RESOURCE ARTICLE

# Evaluating the accuracy of variant calling methods using the frequency of parent-offspring genotype mismatch

Russ J. Jasper[1] | Tegan Krista McDonald[1] | Pooja Singh[1,2,3] | Mengmeng Lu[1] | Clément Rougeux[1] | Brandon M. Lind[4] | Sam Yeaman[1]

[1]Department of Biological Sciences, University of Calgary, Calgary, Alberta, Canada

[2]Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

[3]EAWAG, Swiss Federal Institute of Aquatic Science and Technology, Kastanienbaum, Switzerland

[4]Centre for Forest Conservation Genetics and Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, British Columbia, Canada

**Correspondence**
Russ J. Jasper, Department of Biological Sciences, University of Calgary, Calgary, AB, Canada.
Email: rjjasper@ucalgary.ca

## Abstract

The use of next-generation sequencing (NGS) data sets has increased dramatically over the last decade, but there have been few systematic analyses quantifying the accuracy of the commonly used variant caller programs. Here we used a familial design consisting of diploid tissue from a single lodgepole pine (*Pinus contorta*) parent and the maternally derived haploid tissue from 106 full-sibling offspring, where mismatches could only arise due to mutation or bioinformatic error. Given the rarity of mutation, we used the rate of mismatches between parent and offspring genotype calls to infer the single nucleotide polymorphism (SNP) genotyping error rates of FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan. With baseline filtering HaplotypeCaller and UnifiedGenotyper yielded more SNPs and higher error rates by one to two orders of magnitude, whereas FreeBayes, SAMtools and VarScan yielded lower numbers of SNPs and more modest error rates. To facilitate comparison between variant callers we standardized each SNP set to the same number of SNPs using additional filtering, where UnifiedGenotyper consistently produced the smallest proportion of genotype errors, followed by HaplotypeCaller, VarScan, SAMtools, and FreeBayes. Additionally, we found that error rates were minimized for SNPs called by more than one variant caller. Finally, we evaluated the performance of various commonly used filtering metrics on SNP calling. Our analysis provides a quantitative assessment of the accuracy of five widely used variant calling programs and offers valuable insights into both the choice of variant caller program and the choice of filtering metrics, especially for researchers using non-model study systems.

**KEYWORDS**
bioinformatics, genomics, genotyping, next-generation sequencing, non-model, single nucleotide polymorphism

# 1 | INTRODUCTION

The decreasing cost and ease of producing short-read next-generation sequencing (NGS) data sets has transformed our understanding of organismal diversity across the tree of life. Next-generation sequencing offers new opportunities to empirically test both basic and applied hypotheses relating to the molecular ecology and evolutionary genetics within and among populations. However, transforming abundant raw sequence data into biologically meaningful genetic data is nontrivial. Genotyping errors can be introduced at several steps of NGS data analysis, which by extension may induce biases in subsequent inference. This becomes particularly problematic in non-model organisms with large and complex genomes and fragmented genome assemblies. While SNP errors may be less of an issue for window-based approaches like genome scans, where errors add noise to estimating statistics of interest (e.g., mean $F_{ST}$), they are probably more problematic for analyses that depend on estimating the site frequency spectrum (SFS; e.g., Lapierre et al., 2017) or the number of singleton alleles (Field et al., 2016). Whatever the application, it is clear that understanding the performance and sources of biases and errors from such tools is critical for the success of any NGS-based project.

Various open-source programs have been developed to identify genomic variants, such as single nucleotide polymorphisms (SNPs) or insertions and deletions (indels), from short read data. Commonly used programs include Freebayes (Garrison & Marth, 2012), HaplotypeCaller and UnifiedGenotyper from the Genome Analysis Tool Kit (Van der Auwera & O'Connor, 2020), SAMtools (Li et al., 2009), and VarScan (Koboldt et al., 2012), which are widely used in NGS-based projects in both model and non-model systems. However, it is often difficult to know a priori how well a given tool will perform given the genomic resources available for any particular study organism, or how the study design will interact with the underlying assumptions of such tools, which are often benchmarked with model organism data.

Several studies comparing variant calling programs exist in the literature (Bian et al., 2018; Chen et al., 2019; Cornish & Guda, 2015; Hwang et al., 2015; Sandmann et al., 2019). Most aim to test program efficiency (Hwang et al., 2015) or to evaluate estimations of the precision and sensitivity of the variant calling tools (Sandmann et al., 2019), and are often conducted with human genomic data (Bian et al., 2018; Cornish & Guda, 2015; Sandmann et al., 2019). These studies often conclude that there are substantial differences in precision and sensitivity across tools which depend in large part on aspects of the data such as sample size and coverage as well as the genomic resources available (Cornish & Guda, 2015; Sandmann et al., 2019, but see Bian et al., 2018). As genome-scale data sets are now common for non-model organisms with limited or fragmented reference genomes, it is necessary to expand upon these studies to understand and establish best practices for systems where genomic resources are limited.

Conifers are non-model organisms with genomes recalcitrant to chromosome-scale assembly, especially under budgetary constraint.

For instance, conifers often have exceptionally large genomes (20–40 Gbp; Neale et al., 2017) with histories of whole-genome duplication (Zheng et al., 2015), gene family expansion (Scott et al., 2020; De La Torre et al., 2014), transposable element dynamics (Scott et al., 2020; Wang et al., 2020; Yi et al., 2018), and extensive repeat regions (Wegrzyn et al., 2014). These complexities present a major challenge for NGS data analysis and downstream hypothesis testing in conifers (Lind et al., 2022; Shu & Moran, 2020). Such challenges can be alleviated by quantifying the accuracy of SNP calling when using the above-mentioned variant calling tools. High-quality databases exist for model organisms to calibrate existing programs and account for biases in the data, yet such resources remain elusive for many non-model organisms. A unique biological attribute of conifer reproduction, however, offers us the opportunity to assess the genotyping accuracy of these variant calling tools and to do so in a non-model system. Fertilized conifer seeds contain a megagametophyte, a cluster of haploid tissue surrounding the embryo and supplying it with nutrients. This megagametophyte haploid tissue only contains the maternal contribution of the offspring DNA and has been used to advantage in linkage mapping (Bernhardsson et al., 2019; Neves et al., 2014; Pavy et al., 2017) for example, as the phase is already resolved. By leveraging this information for a set of related individuals (i.e., full-sibling megagametophyte tissues and the maternal tree), we can gain insight into the accuracy of genotype calls by quantifying concordance between the haploid and parental genotypes. Given that the offspring haploid genotype at any given locus could only have been inherited from a single source, if we assume no de novo mutation, then any discordance between the haploid offspring genotype calls and the maternal genotype call must have come as a result of genotyping error.

Here, we used a familial design consisting of diploid tissue from a single lodgepole pine (*Pinus contorta*) parent and the maternally derived megagametophyte haploid tissue from 106 full-sibling offspring to evaluate similarities and differences among SNP sets generated from FreeBayes, HaplotypeCaller, Samtools, UnifiedGenotyper, and VarScan. We used the rate of mismatches between parent and offspring genotype calls to infer the genotype error rate of each variant caller, given the rarity of mutation within one generation and describe how the number of SNPs and proportion of genotyping errors behave under varying threshold levels for the most commonly used quality metrics during filtering steps. Our results shed light on how the choice of variant calling program can affect the result and interpretation of genomic analyses in non-model organisms lacking extensive genomic resources.

# 2 | MATERIALS AND METHODS

## 2.1 | Sample preparation and DNA extraction

Our sample data came from a *Pinus contorta* linkage mapping population consisting of two parental trees located in Bulkley Valley, British Columbia, Canada and the 106 F1 seeds from a single

outcross, provided to us by the University of British Columbia (https://coadaptree.forestry.ubc.ca). Needles were collected from the parental trees and stored at –20°C prior to DNA isolation. The F1 seeds were hydrated at room temperature and the tissue layers separated under a microscope using sterile technique. After the seed coat was completely removed, the megagametophyte tissue was carefully separated from the embryo using a surgical blade and tweezers. DNA was then extracted from the needles of the parental trees or from the megagametophyte tissue of the F1 seeds using the NucleoSpin Plant II Mini kit (Macherey–Nagel GmbH & Co.), following modifications recommended in García and Escribano-Ávila (2016).

## 2.2 | Probe design and sequence capture

Due to the substantial size and complexity of conifer genomes, exome sequence capture is the preferred genotyping method for the taxon (Lind et al., 2022). We designed our capture probes based on previous *Pinus contorta* exome capture probes (Suren et al., 2016) and removed all probes which failed to effectively capture the target sequences. In addition, we included probes corresponding to previously described pathogen response genes (Lu et al., 2021), filtering out those which targeted exon sequences of less than 100 base pairs (bp). We then submitted our probe sequences to Roche NimbleGen for Custom SeqCapEZ probe design. We achieved a final capture space of ~44 Mbp.

Following DNA extraction and probe design, approximately 100 ng of DNA from each sample was used to construct a barcoded library (Kapa, Dual-Indexed Adapter Kit). Sequence capture was performed following the SeqCap EZ HyperCap Workflow User's Guide Version 2.0 (Roche Sequencing Solutions, Inc.). The enriched capture libraries were then multiplexed and sequenced in four lanes, with 23–30 libraries per lane, on an Illumina HiSeq4000 instrument at the Centre d'expertise et de services Génome Québec, Montreal, Canada, resulting in 351 Gbp of 150-bp paired-end reads.

## 2.3 | Sequence read processing, mapping, and SNP calling

We used fastp v0.19.5 (Chen et al., 2018) to trim sample reads, removing adapters and polyG tails, in addition to using a 5 bp sliding window and removing all windows with less than 30 mean quality. We then removed all reads less than 75 bp in length. Following processing with fastp, 95% of bases were greater than 20 quality and 86% greater than 30 quality. We then used BWA-MEM v0.7.17 (Li & Durbin, 2009) to align reads against the congeneric loblolly pine (*Pinus taeda*) reference genome v2.01 (Zimin et al., 2017; https://treegenesdb.org/FTP/Genomes/Pita/v2.01), as a reference genome does not yet exist for *P. contorta*. We used SAMtools v1.9 to

sort aligned reads, only keeping proper pairs, and PICARD v2.18.9 (http://broadinstitute.github.io/picard) to mark and remove duplicates. Where applicable, we converted the BAM files to mpileup files using SAMtools.

We then used the following variant caller programs: FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan, to call SNPs using the same BAM (or mpileup) files as input. After calling SNPs we performed an initial baseline level of filtering, first using criteria specific to each caller based on commonly used practices for the caller, and then secondly filtering each data set by a common set of filtering criteria (Table 1). The common filtering thresholds we used for each caller required that sites were called in both the parent and the F1 sample, did not have greater than 50% missingness, and were not multiallelic. We used both VCFtools v0.1.14 (Danecek et al., 2011) and R v4.0.5 (R Core Team, 2021) to achieve the above-described filtering.

**TABLE 1** Baseline filtering criteria. Set of filtering criteria unique to each variant caller program and set of common filtering criteria used across all programs. Criteria describe the sites removed

| FreeBayes |
| --- |
| Sites with less than 30 quality (QUAL) |
| Genotype calls with less than 5 depth (DP) |
| Genotype calls with less than 20 genotype quality (GQ) |

| HaplotypeCaller |
| --- |
| Sites with greater than 60 Fisher strand (FS) |
| Sites with less than 40 mapping quality (MQ) |
| Sites with less than –12.5 mapping quality rank sum test (MQRankSum) |
| Sites with less than 30 quality (QUAL) |
| Sites with less than 2.0 quality by depth (QD) |
| Sites with less than –8.0 read position rank sum test (ReadPosRankSum) |
| Sites with greater than 3.0 strand odds ratio (SOR) |
| Genotype calls with less than 20 genotype quality (GQ) |

| SAMtools |
| --- |
| Sites with less than 20 quality (QUAL) |
| Genotype calls with less than 5 depth (DP) |

| UnifiedGenotyper |
| --- |
| Sites with greater than 60 Fisher strand (FS) |
| Sites with less than 40 mapping quality (MQ) |
| Sites with less than –12.5 mapping quality rank sum test (MQRankSum) |
| Sites with less than 30 quality (QUAL) |
| Sites with less than 2.0 quality by depth (QD) |
| Sites with less than –8.0 read position rank sum test (ReadPosRankSum) |
| Sites with greater than 3.0 strand odds ratio (SOR) |

| VarScan |
| --- |
| Genotype calls with less than 10 depth (DP) |
| Genotype calls with less than 20 genotype quality (GQ) |
| Heterozygote genotype calls |

| Common filters |
| --- |
| Sites not called in both parent and offspring |
| Sites with greater than 50% missingness |
| Multiallelic sites |

## 2.4 | Program-specific details for SNP calling and filtering

### 2.4.1 | FreeBayes

We used FreeBayes v1.3.1–17-gaa2ace8 to call SNPs separately in both the offspring samples and the parental sample to account for the ploidy difference. The resulting VCF files were merged using VCFtools. Sites with less than 30 quality score, and genotype calls with less than 20 genotype quality and/or less than 5 depth were filtered out, before applying the set of common filters as described above (Table 1).

### 2.4.2 | HaplotypeCaller

We used GATK v4.1 HaplotypeCaller in -ERC mode to generate 107 individual sample g.vcf files for both parent and offspring. Individual g.vcf's were combined in batches of 20 using CombineGVCFs into a single multiple sample g.vcf comprising the full mixed ploidy cohort. GenotypeGVCF was used to call SNPs on the full cohort. Both commands for combining and genotyping are able to handle mixed ploidy of samples. We used the standard GATK hard-filter expression (https://gatk.broadinstitute.org/hc/en-us/articles/360035532412?id=11097) as our initial filtering step, as *P. contorta* lacks a benchmark set of high-quality SNPs for calibration and therefore we could not use variant quality score recalibration (VQSR). After the hard-filter expression we filtered out sites with less than 30 quality score, genotype calls with less than 20 genotype quality, and then applied the common set of filters (Table 1).

### 2.4.3 | SAMtools

We used the bcftools v1.9 utility from SAMtools to call SNPs on the offspring and parent samples as a cohort. Filtering was performed by removing sites with less than 20 quality score, genotype calls with less than 5 depth, and then applying the set of common filters (Table 1). The -G flag was set to the default (i.e., all samples were treated as one population).

### 2.4.4 | UnifiedGenotyper

We used GATK v3.8 UnifiedGenotyper to call SNPs separately in both the offspring samples and the parental sample. As with HaplotypeCaller, because VQSR was not applicable, we filtered using the hard-filter expression (https://gatk.broadinstitute.org/hc/en-us/articles/360035532412?id=11097), further filtered sites with less than 30 quality score, and then filtered using the common set of filters (Table 1).

### 2.4.5 | VarScan

We used VarScan v2.4.4 to call SNPs on the offspring and parental data separately, using a *p*-value of.05. We then merged the resulting VCF files using VCFtools and filtered out all genotype calls with less than 20 genotype quality and/or less than 10 depth. Additionally, as VarScan does not have a ploidy setting (i.e., all genotype calls are diploid) and our offspring sample data were haploid, we removed all heterozygote genotype calls. Finally, we applied the common set of filters (Table 1).

## 2.5 | Comparing variant caller programs

Variant callers differed considerably in the number of sites they called using the baseline filtering (e.g., UnifiedGenotyper yielded ~3 M vs ~100 k for VarScan; Table 2). We found that the mismatch rates between parent and offspring genotype calls roughly scaled with the number of SNPs called (Table 2; Figure S1), and in order to compare mismatch rates between different callers we standardized their output by adjusting the filtering criteria so that each caller yielded a similar number of SNPs (Tables S1–S5). This additional filtering step was conducted in an iterative manner to assess the effect of varying stringency on mismatch rates, using the QUAL, DP, and GQ metrics (but only DP and GQ for VarScan) and varying each metric according to their empirical distribution. For example, to get approximately $x$ SNPs with a given caller, we would increase the QUAL, DP, and GP filtering criteria from the $y$th to the $z$th percentile of each of their respective distributions. We calculated two different mismatch rates, a by-genotype rate and a by-site rate. The

| | Sites | Genotypes | Site mismatch rate | Genotype mismatch rate |
|---|---|---|---|---|
| FreeBayes | $1.19 \times 10^5$ | $1.05 \times 10^7$ | $1.03 \times 10^{-2}$ | $2.39 \times 10^{-3}$ |
| HaplotypeCaller | $2.24 \times 10^6$ | $1.87 \times 10^8$ | $2.41 \times 10^{-1}$ | $1.18 \times 10^{-2}$ |
| SAMtools | $4.59 \times 10^5$ | $4.16 \times 10^7$ | $5.36 \times 10^{-2}$ | $3.08 \times 10^{-3}$ |
| UnifiedGenotyper | $3.49 \times 10^6$ | $2.86 \times 10^8$ | $1.51 \times 10^{-1}$ | $2.03 \times 10^{-2}$ |
| VarScan | $1.16 \times 10^5$ | $9.01 \times 10^6$ | $2.49 \times 10^{-3}$ | $5.48 \times 10^{-4}$ |

**TABLE 2** The number of sites and genotypes called, and the by-site and by-genotype mismatch rates for each variant caller program after base filtering was applied

by-genotype mismatch rate was calculated as the proportion of mismatched parent-offspring genotype calls out of the total number of offspring genotypes called. The by-site mismatch rate was calculated as the proportion of sites with at least one mismatched parent-offspring genotype call out of the total number of sites. For example, if 25/100 sites each had any nonzero number of mismatches, meaning 75/100 sites were completely mismatch free, we would calculate this as a 25% by-site mismatch rate.

# 3 | RESULTS

## 3.1 | Base filter

After applying the base level filters to each caller's SNP set, the two GATK callers, UnifiedGenotyper and HaplotypeCaller, resulted in the greatest number of SNPs called and the highest mismatch rates by site and by genotype (Table 2). While SAMtools, FreeBayes, and VarScan called an order of magnitude fewer SNPs than the GATK callers, they also resulted in mismatch rates 1–2 orders of magnitude lower (Table 2). The strong correlation between the number of SNPs a program called after base filtering and its mismatch rate ($R^2 = 99.4\%$; Figure S1) led us to apply our additional incremental filtering method to better facilitate the comparison among variant callers.

Despite the different variant callers generating SNP sets orders of magnitude different in size, the distributions of parent-offspring genotype mismatches across the sites called were very similar among programs (i.e., heavily right-skewed; Figure S2). UnifiedGenotyper and VarScan, however, produced an overinflation of sites with a 50% genotype mismatch rate, suggesting a higher rate of genotyping error in the parent than seen with the other variant callers (Figure S2).

## 3.2 | Comparing between callers

### 3.2.1 | Sites called

After applying the additional incremental filtering and reducing each SNP set down to approximately $4 \times 10^5$ and $1 \times 10^5$ sites called, HaplotypeCaller, SAMtools, and UnifiedGenotyper were most similar in the specific sites called, sharing approximately 44 and 58% sites, respectively, between all three (Figures S3–S7). Conversely, FreeBayes and VarScan called the most sites unique to a single caller; FreeBayes called 61% unique sites and VarScan called 41% unique sites (Figures S3–S4). We found very little difference when comparing the total sites shared between callers (Figures S3 and S5) and the sites shared with zero genotype mismatches (Figure 1a, S7A). Notably, however, the specific sites each program called that had genotyping mismatches were largely unique to each caller (Figure 1b, S7B).

### 3.2.2 | Mismatch rates

With the additional incremental filtering, the majority of callers dramatically improved in by-site mismatch rates and by-genotype mismatch rates (Figure 2; see also Figure S8; Table S6). UnifiedGenotyper, HaplotypeCaller, and SAMtools all showed striking improvements in mismatch rates with additional incremental filtering until about the $10^{5.5}$ SNP mark, where mismatch rates for all three callers approached relative plateaus. FreeBayes, on the other hand, showed continued improvements in mismatch rates with increased filtering and did not reach a point of diminishing returns. Finally, while VarScan did see some limited improvement in mismatch rates with increased filtering this change was less dramatic than in the other callers. However, note that because VarScan called relatively few SNPs, minimal additional filtering could be applied. After additional incremental filtering to a specific number of sites called, UnifiedGenotyper consistently had the lowest genotype mismatch rate, followed by HaplotypeCaller, VarScan, SAMtools, and FreeBayes (Figure 2a). The number of SNPs called (i.e., the degree of filtering applied) did not appreciably change how the five callers ranked in either mismatch rate metric (Figure 2).

## 3.3 | Comparing within callers

To assess the effects of individual quality metrics on mismatch rates, we explored the effect of varying each metric on the number of genotypes called and the by-genotype mismatch rates for the baseline filter SNP set from each variant caller program (Table 1).

### 3.3.1 | FreeBayes

Increasing stringency in either of DP or GQ resulted in monotonically decreasing genotype mismatch rates in the FreeBayes SNP set (Figure 3a, S9–S10). Similarly, increasing QUAL over lower values decreased the mismatch rate; however, mismatch rates did plateau over very high QUAL values (Figure S11). All three filtering metrics performed well on our data set and would be useful metrics to filter SNPs with increased stringency.

### 3.3.2 | HaplotypeCaller

Filtering by any of QUAL, DP, GQ, or MQ produced monotonic decreases in genotype mismatch rates with the HaplotypeCaller SNP set (Figure 3b, S12–S14, S16). These filtering metrics would be useful for researchers unable to use VQSR and requiring further increased filtering stringency. The metrics FS, MQRankSum, QD, ReadPosRankSum, and SOR did not appreciably improve mismatch rates with further filtering, and in the case of QD slightly increased mismatch rates in our data (Figures S15, S17–S20).
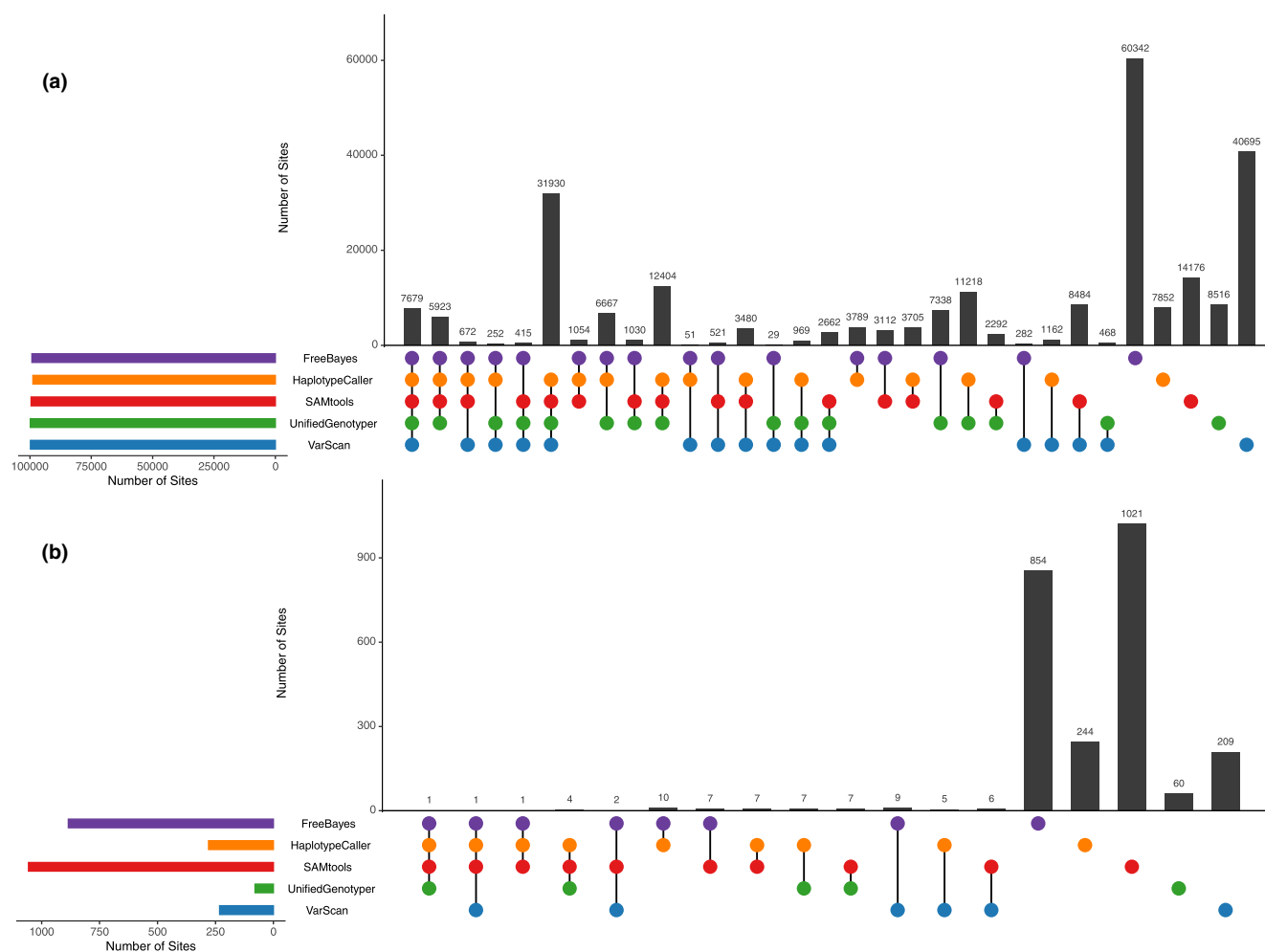
**FIGURE 1** Number of unique variant sites shared between FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan after additional filtering to approximately $1 \times 10^5$ sites called. The number of unique sites with zero genotype mismatches (a) and the number of unique sites with at least one genotype mismatch (b) are shown. The vertical bars show the number of unique sites shared by a combination of variant callers (or single variant caller), and the coloured dots and connecting line below define which combination of variant callers. The horizontal bars at the lower left show the total number of sites without genotype mismatches (a) or with genotype mismatches (b) from each caller. SNP sets were filtered by depth, genotype quality, and quality score where applicable. FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan resulted in 100,041, 99,087, 100,596, 99,871, and 99,983 sites, respectively

### 3.3.3 | SAMtools

Filtering by either QUAL or GQ resulted in monotonic decreases in genotype mismatch rates in the SAMtools SNP set (S21–S22). Increasing DP did improve mismatch rates over lower DP thresholds, although mismatch rates plateaued over higher values (Figure 3c, S23). As with the other callers, these three filtering metrics performed well and would be useful for filtering SNPs with further stringency; however, there may be diminishing returns when filtering by DP with SAMtools.

### 3.3.4 | UnifiedGenotyper

Filtering by any of QUAL, DP, or GQ resulted in decreased genotype mismatch rates with increasing stringency in the UnifiedGenotyper

SNP set (Figure 3d, S24–S26). As with HaplotypeCaller, these three filtering metrics would be useful for researches unable to filter with VQSR and requiring further filtering stringency. None of FS, MQ, MQRankSum, QD, ReadPosRankSum, nor SOR appreciably improved mismatch rates with UnifiedGenotyper, and with our data set filtering by QD or SOR slightly increased mismatch rates (Figures S27–S32).

### 3.3.5 | VarScan

Filtering by either of DP or GQ produced appreciable improvements in genotype mismatch rates in the VarScan SNP set (Figures 3e and S33–S34). While increasing stringency in DP resulted in quite a sharp decrease in mismatch rates over lower DP values, mismatch rates did plateau or increase over higher values (Figures 3e and S33),
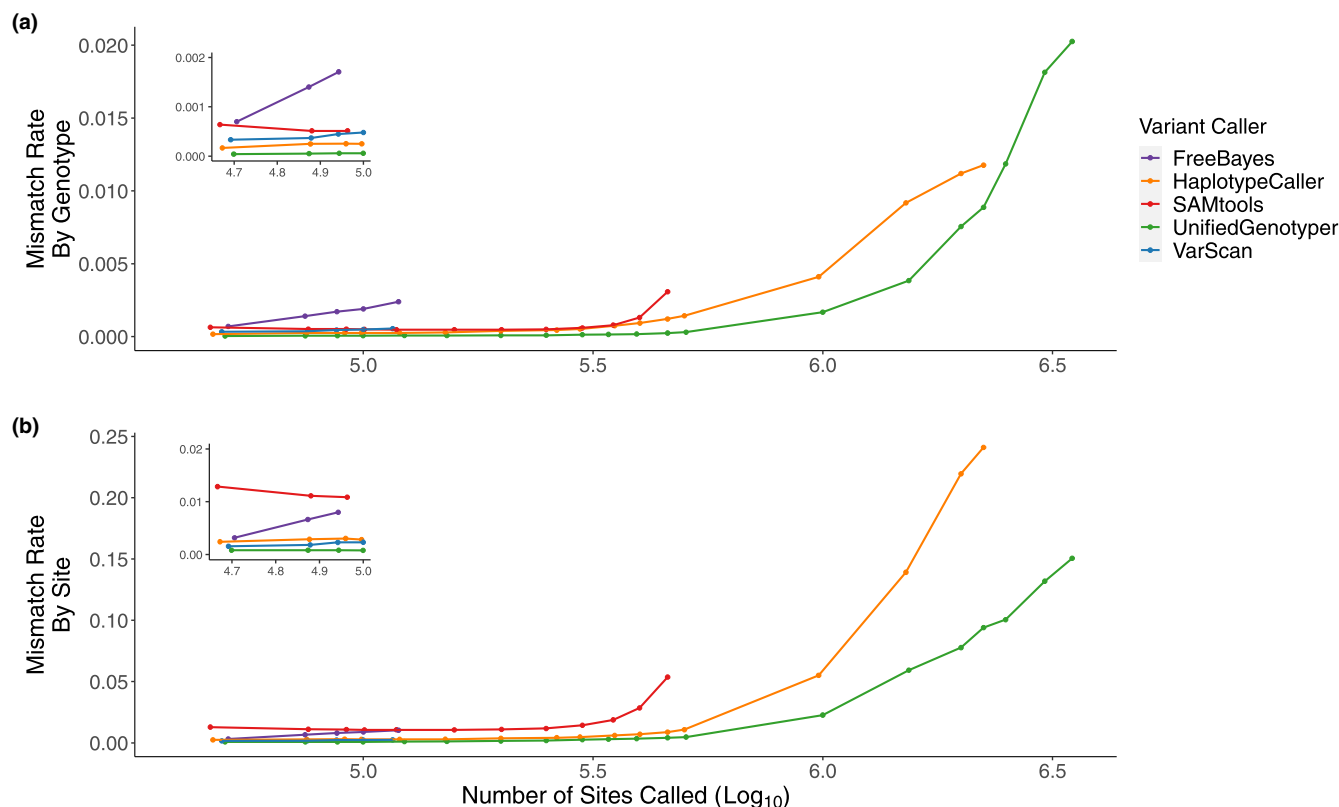
**FIGURE 2** Comparison of mismatch rates by genotype (a) and by site (b) between variant callers at variable numbers of sites called. Mismatch rates were calculated as the proportion of mismatched parent-offspring genotype calls out of the total number of genotypes called (a) and as the proportion of sites with at least one mismatched parent-offspring genotype call out of the total number of sites (b). Inset plots show a magnification of the mismatch rates over 50 to 100 k sites called. Variation in the number of sites called for a particular variant caller was generated by additional incremental filtering to different degrees with depth, genotype quality, and quality score where applicable

somewhat similar to SAMtools (Figures 3c and S23). Rather than suggesting that increasing DP increases error rates in VarScan, however, it is very likely that our mismatch rates are biased over higher values of DP as a consequence of a few erroneously called sites in the parent. For example, at a filtering DP of 80, two sites are responsible for the entirety of the genotype mismatches remaining, one with a parent-offspring mismatch rate of 50% and one with a mismatch rate of 100% (i.e., probably one base miscalled in the parent and both bases miscalled in the parent, respectively). Both DP and GQ appear to be useful for increased filtering with VarScan; however, there could possibly be diminishing returns when filtering by DP with VarScan.

## 4 | DISCUSSION

The reliability of SNP genotypes identified from high-throughput sequencing and the choice of variant caller has been a topic of debate for over a decade (Hwang et al., 2015; Poland & Rife, 2012). Here, we leveraged diploid and haploid sequence data from a single *P. contorta* parent and 106 full-sibling offspring to compare SNP genotyping across five popular variant calling tools: FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan. We used the proportion of mismatches between parent and offspring genotype calls to infer

the genotype error rates of each variant caller, given the rarity of mutation within one generation. Our comparison found large differences in the SNPs called and we evaluated the impact of various filtering metrics on the SNP quality and quantity.

After applying an initial, base level of filtering (Table 1) with each program, we found a large disparity in the number of SNPs called and the error rates between callers (Table 2). As might be expected, we found a strong correlation between the number of SNPs called and the by-genotype error rates ($R^2 = 99.4\%$; Figure S1), which led us to apply our additional incremental filtering to compare callers. For our specific data set, we show that UnifiedGenotyper consistently had the lowest error rates at all degrees of additional filtering stringency (Figure 2). Not only did UnifiedGenotyper have the lowest overall error rates after additional filtering, but it also resulted in the most SNPs called in total, offering ample opportunity to prioritize either the number of SNPs called or the error rates. However, note that UnifiedGenotyper is no longer supported by GATK. After UnifiedGenotyper, both HaplotyperCaller and VarScan performed appreciably well in terms of error rates (Figure 2). VarScan, however, produced quite a low number of SNPs despite initial filtering being relatively lenient (Table 1). After additional filtering, FreeBayes and SAMtools resulted in the highest by-genotype error rate and by-site error rate, respectively (Figure 2).
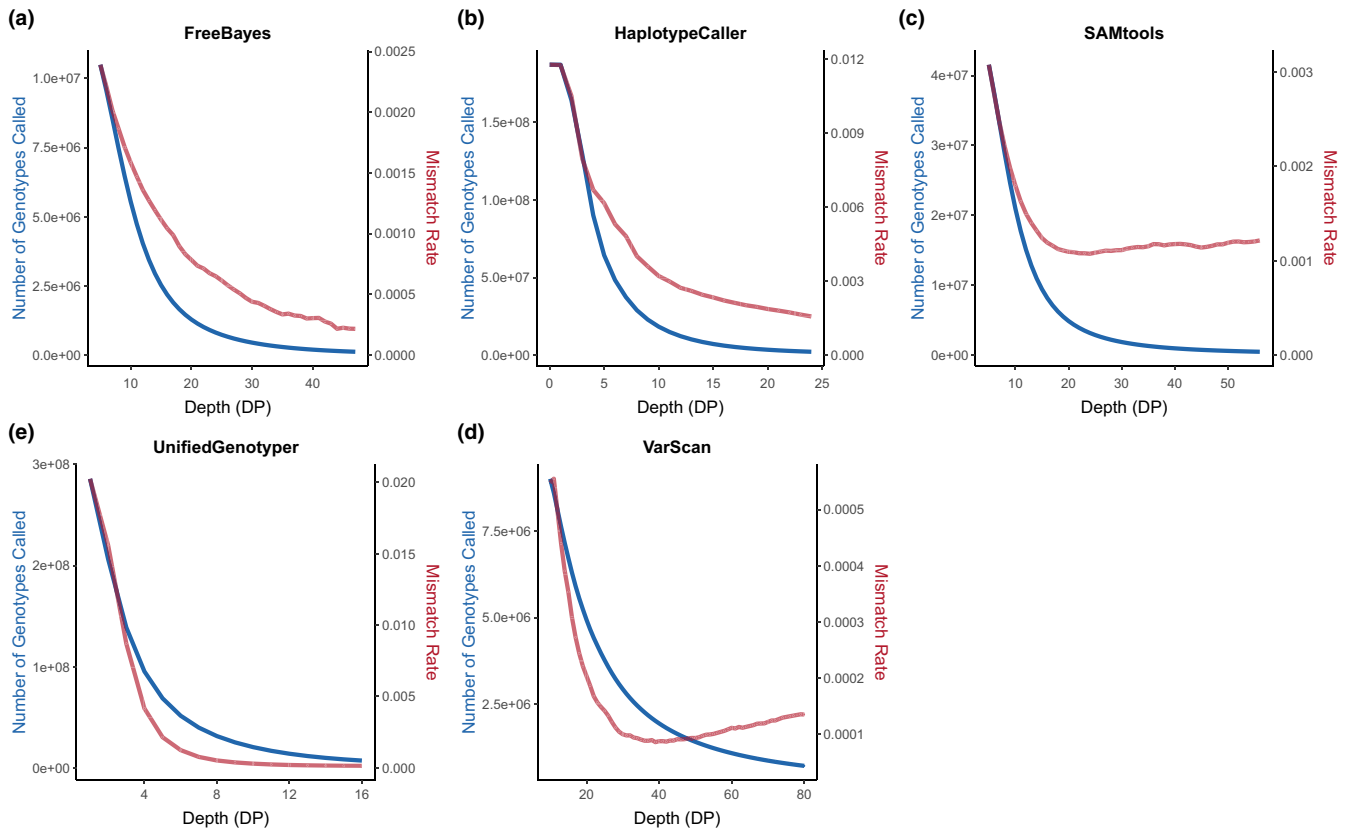
**FIGURE 3** Effect of filtering by depth (DP) on the number of genotypes called (blue) and the genotype mismatch rates (red) after baseline filtering. Scales differ on all three axes for each panel

Our results highlight two different "flavours" of variant callers. Tools such as HaplotypeCaller and UnifiedGenotyper are highly customizable, highly flexible and offer the user the option to prioritize the number of SNPs called or the error rate. However, HaplotypeCaller and UnifiedGenotyper also required extensive additional filtering (Figure S8) beyond baseline to achieve acceptable error rates (Figure 2). The huge number of SNPs these two callers produce, the extent of the filtering required, and the large number of different filter parameters involved, may call for a more experienced user and probably more tinkering to curate a suitable SNP set, especially so in non-model organisms where VQSR is not an option. Other variant callers, such as FreeBayes, SAMtools, and VarScan, achieve decidedly lower error rates when comparing baseline filter SNP sets (Table 2), and probably require much less tinkering and effort to produce an acceptable SNP set. However, these three variant callers may lack the customization and flexibility of the GATK callers, and because they call many fewer SNPs may be at risk of missing some important sites.

When we compared the specific sites produced with each variant caller, we found a large degree of overlap among most programs (3/5) in both the total sites a program called (Figures S3 and S5), and as well in the error-free sites a program called (Figures 1a and S7A). The sites where each caller made genotyping errors, however, were largely specific to the caller used (Figures 1b and S7B), suggesting that the processes by which each caller makes errors differ

mechanistically. Taken together, the high degree of overlap among callers in correctly called sites and the low degree of overlap in erroneously called sites (Figures 1 and S7), suggests that the practice of calling sites with multiple different variant caller tools and using only the SNPs common to all tools may be a highly effective method to improve accuracy. However, while this may reduce error rates, taking the intersection across multiple tools will result in a smaller number of SNPs called.

Across all callers, we found that filtering by QUAL, DP, or GQ gave excellent results in terms of reducing by-genotype error rates in our data set (Figure 3; also see Supporting Information). For those using GATK callers and investigating non-model organisms where VQSR is not applicable, these three metrics may offer the best returns on increasing filtering stringency. While the majority of the GATK filtering metrics did not perform optimally on our data (Figures S15, S17–S20, S27–S32), it should be noted that our results represent a comparison performed on one single data set and that results may vary with other input data. For other variant calling programs, our results suggest that substantial improvements in error rate can be achieved solely through filtering with QUAL, DP, and/ or GQ only.

Our study provides a quantitative assessment of the accuracy of some of the more popular variant caller programs and their commonly used filtering metrics; however, it comes with three main caveats. Our analysis was performed on a linkage mapping population

and as such our results may differ from similar analyses performed on natural populations. For example, a variant caller that is predicated on sample allele frequencies being in Hardy-Weinberg equilibrium will have biased error rates when used on a linkage mapping population where allele frequencies are expected to be 0, 0.5, or 1. As such, it would be valuable to repeat our comparison of variant callers on natural populations in the future. Second, because a *P. contora* reference genome does not yet exist, we aligned our reads against the best available but highly fragmented congeneric loblolly pine (*P. taeda*) reference genome. Any differentiation or lack of synteny between *P. contorta* and *P. taeda*, as well as any assembly errors in the loblolly pine reference genome may have influenced our results. The option to use the reference genome of the correct study species, the quality of that genome, and whether or not the study species is model will all probably have important effects on results. Finally, we chose to include VarScan in our analysis as it is currently a popular variant calling tool, however, VarScan can only call diploid genotypes. Despite these ploidy limitations, our results show that VarScan is still very accurate given haploid input data, with the lowest error rates after baseline filtering out of the variant callers we tested, and comparable error rates after additional incremental filtering. However, because these are diploid genotype calls from haploid input data they still need to be interpreted with some caution. It could prove insightful to repeat a similar analysis in the future, taking into account the genotypes of the pollen donor and diploid offspring to assess variant caller performance given diploid input data.

## CONFLICT OF INTEREST

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

*Pinus contorta* sequence data have been deposited on the Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA) under the BioProject ID: PRJNA764196 and will be released after a 9 month embargo period (Jasper et al., 2021). Code for additional incremental filtering steps was uploaded to GitHub at https://github.com/russjasp/snpcaller_accuracy.

## BENEFIT-SHARING STATEMENT

As a supplement to the data archive, our research group will provide further interpretation and communication regarding the materials and methods discussed here upon request.

## ORCID

*Russ J. Jasper* https://orcid.org/0000-0003-4275-1155
*Tegan Krista McDonald* https://orcid.org/0000-0002-7544-3987
*Pooja Singh* https://orcid.org/0000-0001-6576-400X
*Mengmeng Lu* https://orcid.org/0000-0001-5023-3759
*Clément Rougeux* https://orcid.org/0000-0002-1776-5187
*Brandon M. Lind* https://orcid.org/0000-0002-8560-5417
*Sam Yeaman* https://orcid.org/0000-0002-1706-8699

## REFERENCES

Bernhardsson, C., Vidalis, A., Wang, X., Scofield, D. G., Schiffthaler, B., Baison, J., Street, N. R., García-Gil, M. R., & Ingvarsson, P. K. (2019). An ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway spruce (*Picea abies*). *G3: Genes, Genomes, Genetics, 9*, 1623–1632.

Bian, X., Zhu, B., Wang, M., Hu, Y., Chen, Q., Nguyen, C., Hicks, B., & Meerzaman, D. (2018). Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics, 19*, 1–11. https://doi.org/10.1186/s12859-018-2440-7

Chen, J., Li, X., Zhong, H., Meng, Y., & Du, H. (2019). Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Scientific Reports, 9*, 9345. https://doi.org/10.1038/s41598-019-45835-3

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cornish, A., & Guda, C. (2015). A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed Research International, 2015*, 456479. https://doi.org/10.1155/2015/456479

Danecek, P., Auton, A., Abecasis, G., Albers, C. A, Banks, E., DePristo, M. A, Handsaker, R. E, Lunter, G., Marth, G. T, Sherry, S. T, McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

De La Torre, R. A., Birol, I., Bousquet, J., Ingvarsson, P. K., Jansson, S., Jones, S. J. M., Keeling, C. I., MacKay, J., Nilsson, O., Ritland, K., Street, N., Yanchuk, A., Zerbe, P., & Bohlmann, J. (2014). Insights into conifer giga-genomes. *Plant Physiology, 166*, 1724–1732. https://doi.org/10.1104/pp.114.248708

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., & Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science, 354*, 760–764. https://doi.org/10.1126/science.aag0776

García, C., & Escribano-Ávila, G. (2016). An optimised protocol to isolate high-quality genomic DNA from seed tissues streamlines the workflow to obtain direct estimates of seed dispersal distances in gymnosperms. *Journal of Plant Research, 129*, 559–563. https://doi.org/10.1007/s10265-016-0806-3

Garrison, E., & Marth, G. (2012) *Haplotype-based variant detection from short-read sequencing.* arXiv arXiv:1207.3907.

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports, 5*, 17875. https://doi.org/10.1038/srep17875

Jasper, R. J., McDonald, T. K., Singh, P., Lu, M., Rougeux, C., Lind, B. M., & Yeaman, S. (2021) Data: Evaluating the accuracy of variant calling methods using the frequency of parent-offspring genotype mismatch. *Sequence Read Archive of the National Center for Biotechnology Information*. BioProject ID: PRJNA764196.

Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22, 568–576. https://doi.org/10.1101/gr.129684.111

Lapierre, M., Lambert, A., & Achaz, G. (2017). Accuracy of demographic inferences from the site frequency spectrum: The case of the Yoruba population. *Genetics*, 206, 439–449. https://doi.org/10.1534/genetics.116.192708

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078–2079.

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

Lind, B. M., Lu, M., Obreht Vidakovic, D., Singh, P., Booker, T., Aitken, S., & Yeaman, S. (2022). Haploid, diploid, and pooled exome capture recapitulate features of biology and paralogy in two non-model tree species. *Molecular Ecology Resources*, 22, 1–14. https://doi.org/10.1111/1755-0998.13474

Lu, M., Feau, N., Vidakovic, D. O., Ukrainetz, N., Wong, B., Aitken, S. N., Hamelin, R. C., & Yeaman, S. (2021). Comparative gene expression analysis reveals mechanism of *Pinus contorta* response to the fungal pathogen *Dothistroma septosporum*. *Molecular Plant-Microbe Interactions*, 34, 397–409.

Neale, D. B., Martínez-García, P. J., De La Torre, A. R., Montanari, S., & Wei, X.-X. (2017). Novel insights into tree biology and genome evolution as revealed through genomics. *Annual Review of Plant Biology*, 68, 457–483. https://doi.org/10.1146/annurev-arplant-042916-041049

Neves, L. G., Davis, J. M., Barbazuk, W. B., & Kirst, M. (2014). A high-density gene map of loblolly pine (*Pinus taeda* L.) based on exome sequence capture genotyping. *G3: Genes, Genomes, Genetics*, 4, 29–37.

Pavy, N., Lamothe, M., Pelgas, B., Gagnon, F., Birol, I., Bohlmann, J., Mackay, J., Isabel, N., & Bousquet, J. (2017). A high-resolution reference genetic map positioning 8.8 K genes for the conifer white spruce: Structural genomics implications and correspondence with physical distance. *The Plant Journal*, 90, 189–203.

Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, 5, 92–102. https://doi.org/10.3835/plantgenome2012.05.0005

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Sandmann, S., de Graaf, A. O., Karimi, M., van der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., & Dugas, M. (2019). Evaluating variant calling tools for non-matched next-generation sequencing data. *Scientific Reports*, 7, 43169. https://doi.org/10.1038/srep43169

Scott, A. D., Zimin, A. V., Puiu, D., Workman, R., Britton, M., Zaman, S., Caballero, M., Read, A. C., Bogdanove, A. J., Burns, E., Wegrzyn, J., Timp, W., Salzberg, S. L., & Neale, D. B. (2020). A reference genome sequence for giant sequoia. *G3: Genes, Genomes, Genetics*, 10, 3907–3919. https://doi.org/10.1534/g3.120.401612

Shu, M., & Moran, E. V. (2020). Testing pipelines for genome-wide SNP calling from genotyping-by-sequencing (GBS) data for *Pinus ponderosa*. *Research Square*, 1–21. https://doi.org/10.21203/rs.3.rs-32336/v1

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., Aitken, S. N., & Holliday, J. A. (2016). Exome capture from the spruce and pine giga-genomes. *Molecular Ecology Resources*, 16, 1136–1146. https://doi.org/10.1111/1755-0998.12570

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the cloud*. O'Reilly Media.

Wang, J., Lu, N., Yi, F., & Xiao, Y. (2020). Identification of transposable elements in conifer and their potential application in breeding. *Evolutionary Bioinformatics*, 16, 1–4. https://doi.org/10.1177/1176934320930263

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., Dougherty, W. M., Lin, B. Y., Zieve, J. J., Martínez-García, P. J., & Holt, C. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, 196, 891–909.

Yi, F., Ling, J., Xiao, Y., Zhang, H., Ouyang, F., & Wang, J. (2018). ConTEdb: A comprehensive database of transposable elements in conifers. *Database*, 2018. https://doi.org/10.1093/database/bay131

Zheng, L., Baniaga, A. E., Sessa, E. B., Scascitelli, M., Graham, S. W., Rieseberg, L. H., & Barker, M. S. (2015). Early genome duplications in conifers and other seed plants. *Science Advances*, 1, e1501084. https://doi.org/10.1126/sciadv.1501084

Zimin, A., Stevens, K. A., Crepeau, M. W., Puiu, D., Wegrzyn, J. L., Yorke, J. A., Langley, C. H., Neale, D. B., & Salzberg, S. L. (2017). An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience*, 6, 1–4. https://doi.org/10.1093/gigascience/giw016

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.