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], Mengmeng Lu[1], Clément Rougeux[1], Brandon M. Lind[2], and Sam Yeaman[1]

[1]University of Calgary
[2]The University of British Columbia

February 22, 2024

Abstract

The use of NGS datasets has increased dramatically over the last decade, however, 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 *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 SNP genotyping error rates of FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan. With baseline filtering HaplotypeCaller and UnifiedGenotyper yielded one to two orders of magnitude larger numbers of SNPs and error rates, 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.

Introduction

The decreasing cost and ease of producing short-read next-generation sequencing (NGS) datasets 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 non-trivial. 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. It has become increasingly 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 (Cornish & Guda 2015; Hwang et al. 2015; Bian et al. 2018; Chen et al. 2019; 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). However, these studies are often conducted with human genomic data in mind (Cornish & Guda 2015; Bian et al. 2018; 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 datasets 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 so 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 (De La Torre et al. 2014; Scott et al. 2020), transposable element dynamics (Yi et al. 2018; Scott et al. 2020; Wang et al. 2020), and extensive repeat regions (Wegrzyn et al. 2014). These complexities present a major challenge for NGS data analysis and downstream hypothesis testing in conifers (Shu & Moran 2020; Lind et al. 2021). 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 conifers is that they possess maternally derived haploid megagametophyte tissue that can be reliably excised from embryonic seeds. Megagametophyte tissue can be used in a family design with parent and offspring samples and provide insight into the accuracy of genotype calls by quantifying concordance between haploid offspring genotypes that could arise given the diploid parental genotypes.

Here we used a familial design consisting of diploid tissue from a single lodgepole pine (*Pinus contorta*) parent and maternally derived haploid tissue from 106 full-sibling offspring to evaluate similarities and differences among SNP sets generated from FreeBayes, HaplotypeCaller, Samtools, UnifiedGenotyper, and VarScan. We use 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. We 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.

Methods

Sample Preparation

Our sample data originated from a *Pinus contorta* linkage mapping population consisting of a single parent and its 106 F1 offspring from the interior of British Columbia, Canada, and was provided to us by collaborators at the University of British Columbia (https://coadaptree.forestry.ubc.ca). For each F1 sample, haploid megagametophyte tissue was excised from embryonic seeds for sequencing. Sample preparation, probe design, DNA extraction, and library preparation were performed as described in Lind et al. 2021. DNA samples were sequenced at the Genome Quebec Innovation Centre at McGill University, Montreal, Canada, where they isolated 351 Gbp of ˜150-bp paired-end reads from an Illumina HiSeq4000 instrument.

We used fastp v0.19.5 (Chen et al. 2018) to process and trim sample reads and BWA-MEM v0.7.17 (Li & Durbin 2009) to align them 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 then sorted, indexed, and converted the aligned reads to BAM files with SAMtools v1.9 and processed them with PICARD v2.18.9 (http://broadinstitute.github.io/picard). Where applicable, we converted the BAM files to mpileup files using SAMtools.

We then took these common BAM (or mpileup) files and called SNPs using the following variant caller programs: FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan. After calling SNPs we performed an initial baseline level of filtering consisting of filtering criteria specific to each caller followed 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 v1.4.1106 (R Core Team 2021) to achieve the above-described filtering.

SNP Calling and Baseline Filtering

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. 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).

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).

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).

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).

VarScan

We used VarScan v2.4.4 to call SNPs on the offspring data using a p-value of 0.05 and no additional filters. We then called SNPs in the parental sample in a similar manner and merged the resulting VCF files using VCFtools. SNPs were filtered by removing sites with minor allele frequencies below 0.10%, individual genotype calls with depths below 10, and then applying the common filters. 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 (Table 1).

Comparing variant caller programs

Variant callers differed considerably in the number of sites they called using the baseline filtering (e.g., UnifiedGenotyper yielded ˜3M vs. ˜80k 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; Fig. 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 (Table S1-5). 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$y$ $^{\text{th}}$ to the z $^{\text{th}}$percentile of each of their respective distributions. We calculated two different mismatch rates, a by-genotype rate and a by-site rate, defined respectively as the number of mismatched parent-offspring genotype calls out of the total number of genotypes called and the number of sites with at least one mismatched parent-offspring genotype call out of the total number of sites.

Results

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). SAMtools and FreeBayes called an order of magnitude fewer SNPs than the GATK callers, but they also resulted in mismatch rates an order of magnitude lower (Table 2). Finally, VarScan called the lowest number of SNPs and resulted in the lowest mismatch rates by site and by genotype, all metrics two orders of magnitude lower than the GATK callers (Table 2). The strong correlation between the number of SNPs a program called after base filtering and its mismatch rate ($R^2 = 99.4\%$; Fig. 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; Fig. S2). UnifiedGenotyper, 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 (Fig. S2).

Comparing Between Callers

Sites Called

After applying the additional incremental filtering and reducing each SNP set down to approximately $4x10^5$, $1x10^5$, and $8x10^4$ sites called, HaplotypeCaller, SAMtools, and UnifiedGenotyper were most similar in the specific sites called, sharing at least 44% of sites at each level of filtering (Fig. S3-8). FreeBayes and VarScan called the most sites unique to a single caller; FreeBayes called at least 61% unique sites across these incremental filter levels (Fig. S3-6), and VarScan called 86% unique SNPs in the one comparison it was included in (Fig. S3-4). We found very little difference when comparing the total sites shared between callers (Fig. S3,5,7) and the sites shared with zero genotype mismatches (Fig. 1, S9-10). Notably however, the specific sites each program called that had genotyping mismatches were largely unique to each caller (Fig. 1, S9-10).

Mismatch Rates

With the additional incremental filtering, all callers dramatically improved in by-site mismatch rates and by-genotype mismatch rates, with the exception of VarScan which remained rather insensitive (Table S6; Fig. 2). 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 showed continued improvements in mismatch rates with increased filtering and did not reach a point of diminishing returns. In contrast, VarScan did not substantially improve in mismatch rates with increased filtering, however, 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 (Fig. 2A). The number of SNPs called (i.e., the degree of filtering applied) did not change the ranking of the five callers in terms of mismatch rate by genotype (Fig. 2A).

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).

FreeBayes

Increasing stringency in either of DP or GQ resulted in monotonically decreasing genotype mismatch rates in the FreeBayes SNP set (Fig. 3, S11-12). Similarly, increasing QUAL over lower values decreased the mismatch rate, however, mismatch rates did plateau over very high QUAL values (Fig. S13). All three filtering metrics performed well on our dataset and would be useful metrics to filter SNPs with increased stringency.

HaplotypeCaller

Filtering by any of QUAL, DP, or GQ produced monotonic decreases in genotype mismatch rates with the HaplotypeCaller SNP set (Fig. 3, S14-16). These three filtering metrics would be useful for researchers unable to use VQSR and requiring further increased filtering stringency. The metrics FS, MQ, 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 (Fig. S17-22).

SAMtools

Filtering by either QUAL or GQ resulted in monotonic decreases in genotype mismatch rates in the SAMtools SNP set (Fig. 3, S23-24). Increasing DP did improve mismatch rates over lower DP thresholds, although mismatch rates plateaued over higher values (Fig. 3, S25). 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.

UnifiedGenotyper

Filtering by any of QUAL, DP, or GQ resulted in decreased genotype mismatch rates with increasing stringency in the UnifiedGenotyper SNP set (Fig. 3, S26-28). As with HaplotypeCaller, these three filtering metrics would be useful for researches unable to filter with VQSR and requiring further filtering stringency. Similar to HaplotypeCaller, none of FS, MQ, MQRankSum, QD, ReadPosRankSum, nor SOR appreciably improved mismatch rates, and with our dataset filtering by QD or SOR slightly increased mismatch rates with UnifiedGenotyper (Fig. S29-34).

VarScan

Increasing stringency in DP or increasing stringency in GQ beyond 20 produced appreciable improvements in genotype mismatch rate in the VarScan SNP set (Fig. 3, S35-36). Similar to SAMtools, filtering by DP resulted in mismatch rates plateauing over higher values (Fig. S35). Both DP and GQ appear to be useful

for increased filtering with VarScan, however, there may be diminishing returns when filtering by DP with VarScan.

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 (Poland & Rife 2012; Hwang et al. 2015). 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 finds large differences in the SNPs called and we evaluate 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 ($R^2 = 99.4\%$; Fig. S1), which led us to apply our additional incremental filtering to compare callers. For our specific dataset, we show that UnifiedGenotyper consistently had the lowest error rates at all degrees of additional filtering stringency (Fig. 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 volume 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 (Fig. 2). VarScan, however, produced quite a low volume of SNPs despite initial filtering on minor allele frequency and depth 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 (Fig. 2).

Our results highlight two different flavours of variant callers. Tools like 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 beyond baseline to achieve acceptable error rates, and may require a more experienced user and likely more tinkering to curate a suitable SNP set. Other variant callers like FreeBayes, SAMtools, and VarScan, achieve decidedly lower error rates when comparing baseline filter SNP sets, and likely 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 (60%) in both the total sites a program called (Fig. S3,5,7), and as well in the error-free sites a program called (Fig. 1, S9-10). The sites where each caller made genotyping errors, however, were largely specific to the caller used (Fig. 1, S9-10), 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 (Fig. 1, S9-10), 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 the best results in terms of reducing by-genotype error rates in our dataset (Fig. 3; also see Supp. Mat.). 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, it should be noted that our results represent a comparison performed on one single

dataset 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. Secondly, 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 likely 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. As such, all of our VarScan results are diploid genotype calls from haploid input data and should be interpreted with some caution.

Acknowledgements

We thank Sally Aitken's lab at UBC for providing the *P. contorta*seeds for this study, Dragana Obreht Vidakovic for preparing the sequence capture libraries, Pia Smets and Christine Chourmouzis for technical assistance, the Centre d'expertise et de services Génome Québec for sequencing services, WestGrid and Compute Canada for computational support, and University of Calgary Information Technologies for system support. Grant support for this project was provided by Alberta Innovates (20150252), NSERC Discovery (RGPIN/03950-2017), and Genome Canada, Genome Alberta, and Genome BC (The CoAdapTree project; 241REF). The funding bodies did not have any role in the design of the study, collection, analysis, or interpretation of data in writing the manuscript.

References

Bian, X., B. Zhu, M. Wang, Y. Hu, Q. Chen, C. Nguyen, B. Hicks, and D. Meerzaman 2018 Comparing the performance of selected variant callers using synthetic data and genome segmentation. BMC bioinformatics, 19:429.

Chen, J., X. Li, H. Zhong, Y. Meng, and H. Du 2019 Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Scientific reports, 9:9345.

Chen, S., Y. Zhou, Y. Chen, and J. Gu, 2018 fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890.

Cornish, A., and C. Guda 2015 A comparison of variant calling pipelines using genome in a bottle as a reference. BioMed Research International, 2015:456479.

Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group, 2011 The variant call format and VCFtools. Bioinformatics 27:2156–2158.

De La Torre, R.A., I. Birol, J. Bousquet, P.K. Ingvarsson, S. Jansson, S.J.M. Jones, C.I. Keeling, J. MacKay,

O. Nilsson, K. Ritland, N. Street, A. Yanchuk, P. Zerbe, and J. Bohlmann 2014 Insights into conifer giga-genomes. Plant Physiology 166:1724–1732.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv arXiv:1207.3907.

Hwang, S., E. Kim, I. Lee, and E.M. Marcotte 2015 Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific reports, 5:17875.

[dataset] Jasper R.J., T.K. McDonald, P. Singh, M. Lu, C. Rougeux, B.M. Lind, and S. Yeaman 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., Q. Zhang, D. Larson, D. Shen, M. McLellan, L. Lin, C. Miller, E. Mardis, L. Ding, and R. Wilson 2012 VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 22:568–576.

Li, H., and R. Durbin 2009 Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754–60.

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

Lind, B.M., M. Lu, D. Obreht Vidakovic, P. Singh, T. Booker, S. Aitken and S. Yeaman 2021 Haploid, diploid, and pooled exome capture recapitulate features of biology and paralogy in two non-model tree species. Molecular Ecology Resources 00:1–14.

Neale, D.B., P.J. Martínez-García, A.R. De La Torre, S. Montanari and X.-X. Wei 2017 Novel Insights into Tree Biology and Genome Evolution as Revealed Through Genomics. Annual Review of Plant Biology 68:457–483.

Poland, J.A., and T.W. Rife 2012 Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome 5:92–102.

R Core Team 2021 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sandmann, S., A.O. de Graaf, M. Karimi, B.A. van der Reijedn, E. Hellström-Lindberg, J.H. Jansen and M. Dugas 2019 Evaluating variant calling tools for non-matched next-generation sequencing data. Scientific reports, 7:43169.

Scott, A.D., A.V. Zimin, D. Puiu, R. Workman, M. Britton, S. Zaman, M. Caballero, A.C. Read, A.J. Bogdanove, E. Burns, J. Wegrzyn, W. Timp, S.L. Salzberg and D.B. Neale 2020 A reference genome sequence for giant sequoia. G3: Genes, Genomes, Genetics 10:3907–3919.

Shu, M., and E.V. Moran 2020 Testing pipelines for genome-wide SNP calling from genotyping- by-sequencing (GBS) data for Pinus ponderosa. Research Square 1–21.

Van der Auwera, G.A., and B.D. O'Connor 2020 Genomics in the cloud. O'Reilly Media, USA.

Wang, J., N. Lu, F. Yi, and Y. Xiao 2020 Identification of transposable elements in conifer and their potential application in breeding. Evolutionary Bioinformatics 16:1–4.

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

Yi F., J. Ling, Y. Xiao, H. Zhang, F. Ouyang, and J. Wang 2018 ConTEdb: a comprehensive database of transposable elements in conifers. Database 2018.

Zheng, L., A.E. Baniaga, E.B. Sessa, M. Scascitelli, S.W. Graham, L.H. Rieseberg, and M.S. Barker 2015 Early genome duplications in conifers and other seed plants. Science advances 1:e1501084.

Zimin, A., K.A. Stevens, M.W. Crepeau, D. Puiu, J.L. Wegrzyn, J.A. Yorke, C.H. Langley, D.B. Neale, and S.L. Salzberg 2017 An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience 6:1–4.

Conflict of Interests

The authors declare no conflicts of interest.

Data Accessibility

Pinus contorta sequence data was deposited on the Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA) under the BioProject ID: PRJNA764196 (Jasper et al. 2021). Code for additional incremental filtering steps was uploaded to GitHub at https://github.com/russjasp/snpcaller_-accuracy.

Tables

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 10 depth (DP) Genotype calls with less tha
HaplotypeCaller	Sites with greater than 60 Fisher strand (FS) Sites with less than 40 mapping quality (MQ) Sites with less
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 les
VarScan	Sites with less than 0.10% minor allele frequency (MAF) Genotype calls with less than 10 depth (DP) Heterozygot
Common Filters	Sites not called in both parent and offspring Sites with greater than 50% missingness Multiallelic sites

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.

	Sites	Genotypes	Site mismatch rate	Genotype mismatch rate
FreeBayes	1.19×10^{5}	1.05×10^{7}	1.03×10^{-2}	2.39×10^{-3}
HaplotypeCaller	2.24×10^{6}	1.87×10^{8}	2.41×10^{-1}	1.18×10^{-2}
SAMtools	4.59×10^{5}	4.16×10^{7}	5.36×10^{-2}	3.08×10^{-3}
UnifiedGenotyper	3.49×10^{6}	2.86×10^{8}	1.51×10^{-1}	2.03×10^{-2}
VarScan	8.79×10^{4}	4.41×10^{6}	9.79×10^{-4}	3.97×10^{-4}

Figures

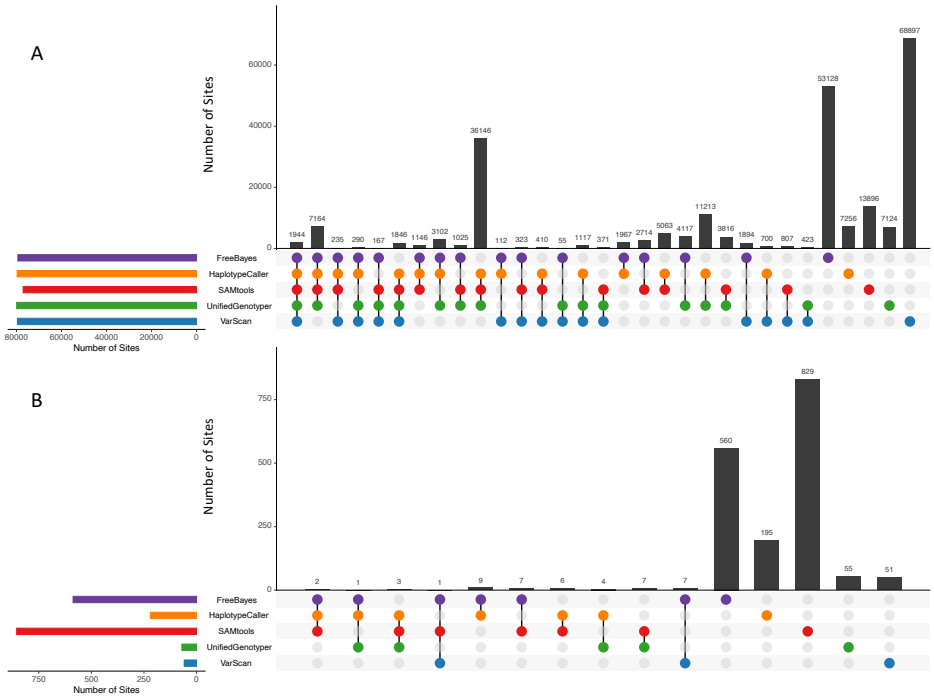

Fig. 1: Number of unique sites shared between FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan after additional filtering to approximately 8x10^4 sites called. The number of unique sites with zero genotype mismatches (A) and the number of unique sites with at least one genotype mismatch (B) are shown. SNP sets were filtered with depth, genotype quality, and quality score where applicable. FreeBayes, HaplotypeCaller, SAMtools, UnifiedGenotyper, and VarScan resulted in 79,970, 79,931, 77,928, 79,990, 79,650 sites respectively.

Hosted file

`image1.emf` available at https://authorea.com/users/439401/articles/540356-evaluating-the-accuracy-of-variant-calling-methods-using-the-frequency-of-parent-offspring-genotype-mismatch

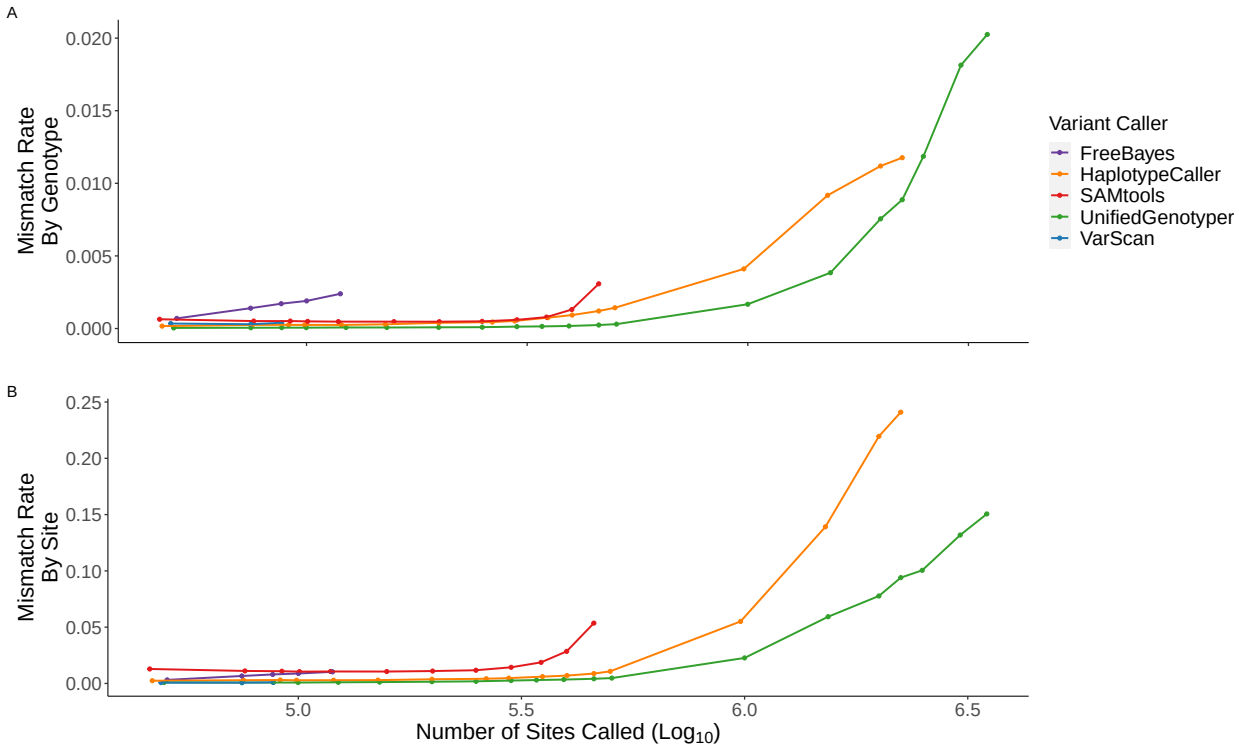

Fig. 2: Comparison of mismatch rates by genotype (A) and by site (B) between SNP callers at variable numbers of sites called. Mismatch rates were calculated as the number of mismatched parent-offspring genotype calls out of the total number of genotypes called (A) and as the number of sites with at least one mismatched parent-offspring genotype call out of the total number of sites (B). Variation in the number of sites called for a particular SNP caller was generated by additional incremental filtering to different degrees with depth, genotype quality, and quality score where applicable.

Hosted file

`image2.emf` available at https://authorea.com/users/439401/articles/540356-evaluating-the-accuracy-of-variant-calling-methods-using-the-frequency-of-parent-offspring-genotype-mismatch

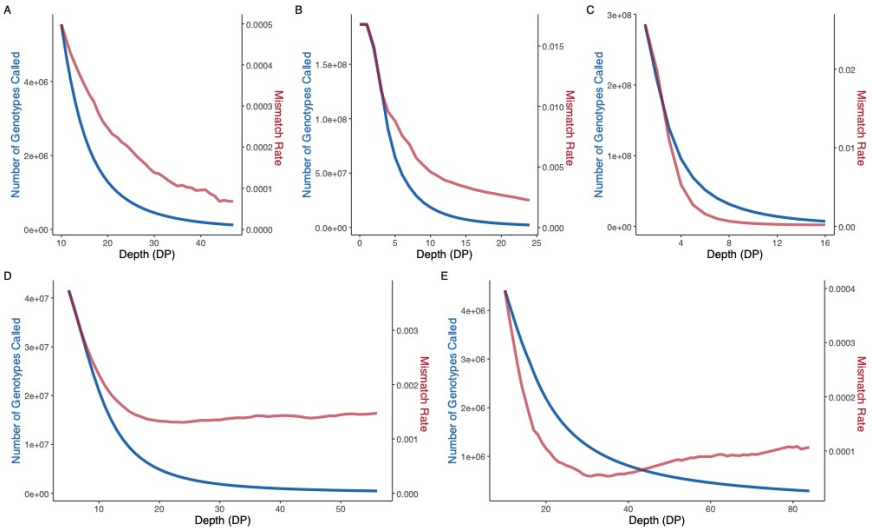

Fig. 3: Effect of filtering by depth (DP) on the number of genotypes called (blue) and the genotype mismatch rates (red) after baseline filtering in FreeBayes, HaplotypeCaller, UnifiedGenotyper, SAMtools, and VarScan (A-E). Note: Scales differ on all three axes for each panel.