client Report: HAM2005-062*.

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

Micheletti, S. J., & Narum, S. R. (2018). Utility of pooled sequencing for association mapping in nonmodel organisms. *Molecular Ecology Resources*, *18*(4), 825–837. https://doi.org/10.1111/1755-0998.12784

Neethiraj, R., Hornett, E. A., Hill, J. A., & Wheat, C. W. (2017). Investigating the genomic basis of discrete phenotypes using a Pool-Seq-only approach: New insights into the genetics underlying colour variation in diverse taxa. *Molecular Ecology*, *26*(19), 4990–5002. https://doi.org/10.1111/mec.14205

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St. Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L. I., … Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, *337*(6090), 100–104. https://doi.org/10.1126/science.1217876

Nevado, B., & Perez-Enciso, M. (2015). Pipeliner: Software to evaluate the performance of bioinformatics pipelines for next-generation resequencing. *Molecular Ecology Resources*, *15*(1), 99–106. https://doi.org/10.1111/1755-0998.12286

Olazcuaga, L., Loiseau, A., Parrinello, H., Paris, M., Fraimout, A., Guedot, C., Diepenbrock, L. M., Kenis, M., Zhang, J., Chen, X., Borowiec, N., Facon, B., Vogt, H., Price, D. K., Vogel, H., Prud'homme, B., Estoup, A., & Gautier, M. (2020). A whole-genome scan for association with invasion success in the fruit fly *Drosophila suzukii* using contrasts of allele frequencies corrected for population structure. *Molecular Biology and Evolution*, *37*, 2369–2385. https://doi.org/10.1093/molbev/msaa098

Peischl, S., Dupanloup, I., Foucal, A., Jomphe, M., Bruat, V., Grenier, J.-C., Gouy, A., Gilbert, K. J., Gbeha, E., Bosshard, L., Hip-Ki, E., Agbessi, M., Hodgkinson, A., Vézina, H., Awadalla, P., & Excoffier, L. (2018). Relaxed selection during a recent human expansion. *Genetics*, *208*(2), 763–777. https://doi.org/10.1534/genetics.117.300551

Pespeni, M. H., Sanford, E., Gaylord, B., Hill, T. M., Hosfelt, J. D., Jaris, H. K., LaVigne, M., Lenz, E. A., Russell, A. D., Young, M. K., & Palumbi, S. R. (2013). Evolutionary change during experimental ocean acidification. *Proceedings of the National Academy of Sciences USA*, *110*(17), 6937–6942. https://doi.org/10.1073/pnas.1220673110

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., & Pérez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics*, *13*, 239. https://doi.org/10.1186/1471-2105-13-239

Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., & Fischer, M. C. (2013). Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS One*, *8*(11), e80422. https://doi.org/10.1371/journal.pone.0080422

Ronen, R., Udpa, N., Halperin, E., & Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics*, *195*(1), 181–193. https://doi.org/10.1534/genetics.113.152587

Ryu, S., Han, J., Norden-Krichmar, T. M., Schork, N. J., & Suh, Y. (2018). Effective discovery of rare variants by pooled target capture sequencing: A comparative analysis with individually indexed target capture sequencing. *Mutation Research*, *809*, 24–31. https://doi.org/10.1016/j.mrfmmm.2018.03.007

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals—Mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, *15*(11), 749–763. https://doi.org/10.1038/nrg3803

Shockey, A. C., Dabney, J., & Pepperell, C. S. (2019). Effects of host, sample, and *in vitro* culture on genomic diversity of pathogenic Mycobacteria. *Frontiers in Genetics*, *10*, 477. https://doi.org/10.3389/fgene.2019.00477

Smit, A. F., Hubley, R., & Green, P. (2013-2015). RepeatMasker Open-4.0. Retrieved from http://www.repeatmasker.org

Stevenson, M., Nunes, T., Sanchez, J., Thornton, R., Reiczigel, J., Robison-Cox, J., & Sebastiani, P. (2019). EpiR: An R package for the analysis of epidemiological data. Retrieved from https://cran.r-project.org/web/packages/epiR/index.html

Tennessen, J. A., Madeoy, J., & Akey, J. M. (2010). Signatures of positive selection apparent in a small sample of human exomes. *Genome Research*, *20*(10), 1327–1334. https://doi.org/10.1101/gr.106161.110

Tilk, S., Bergland, A., Goodman, A., Schmidt, P., Petrov, D., & Greenblum, S. (2019). Accurate allele frequencies from ultra-low coverage pool-seq samples in evolve-and-resequence experiments. *G3 (Bethesda)*, *9*(12), 4159–4168. https://doi.org/10.1534/g3.119.400755

Vieira, F. G., Fumagalli, M., Albrechtsen, A., & Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Research*, *23*(11), 1852–1861. https://doi.org/10.1101/gr.157388.113

Vignal, A., Boitard, S., Thébault, N., Dayo, G.-K., Yapi-Gnaore, V., Youssao Abdou Karim, I., Berthouly-Salazar, C., Pálinkás-Bodzsár, N., Guémené, D., Thibaud-Nissen, F., Warren, W. C., Tixier-Boichard, M., & Rognon, X. (2019). A guinea fowl genome assembly provides new evidence on evolution following domestication and selection in galliformes. *Molecular Ecology Resources*, *19*(4), 997–1014. https://doi.org/10.1111/1755-0998.13017

Vy, H. M., & Kim, Y. (2015). A composite-likelihood method for detecting incomplete selective sweep from population genomic data. *Genetics*, *200*(2), 633–649. https://doi.org/10.1534/genetics.115.175380

Wei, Z., Wang, W., Hu, P., Lyon, G. J., & Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research*, *39*(19), e132. https://doi.org/10.1093/nar/gkr599

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., Khor, C. C., Petric, R., Hibberd, M. L., & Nagarajan, N. (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201. https://doi.org/10.1093/nar/gks918

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., Belbin, G. M., Bien, S. A., Cheng, I., Cullina, S., Hodonsky, C. J., Hu, Y., Huckins, L. M., Jeff, J., Justice, A. E., … Carlson, C. S. (2019). Genetic analyses

of diverse populations improves discovery for complex traits. *Nature*, *570*(7762), 514–518. https://doi.org/10.1038/s41586-019-1310-4

Zhu, Y., Bergland, A. O., González, J., & Petrov, D. A. (2012). Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. *PLoS One*, *7*(7), e41901. https://doi.org/10.1371/journal.pone.0041901

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Guirao-Rico S, González J. Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. *Mol Ecol Resour*. 2021;21:1216–1229. https://doi.org/10.1111/1755-0998.13343