Genome complexity, not ploidy, dictates long-read variant-calling accuracy

Authors: Yoshinori Fukasawa[1,2,*]

Affiliation: [1]Center for Bioscience Research and Education, Utsunomiya University, Tochigi, Japan, [2]Graduate School of Regional Development and Creativity, Utsunomiya University, Tochigi, Japan

***Corresponding author**

Abstract

Accurate characterization of genetic variation is fundamental to genomics. While long-read sequencing technologies promise to resolve complex genomic regions and improve variant detection, their application in polyploid and complex genomes remains challenging. Here, we systematically investigate the factors influencing variant calling accuracy using long reads. Using human trio data with known variants to simulate variable ploidy levels (diploid, tetraploid, hexaploid), we demonstrate that while variant sites can often be identified accurately, genotyping accuracy significantly decreases with increasing ploidy due to allelic dosage uncertainty. This highlights a specific challenge in assigning correct allele counts in polyploids even with high depth, separate from the initial variant discovery. We then assessed variant detection performance in genomes with varying complexity: the relatively simple diploid *Fragaria vesca*, the tetraploid *Solanum tuberosum*, and the highly repetitive diploid *Zea mays*. Our results reveal that overall variant calling accuracy correlates more strongly with inherent genome complexity (e.g., repeat content) than with ploidy level alone. Furthermore, we identify a critical mechanism impacting variant discovery: structural variations between the reference and sample genomes, particularly those containing repetitive elements, induce spurious read mapping. This leads to false variant calls, constituting a distinct and more dominant source of error than allelic-dosage uncertainty. Our findings underscore the multifaceted challenges in long-read variant analysis and highlight the need for ploidy-aware genotypers, complexity-informed variant callers, and bias-aware mapping strategies to fully realize the potential of long reads in diverse organisms.

Background

The comprehensive identification and genotyping of genetic variants, including single nucleotide variants (SNVs), small insertions/deletions (indels), and structural variations (SVs), are crucial for understanding genotype-phenotype relationships, population genetics, and evolutionary processes [1]. The advent of long-read sequencing technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), has significantly advanced variant detection capabilities. Long reads can span repetitive regions, resolve complex structural rearrangements, and facilitate haplotype phasing, overcoming limitations inherent to short-read sequencing [2,3].

Third-generation long-read sequencing technologies, namely Pacific Biosciences HiFi (high-fidelity) and Oxford Nanopore duplex reads, have recently transformed variant detection capabilities. PacBio HiFi reads (~15–20 kb in length) now exceed 99.9% per-base accuracy [2], and Nanopore duplex sequencing (reading both DNA strands) achieves a similar level of quality [4]. These improvements allow reads to span and disambiguate many repetitive regions and structural variants, markedly enhancing variant calling sensitivity in regions that were previously “difficult” [5].

Despite these advantages, accurate variant calling in non-model organisms, particularly those with polyploid or highly complex genomes, remains a significant hurdle [6]. Polyploidy, the state of having more than two complete sets of chromosomes, is widespread in plants and also occurs in some animals and fungi. It introduces challenges related to discriminating between homologous, paralogous, and homeologous sequences, and accurately determining allelic dosage (the number of copies of each allele at a heterozygous site) [7]. While long reads can help distinguish haplotypes, accurately genotyping variants remains difficult [8].

Genome complexity, encompassing factors like high repeat content, elevated heterozygosity, segmental duplications, and large repertoires of SVs relative to the reference, further complicates variant analysis irrespective of ploidy [9]. These features can lead to ambiguous read mapping, collapsed assemblies, and erroneous variant calls [10]. While long reads are expected to ameliorate some mapping issues associated with repeats, the interplay between ploidy, inherent genome complexity, and the accuracy of long-read based variant calling requires systematic investigation. Specifically, it is unclear

whether ploidy itself or the associated/independent genome complexity is the primary determinant of performance degradation.

Furthermore, reference genomes often represent only one haplotype or accession, while significant structural variation exists between individuals or accessions within a species [11]. The impact of such reference-sample divergence, particularly SVs containing repetitive elements, on long-read mapping fidelity and subsequent variant discovery (identifying the presence of a variant allele) needs careful characterization. Mis-mapping influenced by SVs could potentially create false variant signals or mask true variants, representing a distinct error source from genotyping uncertainty.

Accurate variant calling in polyploids is notably complicated by issues such as read mapping errors and allelic dosage uncertainty stemming from their complex genomic nature [12]. However, despite the promise of long-read sequencing to potentially mitigate some of these challenges, the specific impact of polyploidy on long-read variant calling accuracy, particularly the precise nature of remaining obstacles and overall performance benchmarks, remains largely uninvestigated.

This study aims to dissect the relative contributions of ploidy, genome complexity, and reference-sample structural divergence to the accuracy of variant calling using long reads. We leverage human trio data with high confidence variant calls as a ground truth to assess the specific impact of allelic dosage uncertainty in simulated polyploid scenarios. We then compare variant calling performance across plant genomes exhibiting different levels of ploidy and complexity. Finally, we investigate the mechanisms by which SVs between reference and sample genomes affect read mapping and variant discovery accuracy.

Results

Reassessing Variant Calling Parameters for Long-read Alignments

Popular variant-calling tools, GATK and Freebayes, have consistently shown stable software performance even when analyzing polyploids using long-read mapping data, and were therefore selected for this study. Although AI-based methods optimized specifically for diploid genotyping using long-read mapping data have been developed and have demonstrated remarkable performance [13–15], these approaches currently remain inapplicable to polyploid genotyping. Optimized parameter settings and filtration methods for PacBio HiFi data have been previously reported for GATK. Therefore, we first evaluated whether these recommended settings would also be applicable to Freebayes, for which

comparable community guidelines currently do not exist for highly accurate long read data. HiFi-GATK removes reads with a mapping quality (MQ) lower than 60, whereas Freebayes probabilistically down-weights low-MQ reads to retain as much information as possible. Therefore, we applied a post-hoc filtering step to Freebayes results by varying MQM thresholds. When low-MQ reads were included as input to GATK, the results were largely unchanged, although a very slight decrease in F1 score was observed. Therefore, we retained the parameter setting for HiFi-GATK. For Freebayes, although no changes were observed at threshold 50, raising MQM to 60 slightly increased precision but considerably reduced recall, resulting in an overall decrease in the F1 score. FreeBayes achieved an overall F1 of 0.943 (precision 0.920, recall 0.967). Increasing the MQM threshold to 60 slightly lowered F1 (Figure 1). We also evaluated whether the variant-type-specific filtration approach adopted by HiFi-GATK could improve the performance of Freebayes. However, this approach resulted in only a slight increase in precision accompanied by a larger reduction in recall (Figure 1). Overall, parameters and filtration methods used in the previous study demonstrated the best performance; thus, we retained the parameters default for Freebayes [16].

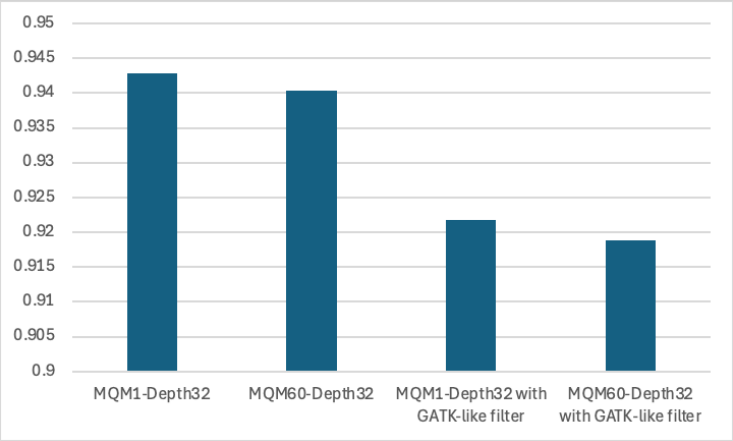

Figure 1 Imposing input constraints yields no appreciable improvement in the FreeBayes model.

We also examined statistics pertaining to the alignments. Upon evaluating alignment quality, we found that the average mapping rate was consistently high (>99%) across all genomes, although F. vesca showed a slightly lower rate at 98%. Interestingly, among mapped reads, the proportion with a mapping quality (MQ) score of 60 was slightly higher in plant genomes (F. vesca: 96.5%, S. tuberosum: 97.0%, Z. mays: 96.5%) compared to the human genome (94.1%). This result contrasts with our expectation that mapping quality

would decrease in more complex plant genomes. Additionally, while the proportion of reads with MQ=0 was low in all species, reads with MQ=1 were slightly more prevalent in human (3%) compared to plant genomes (~1%). The implications of these MQ characteristics on variant calling will be discussed in a later section.

Allelic Dosage Uncertainty Limits Genotyping Accuracy in Polyploids

To distinguish between sequence-based features and allele-dosage uncertainty arising from polyploidy, we performed evaluations using the well-established human trio genome as a standard reference. Following the approach established in a previous study [16], we constructed synthetic polyploid test samples with known truth using the Ashkenazi trio (HG002/HG003/HG004 from Genome in a Bottle) (see Methods). In brief, we merged the parental genomes to simulate an autotetraploid (4x) and combined both parents plus child for a hexaploid (6x) genomes. PacBio HiFi reads for the relevant individuals were pooled and randomly subsampled to generate 10, 30, 50, 70, and 90× coverage for each synthetic polyploid. Variants were then called on these datasets assuming ploidy 2, 4, or 6.

Although performance declines as ploidy increases, ensuring sufficient coverage per haploid genome can partially mitigate this loss (Figure 2). GATK performance was comparable to, and only marginally better than, the Illumina results reported in prior studies [16]. Overall, GATK maintained high genotyping accuracy in diploids and, given adequate coverage, showed only a modest decline even in tetraploids. For FreeBayes, performance in diploids showed little variation, but it deteriorated markedly as ploidy increased (Figure 2). Although sufficient sequencing depth is important, the results suggest additional factors that coverage alone cannot compensate for. Variant-type stratification revealed that FreeBayes's performance decline is driven largely by indels (Figure S1 and S2). Indels are more susceptible than SNVs to platform-specific sequencing biases, which likely accounts for the reduction.

Allele detection exhibited the same pattern. With GATK, reliable detection was achievable at lower sequencing depths, and once a moderate depth was reached, accuracy remained high irrespective of ploidy level (Figure 3). For allele detection, performance for both SNVs and indels reaches a plateau at about 30 × coverage (Figure S3). In tetraploids, however, raising coverage beyond this level for indels correlates with the calling of spurious variants and, intriguingly, a decline in precision (Figure S4).

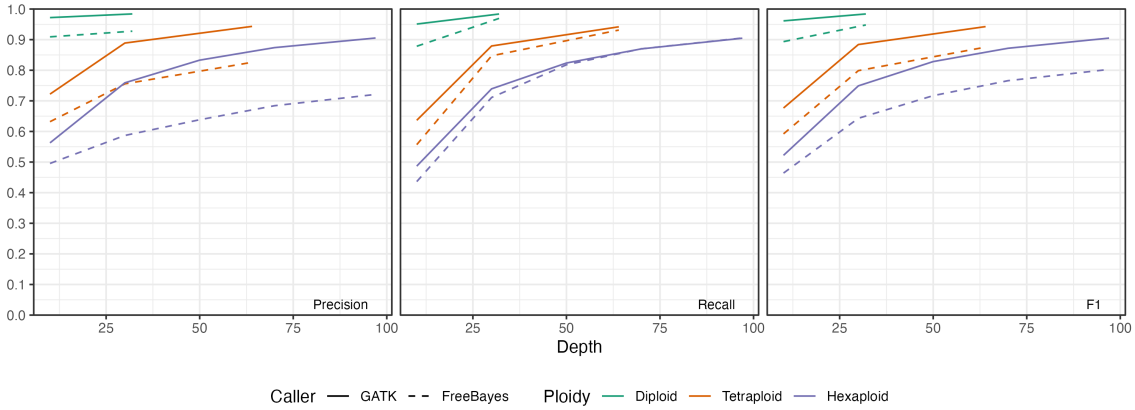

Figure 2 Performance of small variant genotyping on the synthetic human polyploid genome using high-accuracy long reads. Precision, recall, and F1 scores are presented.

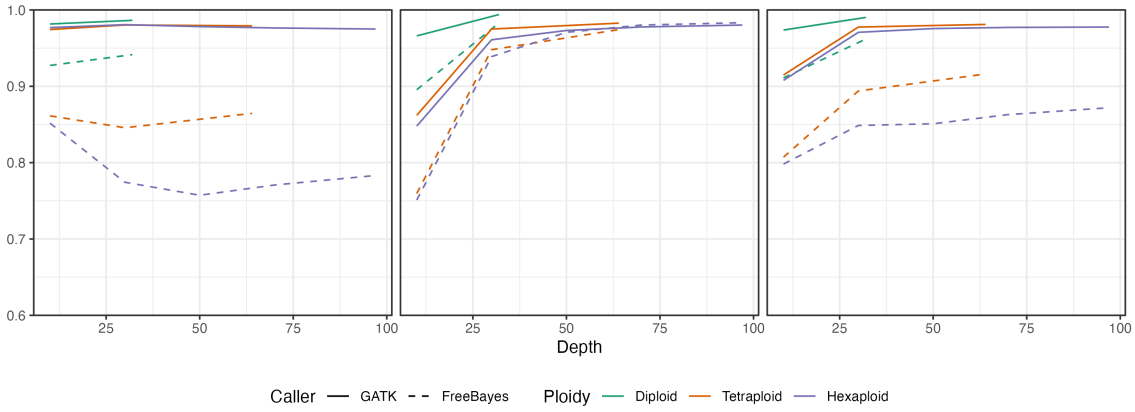

Figure 3 Performance of small variant detection on the synthetic human polyploid genome using high-accuracy long reads. Precision, recall, and F1 scores are presented.

Critically, one of major errors in the polyploid calls was genotyping errors. The variant was often flagged by the caller, but the exact genotype (allelic copy number) was mis-assigned. This aligns with theoretical expectations – as ploidy rises, each haplotype is covered by fewer reads on average, making it harder to distinguish a real low-frequency variant from sequencing error noise [12]. Nonetheless, the fact that precision and recall only dipped a few percentage points by tetraploidy indicates that allelic dosage uncertainty, while present, is manageable with GATK at typical coverage. A further salient issue is the difficulty of genotyping indels, a limitation most pronounced with FreeBayes. By contrast, for SNVs FreeBayes performs on par with, or marginally better than, GATK. SNV detection was reliable with both tools, and GATK proved more robust for indel calls. Encouraged by

evidence that performance remains strong even in highly polyploid genomes, we extended our analyses to genomes with distinct characteristics.

Reference bias and repetitive sequences are the dominant sources of errors

By employing synthetic polyploids derived from human genomes, we confirmed that variant detection performance remains robust and established a baseline for evaluating genotyping accuracy across increasing ploidy levels. Using these results as a reference, we further evaluated performance on genomes with sequence characteristics distinct from those of humans. To test whether the error patterns detected in the synthetic human benchmarks generalize to non-human genomes with contrasting architectures, we next analyzed three plant species chosen to span a gradient of size, ploidy, and repeat content—*Fragaria vesca* (220 Mb, diploid), *Solanum tuberosum* (840 Mb, tetraploid), and *Zea mays* (2.1 Gb, diploid but highly repetitive). These species also offer agronomic relevance and have recently obtained, near-complete reference assemblies, allowing variant calls to be assessed under conditions comparable to the human dataset.

Even with long reads, alignment issues caused by an imperfect reference genome emerged as a primary contributor to false variant calls and missed variants. In the human dataset, which uses a high-quality reference (GRCh38) and where the sample is relatively well represented by that reference, difficult regions were limited. In the case of non-human organisms, a more significant concern likely arises from biases caused by differences between the reference genome and the genome under investigation. Whenever the sample genome contained an insertion or deletion not present in the reference, the aligner could introduce misalignments, causing reads to become gapped or clipped [12]. Consequently, the variant caller might identify spurious single-base differences adjacent to these alignment gaps. This phenomenon is a manifestation of reference bias, where the alignment favors the reference structure of the genome [17]. Given concerns that such reference-bias may pose practical challenges in non-human genomes [12], we proceeded with further validation analyses using the selected genomes.

To define high-confidence variants relative to the reference genome, we obtained long-read sequencing data derived from the same individual used for a recently published, high-quality genome assembly. Although completely eliminating sequencing errors is still challenging, recent genome assemblies have expected error rates below 1%, a level sufficiently low that it does not substantially impact the performance evaluations conducted in this study. We performed chromosome-level comparisons between these highly accurate

genome assemblies and representative reference sequences from the same species. A similar approach has already been adopted for benchmarking human genomes [18,19]; with the recent availability of high-quality genome assemblies, this methodology has become increasingly practical for wide range of organisms. By defining high-confidence genomic regions based on these comparisons, we established variants within these regions as benchmarks for subsequent analyses (Figure 4).

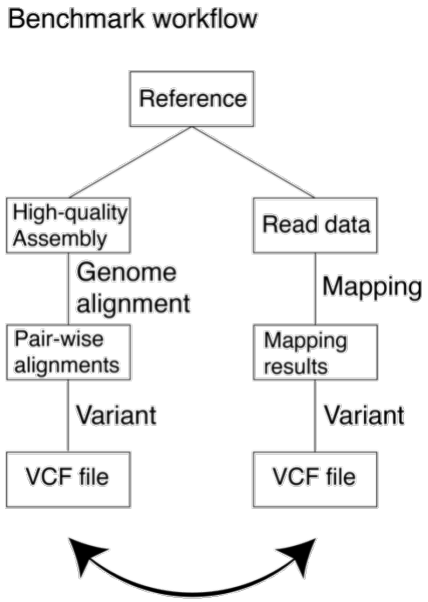

Figure 4 Schematic overview of a semi-automated method for species lacking benchmark data

To test performance in real-world complex genomes, we applied our long-read variant calling pipeline to the three plant genomes of increasing complexity. These genomes span a range from ~220 Mb (strawberry) to ~2.1 Gb (maize) and illustrate different challenges (Table 1). *F. vesca* has 2 sets of 7 chromosomes (2n=14) and a compact genome with relatively low repetitive content [20]. Cultivated potato is 2n=4x=48 (consisting of four homologous sets of 12 chromosomes) with a genome around 840 Mb; it contains a moderate proportion of repeats and high heterozygosity between homeologous chromosomes [21]. Maize, while diploid (2n=20), has a very large genome (~88% transposable elements) and high structural variation between different inbred lines [22,23]. We aligned reads to the respective reference genomes (the *Fragaria vesca* v6.0 reference genome, the *Solanum tuberosum* DM1-3 516 R44 v6.1 reference genome, and the *Zea mays* B73 v5 reference genome). Variant calling was performed with both GATK and FreeBayes (with ploidy set to 2 for strawberry and maize, 4 for potato). We evaluated the

245 resulting variant calls against a set of putative truth variants for each species (see Methods
246 for truth set derivation).
247
248 Table 1 Summary of plant genomes used for benchmarking analyses.

Species (common name)	Ploidy	Assembly version (accession/Database)	Haploid size (Mb)	Estimated repeat content (%)	Reference
Fragaria vesca (woodland strawberry)	2 ×	drFraVesc1 (GCA_964146915,1/GCA_964165485.1)	220	~35	[24]
Solanum tuberosum (potato)	4 ×	Otava v1.0 (Spud DB)	840	~66	[21,25]
Zea mays (maize)	2 ×	Mo17 CAU T2T (GCA_022117705.1)	2,100	~88	[26]

249
250
251 As anticipated, variant calling accuracy showed an inverse correlation with genome
252 complexity. For the simplest genome, *F. vesca*, we obtained accuracy on par with human
253 benchmarks: Overall F1 was lower than that of human, and SNV F1 was ~95% (Figure 5
254 and S5-S7). This suggests that a high-quality diploid plant genome of small size can be
255 surveyed for variants with nearly the same confidence as a human genome using long
256 reads. The errors that did occur in *F. vesca* were mainly in small repetitive sequences and
257 collapsed tandem repeats. In tetraploid potato, performance was clearly reduced relative to
258 *F. vesca* and synthetic-tetraploid human results. SNV precision was ~70% and recall ~85%
259 in the potato sample, indicating a huge drop in both metrics (Figure 5 and S8-S10). Finally,
260 maize showed the greatest challenges. Because of Zea mays' highly complex genome,
261 vcfeval failed to process VCF file containing indels, preventing their assessment. For SNVs,
262 genotyping performance in *Zea mays* fell sharply compared with diploid human and *F.*
263 *vesca*. The decline was driven mainly by lower precision—that is, a large increase in false-
264 positive SNV calls (Figure S11). Under the same analysis pipeline, *Z. mays* produced false-
265 positive calls at rates tens of times higher than those observed for either the human or *F.*
266 *vesca* genomes. The trend was clear – as genome size, repeat content, and structural
267 heterogeneity increased, variant calling accuracy decreased (Figure 5).

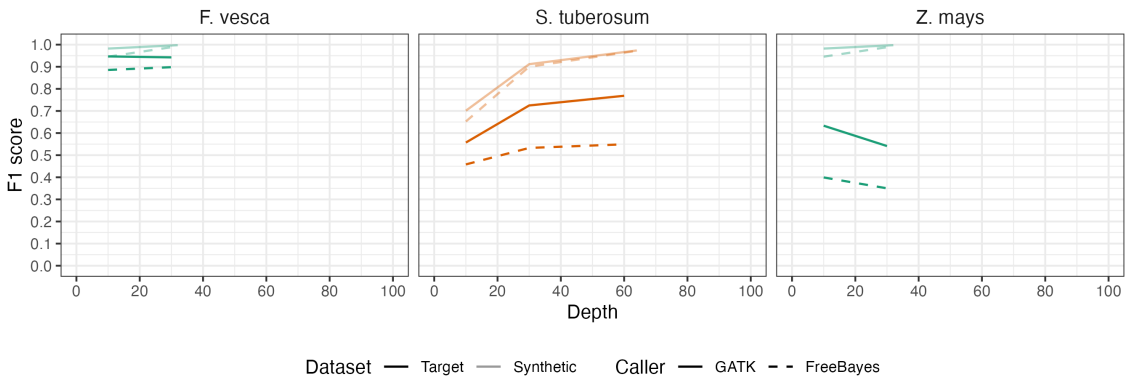

Figure 5 Performance of SNV genotyping (F1 score) in *Fragaria vesca*, *Solanum tuberosum*, and *Zea mays*. The corresponding results for the synthetic human datasets are displayed as a semi-transparent reference trace. F1 scores for each species are presented.

Notably, the *maize* results underscore that a highly repetitive, structurally variable diploid can be even more challenging for variant calling than a less repetitive polyploid even using accurate long reads. The maize line Mo17 (an inbred distinct from B73) carries many structural variants relative to the B73 reference [26]. Decades of maize genomic research have documented that any two maize lines can differ by presence/absence of thousands of genomic segments [27]. Our variant calling analysis reflected this: many false positives in maize were in regions where the sample's reads were erroneously aligned due to sequence present in Mo17 but absent in the B73. In potato, the reference genome is a doubled monoploid line (essentially a haploid-derived assembly) that is divergent from the heterozygous tetraploid cultivar Otava [21]. Therefore, potato also exhibited reference-mismatch issues, though on a smaller scale than maize. In summary, our analyses of plant genomes demonstrate that, with residual read and base calling errors largely mitigated by HiFi and likely ONT duplex sequencing, the principal constraint on variant-calling accuracy is how faithfully the reference genome reflects the sample genome's repetitive and complex architecture.

PAV between reference and sample induces false variant calls

To better understand the errors in the plant genome variant calls, we performed a detailed error analysis, focusing on false positives unique to the long-read calls. A consistent finding was that many false SNV/indel calls were concentrated around unmodeled structural variations between the reference and the sample genome. In the maize dataset, for example, we observed clusters of false SNVs in the vicinity of large presence/absence

variants. One illustrative case was a ~55 kb SV present in the chromosome 1 of Mo17 but absent in B73 reference (Figure 6). In the alignment, reads from the large variant site partially aligned into the reference transposon sequence. The variant caller then reported numerous SNVs and small indels in that region – all of which were artifacts caused by the reads being aligned against sequence that the sample doesn't actually have. Essentially, the alignment constrained the reads to fit the reference, and the variant caller interpreted the resulting mismatches as true variants. This is a classic case of reference bias leading to false positives. Had the transposon absence in the sample been accounted for, those reads would simply not align there (or would be gapped out), and no spurious SNVs would be called. Such alignment artefacts were a leading cause of false variant calls in maize and also contributed in the potato analysis. Although far less common, clusters of these false-positive SNVs were also detected in *F. vesca*.

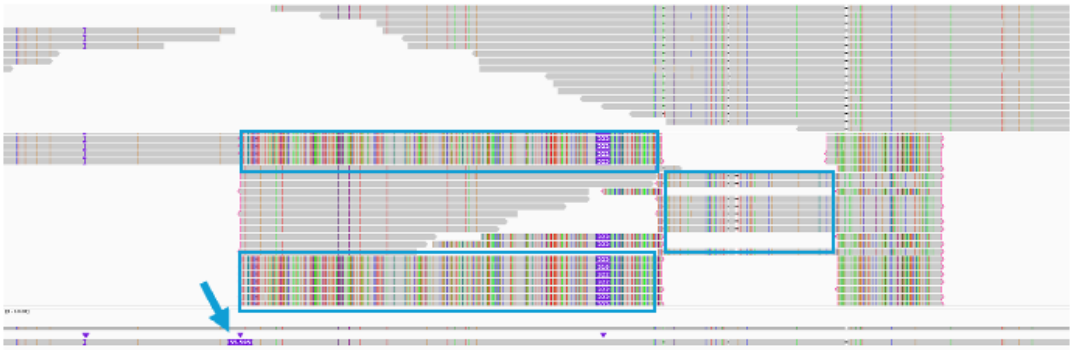

Figure 6 Repeat sequences originating from a large structural variant evade existing filters and become mixed with true alignments. The light-blue arrow marks a 55 kbp insertion in Mo17, and the light-blue box highlights reads derived from this insert.

The repetitive nature of plant genomes exacerbates these issues. Because maize is ~88% transposable elements [22], many of these structural differences occur in repetitive contexts. If a repeat is not present in the reference at a given locus but is present in the sample, reads may align to a homologous repeat elsewhere or in a mis-oriented way, generating spurious variant signals.

To verify this directly, we simulated reads from the Mo17 genome and subjected them to the same analytical workflow. We produced whole-genome data at ~10 and 30× coverage whose length and quality profiles matched those of PacBio HiFi reads. A read was retained if more than 50 % of its bases mapped to regions where sequence homology between Mo17 and B73 could be unambiguously established. Reads that failed to meet this

threshold were discarded, because they were likely to originate predominantly from presence/absence-variation (PAV) segments. Because this threshold is deliberately conservative, the sensitivity is lower than would be expected for a diploid genome; the critical point, however, is that the precision rises to the desired level (Figure 7). Furthermore, because alignments that would otherwise generate false positives have been filtered out, the atypical decline in performance observed at high read depth has been also eliminated.

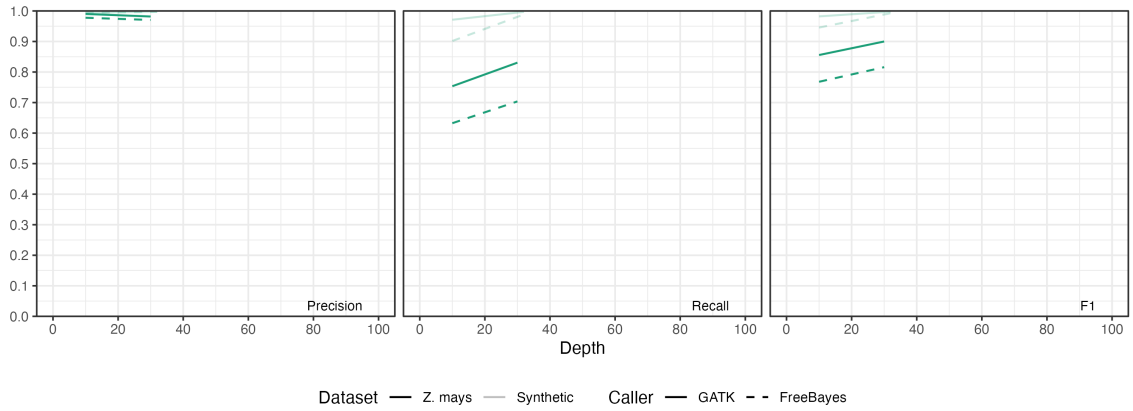

Figure 7 Substantial performance improvement through additional read filtering. Simulated reads originating from regions devoid of PAV achieved the expected precision and improved recall. Precision, recall, and F1 scores are presented.

In *F. vesca*, the closely related reference and lower repeat content meant structural variant-induced errors were smaller. The majority of false positives in *F. vesca* were small indels in homopolymer runs or microsatellites where the exact alignment of a read could shift, a much more localized issue. Taken together, our analyses indicate that structural variants between the reference and sample, especially in repetitive regions, are a major driver of false variant calls. This finding reinforces the importance of high-quality reference genomes that capture population diversity. In crops like maize, where any given accession can have unique sequence content not in the reference (and vice versa) [27], a single reference genome will inevitably cause reference-bias artifacts in variant calling. Long-read sequencing improves read placement compared to short reads, but as our results show, it does not completely solve the reference divergence problem. These issues need to be addressed through improved algorithms or reference representations to avoid spurious variant calls that could mislead downstream analysis.

Discussion

Our study demonstrates both the power and the remaining limitations of long-read sequencing for small variant detection in complex genomes. On the positive side, we show that PacBio HiFi, and likely ONT duplex long reads, enable highly accurate SNV calls even in polyploid genomes. In our human-based simulations, variant detection performance remained accurate from diploid to hexaploid, indicating that modern variant calling algorithms can accommodate the additional allele diversity with minimal loss of sensitivity or precision. This is an encouraging result for researchers working with polyploid species as it suggests that the fundamental task of identifying SNVs does not break down simply because there are more allele copies. Traditional short-read studies had hinted at decent performance in polyploids given sufficient depth [16], but long reads make this even more attainable by providing clearer haplotype information and fewer ambiguities per variant site. We found that genotype (allelic dosage) errors are the main issue in polyploid calling, rather than outright missed variants. This means that for many applications, simply knowing that a variant exists (regardless of exact zygosity) is possible with high confidence. If precise dosage is needed (e.g., distinguishing 1/4 vs 2/4 allele in a tetraploid), one may need higher coverage or specialized algorithms to reduce those errors. Although genotyping accuracy still has room for improvement, the ability to generate haplotype-aware assemblies suggests that phasing-oriented strategies, such as local de novo assembly, would be particularly effective. Nonetheless, the ability to get >95% recall and precision for SNV discovery in a tetraploid like potato using ~30× long-read data is a significant step forward for plant genomics.

The more sobering finding is that genome complexity, particularly the abundance of repeats and structural variation, remains the critical bottleneck of variant calling. Our analyses across strawberry, potato, and maize, albeit to varying degrees, consistently point to reference bias and repetitive sequence alignment problems as the major sources of error, even with accurate long reads. This implies that simply generating longer reads is not a complete solution; the bioinformatics must also evolve to tackle these challenges. The inter-species differences in variant-calling accuracy observed in this study appear to stem greatly from structural divergence from the reference and from species-specific repeat loads. Paradoxically, however, our initial alignment survey showed a slightly higher fraction of MQ 60 reads in the plant genomes than in the human genome. This pattern cannot be explained by overall genome complexity alone and likely reflects a composite effect of reference quality, mapper heuristics, and read characteristics. One possibility is that certain

high-copy repeat families in the human genome create greater mapping ambiguity than the more diverse repeat landscape of the plants. If that interpretation is correct, then our genotyping and simulation results together imply that the mapping-quality metric does not fully penalize the alignment ambiguity introduced by the dispersed repeat elements that pervade plant genomes. This shortcoming could account for the inflated proportion of MQ 60 reads. Although a detailed investigation is beyond the scope of the present work, these findings underscore the need for caution when interpreting mapping-quality statistics. Another implication for future research is the importance of developing or using pangenome references or graph-based alignment approaches. A single linear reference is a bottleneck: if an accession's genome differs significantly (as in maize), the aligner contorts reads to fit the reference, leading to mistakes. Graph genome representations, which incorporate alternate haplotypes, can in principle reduce reference bias by allowing reads to align to the correct version of a sequence present in the population [28,29].

Another area for improvement is the handling of repetitive and multi-copy regions. Our results suggest that variant callers could benefit from explicitly modeling repeat copies or using read-based partitioning to distinguish paralogous loci. For instance, if reads were tagged by haplotype (using algorithms that phase reads, or "haplotagging"), one could potentially separate reads originating from different homeologs in a polyploid and align/call variants on each separately. This approach has been used in diploid assemblies and could be extended to polyploids. In fact, some studies have shown that haplotagging PacBio reads and then running DeepVariant can yield very high accuracy for diploid genomes [15], essentially because the reads are grouped by haplotype. For polyploids, grouping by haplotype (or subgenome) might reduce the misalignment of reads between similar copies. New variant callers such as Octopus (which was designed with polyploid support) are also incorporating statistical models to better distinguish true variant signals from sequencing noise across multiple allele copies (e.g., using probabilistic genotyping models). Continued innovation in these algorithms will be important. Our use of GATK and FreeBayes was partly due to their availability and maturity; however, newer tools and machine learning models could potentially improve genotyping of complex regions if adapted for polyploidy.

It is worth noting that we excluded DeepVariant from our comparisons because it does not natively support ploidies above 2. DeepVariant has exhibited excellent performance in diploid human studies, setting a high benchmark for SNV/indel calling. Extending such deep learning methods to polyploid organisms is a tantalizing prospect. If a model could be trained to call variants in, say, a tetraploid potato, it might better utilize read evidence

across the four alleles and learn to recognize patterns of allele balance that indicate true variants versus errors. At present, the lack of polyploid training data and the complexity of multi-allelic genotype output are barriers. Another future direction is improving indel calling in homopolymer/repeat contexts. We observed that indels remain harder than SNVs to call using GATK/Freebayes that are developed for short reads. Techniques that realign reads locally (as done in the Longshot algorithm for diploids [17]) or incorporate de novo assembly refinement might help resolve complex indel representations.

From an applied perspective, our findings have practical ramifications for crop genomics and breeding. For a crop like potato, our results imply that one can achieve very reliable variant calls across most of the genome using long-read sequencing, which is great for genome-wide association studies or marker discovery in polyploids. Caution is needed in highly duplicated gene families or in regions where the reference might be missing some cultivars' sequence – those will need extra validation. For maize and other large-genome cereals, it's clear that relying on a single reference genome will introduce erroneous variants. A pangenome approach (which the maize community has already begun building) will likely be the way forward; our data provides quantitative evidence of how much improvement is needed. In essence, the accuracy of small variant detection in complex plant genomes is now largely a function of reference quality and completeness. As reference genomes become more representative (e.g., the recent telomere-to-telomere assemblies and pan-genome drafts), variant calling will correspondingly improve.

A critical point is that the problem observed in this study is not automatically remedied by simply adopting a pangenome reference. If a sample genome contains regions that are absent from the pangenome graph, the mapper's scoring scheme can still place reads from those regions on graph nodes that share only local similarity, thereby perpetuating the alignment errors and reference bias we seek to avoid. The pangenome graph must span the widest possible range of genomic diversity. Equally important, mapping and variant-calling algorithms need safeguards that catch the mis-alignments and false calls that can still slip through, even with a well-built pangenome.

In conclusion, long-read sequencing technologies have substantially closed the gap in small-variant detection accuracy between model and non-model genomes. We achieved near-human diploid levels of variant calling performance in a tetraploid and demonstrated the feasibility of SNV/indel discovery in ultra-repetitive genomes like maize. Challenges remain predominantly in the bioinformatic domain: overcoming reference bias and correctly

parsing repetitive sequence signals. Addressing these will require community efforts in assembling diverse genomes, adopting graph/pangenome references, and enhancing sequence analysis algorithms. The outcome will directly benefit plant and animal genomics research, enabling more accurate genotyping for evolutionary studies, crop improvement, and beyond.

Materials and Methods

Data sets and genome references

Human polyploid benchmarking data

We used the well-characterized NA12878 Ashkenazi trio from the Genome in a Bottle (GIAB) project for our human variant calling benchmarks. High-quality variant truth sets are available for the child (HG002) and parents (HG003, HG004) on the GRCh38 reference genome. We obtained PacBio HiFi whole-genome sequencing reads (CCS reads) for these samples from public human pangenome consorthium releases (coverage ~30+× per genome) [2,30]. The GRCh38 reference assembly (with decoy sequences and ALT contigs removed) was used for alignment and evaluation. High-confidence variant calls for HG002, HG003, HG004 (GIAB v4.2.1) were downloaded to serve as ground truth [31].

Plant genomes and sample selection

We selected three plant species to represent a gradient of genome complexity. *Fragaria vesca* (woodland strawberry) is a diploid plant with a small genome (~240 Mb, 7 chromosomes); we used the *F. vesca* reference genome ver. 6 [32] and used a recently released independent assembly as a proxy for the sample under investigation. *Solanum tuberosum* (potato) is an autotetraploid (4n=4x=48) with a ~840 Mb genome; we used the DM1-3 516 R44 v6.1 reference (a doubled monoploid line [33]) and tetraploid variety Otava that has been assembled in a haplotype aware methodology [21]. *Zea mays* (maize) is diploid (2n=20) with an ~2.3 Gb genome rich in repeats (~88% transposable elements); we used the B73 ver. 5 reference genome and selected another inbred line Mo17 as a target genome.

Construction of synthetic polyploids and truth sets

For the human benchmark, we created synthetic polyploid truth sets by leveraging the known high-confidence variants of the GIAB trio. The tetraploid truth set was defined by taking the union of variant sites from HG003 and HG004 (the two parents), restricted to regions where both had confident calls. Similarly, the hexaploid truth set was the union of

HG002, HG003, HG004 variant sites. This approach follows Cooke et al. (2022) who merged GIAB callsets to benchmark polyploid calling [16]. Because the GIAB truth data are highly reliable, we treated this union as ground truth for variant existence. We simulated the sequencing of these polyploids by combining reads: all HiFi reads from HG003 and HG004 were merged into one dataset for the tetraploid, and adding HG002 reads to that for the hexaploid. The read depth per haplotype was thereby roughly equal (e.g., ~15× per haplotype in the tetraploid, summing to 60× total coverage).

For the plant genomes, establishing a truth set is more challenging due to lack of a priori ground truth. We aligned the high-quality assemblies to the reference genome and defined small variants located within the most confidently aligned regions as benchmarks for downstream analyses [18]. We used dipcall ver. 0.3 for the analysis. Because dipcall was originally developed for diploid genomes, we extended its functionality to enable similar analyses for the tetraploid potato genome. *F. vesca* is self-compatible and known for its low heterozygosity. Additionally, a haplotype-resolved genome assembly was recently released by the Tree of Life Project (specific accession details were not provided). We utilized this haplotype-resolved assembly for comparative analysis, using the representative accession, Hawaii-4, as the reference genome [20]. For *S. tuberosum*, we performed a similar analysis by aligning the recently reported haplotype-resolved assembly of the Otava cultivar [21] to the reference genome of doubled monoploid DM1-3 516 R44 [33]. For *Z.mays*, we took advantage of the availability of a reference-quality telomere-to-telomere assembly of the Mo17 inbred [26]. We aligned the Mo17 assembly to the 5th version of B73 reference to derive a comprehensive set of variants. In all cases, we restricted our precision/recall calculations to regions deemed high confidence for truth. For human, this was the GIAB high-confidence regions. These evaluation masks ensure that we are not unfairly penalizing the variant callers for variants that are essentially unresolvable or not represented in our truth data.

HiFi simulation for the *Zea mays* genome.

To simulate HiFi from ccs calling step, subreads were simulated from the *Z. mays* inbred line Mo17 to a depth of 30 × using PBSim3 [34] using quality score model under whole-genome-shotgun settings (mean template length = 1,200 bp; pass-num = 10). Circular consensus sequence (CCS) reads were then generated from the simulated subreads with ccs command in SMRT Link 12.

Alignment and variant calling pipeline

Read alignment

We aligned long reads to their respective reference genomes using minimap2 ver. 2.24 [35]. For PacBio HiFi reads, we used the same parameters applied in pbmm2, which employs sensitive settings for high-accuracy reads. Alignments were output in SAM/BAM format and sorted and indexed with samtools. Distribution of mapping quality was computed using stats subcommand in samtools ver. 1.17.

Variant calling

We evaluated two variant calling tools that support polyploid genotyping: GATK HaplotypeCaller (from GATK ver. 4.1.4) and FreeBayes (v1.3.2) [36,37]. GATK HaplotypeCaller was executed with ploidy set to the appropriate value for each sample (i.e., -ploidy 2 for diploids) and '-pcr-indel-model AGGRESSIVE' and '-minimum-mapping-quality 60' [2]. FreeBayes was run with the --ploidy parameter similarly set, and the other parameters were set to default. For the human trio-derived datasets, we ran variant calling in three scenarios: diploid (ploidy=2), tetraploid (ploidy=4), and hexaploid (ploidy=6) on the synthetic merged read sets. We did not include DeepVariant in our pipeline because it is designed for diploid/haploid cases only and cannot output polyploid genotypes. For reference, we did run DeepVariant on the diploid *F. vesca* and maize datasets as a supplementary experiment, and it achieved similar SNV results to GATK, but we focus on GATK/FreeBayes for consistency across ploidies.

Variant filtering

The raw variant calls were filtered to produce high-confidence call sets for evaluation. We applied hard filters on variant quality metrics. For GATK, we applied the community-recommended per-type QD hard filters: variants were filtered if QD < 2 for SNVs and indels greater than 1 bp, and QD < 5 for 1-bp indels. FreeBayes documentation states that hard input cut-offs seldom improve accuracy and recommends post-hoc site filtering. Pilot tests confirmed that raising the MQM threshold to 60 or applying per-type QD filters, as done in GATK, consistently resulted in reduced F1 scores (Table S1). We therefore applied the same hard-read filters for FreeBayes as previously described [16]. We also did not evaluate any variant calls in regions of low reliability: specifically, we extracted variants with intervals that had reliable alignments between reference and target genome assemblies using minimap2 called by dipcall.

Evaluation metrics and analysis

We calculated precision, recall, and F1-score for variant calls by comparing the caller's VCF to the truth set VCF for each sample. RTG Tools vcfeval was applied for

benchmarking, which account for potential left-shifting or equivalent representations of indels. Genotype concordance (the correctness of allele dosage) was evaluated for the polyploid human samples by comparing the called genotype to the truth genotype. To investigate error causes, we manually inspected alignments and variant calls in IGV for a number of discordant sites, especially clusters of false positives in the plant genomes. We annotated whether false calls coincided with structural variants or repeats.

All analysis code, including custom scripts for merging truth sets and evaluating polyploid calls, is available in a public GitHub repository (to be provided upon publication). By combining controlled experiments on human data with real-world tests on plants, our methods provide a general template for assessing variant calling in non-model genomes, and our pipeline can be readily adapted to other species or sequencing technologies in future work.

Funding

This study was partly supported by JSPS KAKENHI Grant Number 23K19338 to YF and the Utsunomiya University Strawberry Project.

Acknowledgments

Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Contributions

YF designed the research, performed experiments, conducted analyses, interpreted results, and wrote the manuscripts.

Competing interests

The authors declare no competing interests.

References

1. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

2. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.

3. Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021;18:1322–32.

4. Sarashetti P, Lipovac J, Tomas F, Šikić M, Liu J. Evaluating data requirements for high-quality haplotype-resolved genomes for creating robust pangenome references. Genome Biol. 2024;25:312.

5. Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. Hum Genome Var. 2024;11:18.

6. Yao Z, You FM, N'Diaye A, Knox RE, McCartney C, Hiebert CW, et al. Evaluation of variant calling tools for large plant genome re-sequencing. BMC Bioinformatics. 2020;21:360.

7. Garcia AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, et al. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci Rep. 2013;3:3399.

8. Motazedi E, de Ridder D, Finkers R, Baldwin S, Thomson S, Monaghan K, et al. TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. Bioinformatics. 2018;34:3864–72.

9. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.

10. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843–51.

11. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the major structural variant alleles of the human genome. Cell. 2019;176:663-675.e19.

12. Phillips AR. Variant calling in polyploids for population and quantitative genetics. Appl Plant Sci. 2024;12:e11607.

13. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.

14. Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nat Comput Sci. 2022;2:797–803.

15. Abdelwahab O, Torkamaneh D. Artificial intelligence in variant calling: a review. Front Bioinform [Internet]. 2025;5. Available from: http://dx.doi.org/10.3389/fbinf.2025.1574359

16. Cooke DP, Wedge DC, Lunter G. Benchmarking small-variant genotyping in polyploids. Genome Res. 2022;32:403–8.

17. Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nat Commun. 2019;10:4660.

18. Chin C-S, Wagner J, Zeng Q, Garrison E, Garg S, Fungtammasan A, et al. A diploid assembly-based benchmark for variants in the major histocompatibility complex. Nat Commun. 2020;11:4794.

19. Jarvis ED, Formenti G, Rhie A, Guarracino A, Yang C, Wood J, et al. Semi-automated assembly of high-quality diploid human reference genomes. Nature. 2022;611:519–31.

20. Song Y, Peng Y, Liu L, Li G, Zhao X, Wang X, et al. Phased gap-free genome assembly of octoploid cultivated strawberry illustrates the genetic and epigenetic divergence among subgenomes. Hortic Res. 2024;11:uhad252.

21. Sun H, Jiao W-B, Krause K, Campoy JA, Goel M, Folz-Donahue K, et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat Genet. 2022;54:342–8.

22. Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. Genome Biol Evol. 2011;3:219–29.

23. Munasinghe M, Read A, Stitzer MC, Song B, Menard CC, Ma KY, et al. Combined analysis of transposable elements and structural variation in maize genomes reveals genome contraction outpaces expansion. PLoS Genet. 2023;19:e1011086.

24. Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. Proc Natl Acad Sci U S A. 2022;119:e2115642118.

25. Hirsch CD, Hamilton JP, Childs KL, Cepela J, Crisovan E, Vaillancourt B, et al. Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato

659 breeding. Plant Genome. 2014;7:lantgenome2013.12.0042.

660 26. Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, et al. A complete telomere-to-telomere
661 assembly of the maize genome. Nat Genet. 2023;55:1221–31.

662 27. Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, et al. Maize inbreds exhibit high levels of
663 copy number variation (CNV) and presence/absence variation (PAV) in genome content.
664 PLoS Genet. 2009;5:e1000734.

665 28. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation
666 graph toolkit improves read mapping by representing genetic variation in the reference. Nat
667 Biotechnol. 2018;36:875–9.

668 29. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph
669 construction from genome alignments with Minigraph-Cactus. Nat Biotechnol.
670 2024;42:663–73.

671 30. Zook J, Catoe D, McDaniel J, Vang LK, Spies N, Sidow A, et al. Extensive sequencing
672 of seven human genomes to characterize benchmark reference materials. Sci Data.
673 2015;3:1–26.

674 31. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, et al. An open
675 resource for accurately benchmarking small variant and reference calls. Nat Biotechnol.
676 2019;37:561–6.

677 32. Zhou Y, Xiong J, Shu Z, Dong C, Gu T, Sun P, et al. The telomere-to-telomere genome
678 of Fragaria vesca reveals the genomic evolution of Fragaria and the origin of cultivated
679 octoploid strawberry. Hortic Res. 2023;10:uhad027.

680 33. Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, Vaillancourt B, et al. Construction
681 of a chromosome-scale long-read reference genome assembly for potato. Gigascience.
682 2020;9:giaa100.

683 34. Ono Y, Hamada M, Asai K. PBSIM3: a simulator for all types of PacBio and ONT long
684 reads. NAR Genom Bioinform. 2022;4:lqac092.

685 35. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
686 2018;34:3094–100.

687 36. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
688 Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA

689 sequencing data. Genome Res. 2010;20:1297–303.

690 37. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing
691 [Internet]. arXiv [q-bio.GN]. 2012 [cited 2025 May 11]. Available from:
692 http://arxiv.org/abs/1207.3907