

Benchmarking small-variant genotyping in polyploids

Daniel P. Cooke,¹ David C. Wedge,² and Gerton Lunter^{1,3}

¹MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, United Kingdom; ²Manchester Cancer Research Centre, University of Manchester, Manchester M20 4GJ, United Kingdom; ³Department of Epidemiology, University Medical Center Groningen, 9713 GZ Groningen, The Netherlands

Genotyping from sequencing is the basis of emerging strategies in the molecular breeding of polyploid plants. However, compared with the situation for diploids, in which genotyping accuracies are confidently determined with comprehensive benchmarks, polyploids have been neglected; there are no benchmarks measuring genotyping error rates for small variants using real sequencing reads. We previously introduced a variant calling method, Octopus, that accurately calls germline variants in diploids and somatic mutations in tumors. Here, we evaluate Octopus and other popular tools on whole-genome tetraploid and hexaploid data sets created using *in silico* mixtures of diploid Genome in a Bottle (GIAB) samples. We find that genotyping errors are abundant for typical sequencing depths but that Octopus makes 25% fewer errors than other methods on average. We supplement our benchmarks with concordance analysis in real autotriploid banana data sets.

[Supplemental material is available for this article.]

Polyploidy is common in many plant species, including important agricultural crops such as wheat, potato, oat, coffee, rapeseed, cotton, banana, and sugar cane (Song et al. 2012). In mammals, polyploidization regularly occurs during tumorigenesis but has also been shown to be a normal part of development in some mouse and human tissues (Velicky et al. 2018). Molecular markers have been widely used for decades in artificial polyploid crop breeding to assist in selection of more desirable traits such as better resilience to climate change and disease. More recently, genotyping by sequencing has been applied for marker-assisted and genomic selection (He et al. 2014; Kim et al. 2016; Hickey et al. 2019), and the assembly of high-quality plant reference genomes (Jackson et al. 2011; Potato Genome Sequencing Consortium 2011; D'Hont et al. 2012; International Wheat Genome Sequencing Consortium [IWGSC] et al. 2018; Zhuang et al. 2019), together with developments in resequencing, promises new strategies for quantitative trait analysis with a wider variety of genetic variants and better linkage information than is currently possible (Jackson et al. 2011; Uitdewilligen et al. 2013; Bourke et al. 2018; Kyriakidou et al. 2018).

Despite these advances, methods for genotyping polyploids from sequencing data have received little scrutiny in comparison to those for diploids (Bourke et al. 2018; Zook et al. 2019; Li et al. 2018; Krusche et al. 2019). Variant calling and genotyping in polyploids are more difficult than in diploids primarily because the number of possible genotypes at a given loci is combinatorial in the ploidy and number of distinct alleles, and a sequencing read cannot distinguish identical allele copies in the absence of physical linkage with other heterozygous alleles. It therefore becomes harder to determine the allele-specific copy number for a fixed read depth as the ploidy increases. The lower per-allele coverage also makes differentiating true variation from sequencing error less certain. Haplotype-based methods increase power to genotype individual alleles by jointly evaluating combinations of several proximal alleles (haplotypes). They are now standard for diploid calling (Garrison and Marth 2012; Rimmer et al. 2014; Poplin

et al. 2017, 2018; Kim et al. 2018; Cooke et al. 2021) and are becoming more common for somatic mutation calling in tumors (Cooke et al. 2021). Unfortunately, only a minority are capable of polyploid calling (Garrison and Marth 2012; Poplin et al. 2017; Cooke et al. 2021), and none have been rigorously tested for this purpose. Specialized methods for polyploid genotyping have been developed (Blischak et al. 2018; Gerard et al. 2018; Clark et al. 2019) but are only suitable for biallelic SNPs. Furthermore, existing benchmarks of polyploid-calling methods fall short of the standard demanded for diploid calling (Uitdewilligen et al. 2013; Clevenger et al. 2015; Krusche et al. 2019; Yao et al. 2020). In particular, we are not aware of any that consider insertions and deletions (indels), genotyping errors in real sequencing data, or representation differences between callers (Krusche et al. 2019). Polyploid genotyping error rates from sequencing are therefore highly uncertain, undermining developments that depends on them.

We sought to address some of these issues by conducting an in-depth assessment of polyploid small variant calling using an independent and comprehensive ground truth, real sequencing data, and haplotype-aware comparisons.

Results

Synthetic polyploid genomes

We created synthetic tetraploid and hexaploid samples with high-quality truth sets by merging GIAB v4.2.1 (Zook et al. 2019) GRCh38 variants for the human diploid samples HG002, HG003, and HG004. We chose HG003 and HG004 for the tetraploid sample: the two unrelated parents of HG002. Evaluation regions were defined by intersecting (Quinlan and Hall 2010) the GIAB high-confidence regions for each sample, resulting in 2.50-Gb (86% non-N primary reference) confident tetraploid bases containing 5,010,307 variants and 2.48-Gb (85% non-N primary reference) confident hexaploid bases containing 4,951,498 variants. We constructed polyploid Illumina NovaSeq whole-genome test data by mixing reads generated independently for each sample

Corresponding author: daniel.cooke@me.com

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.275579.121>. Freely available online through the *Genome Research* Open Access option.

© 2022 Cooke et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

with consistent PCR-free library preparation and depths (Methods). Each individual sequencing run targeted 35× coverage, resulting in 70× coverage tetraploid samples and 105× coverage hexaploid samples. We confirmed total read counts were similar for each contributing sample, ensuring realistic heterozygous allele frequencies. We then randomly down-sampled the full data sets, starting from 10× in 10× intervals to the full coverage, resulting in 6 + 10 = 16 polyploid data sets. All reads were mapped to GRCh38 with BWA-MEM (Methods) (Li 2013).

Polyploid genotyping accuracy from short-read whole-genome sequencing

We evaluated three popular germline variant callers that support polyploid genotypes—Octopus (Cooke et al. 2021), GATK4 (Poplin et al. 2017), and FreeBayes (Garrison and Marth 2012)—on all synthetic polyploid Illumina data sets, as well as in the diploid HG002 sample to get performance baselines. Other notable germline callers, such as DeepVariant (Poplin et al. 2018), Strelka2 (Kim et al. 2018), and Platypus (Rimmer et al. 2014) were not included because they do not support polyploid calling. We also ignored methods that call polyploid SNVs but not indels, such as polyRAD (Clark et al. 2019). Other than specifying the ploidy and requesting genotype qualities from FreeBayes, we used default setting for all callers (Methods). Octopus calls were hard-filtered with default thresholds; GATK4 and FreeBayes calls were hard-filtered us-

ing recommended thresholds (Methods). Variants were compared using RTG Tools *vcfeval* (Cleary et al. 2015) based on both genotype and allele matches (Methods).

Genotyping accuracy was considerably worse for polyploids compared with diploids (Fig. 1A; Supplemental Table S1). For 30× sequencing depth, on average one of 200 diploid genotype calls were incorrect, in contrast with one of 11 for tetraploid and one of six for hexaploid. Sensitivity was similarly affected; there were 8× and 16× more false negatives on average for tetraploid and hexaploid, respectively, compared with diploid, for 30× sequencing. There were also more substantial differences in accuracy between callers for polyploids compared with diploids. Sensitivity was greater for SNVs than indels, and there was greater disparity between callers for indels (Supplemental Fig. S1; Supplemental Table S2). Overall, Octopus made 26% fewer errors than GATK4 and 30% fewer errors than FreeBayes. However, performance differences varied across depths; the largest F-measure difference between callers occurred at moderate sequencing depth: 30× for tetraploid, 50× for hexaploid. The F-measure showed a typical logarithmic relationship with sequencing depth for both tetraploid and hexaploid samples but also showed a suboptimal response considering ploidy; the F-measure lost from doubling the ploidy was not recovered by doubling the depth, and the differential increased with depth. Stratifying evaluation by GIAB/GA4GH “difficult” regions showed similar results to those found genome-wide (Supplemental Fig. S2; Supplemental Table S1).

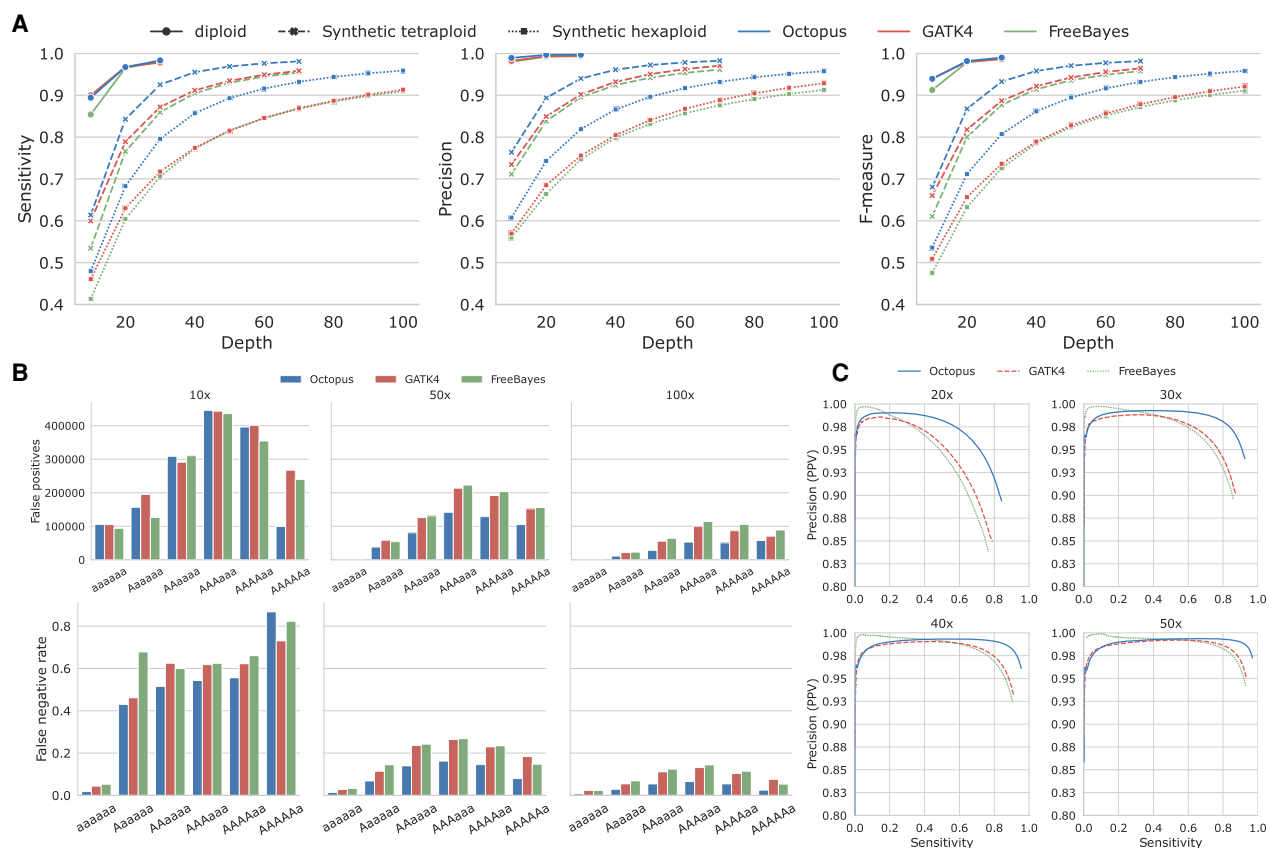


Figure 1. Genotyping accuracy in synthetic polyploids. (A) Sensitivity and precision by depth for each caller on real diploid, as well as synthetic tetraploid and hexaploid Illumina data sets. (B) Counts of false-positive biallelic calls stratified by depth and genotype (*top*). False-negative rates at biallelic sites stratified by depth and genotype (*bottom*). (C) Precision-recall curves for a various tetraploid sequencing depths. Score metrics used to generate the curves were RFGQ (Octopus), GQ (GATK4), and GQ (FreeBayes).

Most false-positive genotype calls resulted from calling the incorrect number of variant alleles (allele-specific copy number error): 89% of false-positive biallelic genotype calls (97% of all false positives) were owing to genotyping errors. The most common false positive for all depths was the balanced heterozygotes: AAaa and AAAaaa (Fig. 1B; Supplemental Fig. S3; Supplemental Table S3), 94% of which were owing to incorrect variant allele copy number. A larger fraction of these was made when the true genotype had a -1 variant allele copy rather than a $+1$ copy (65% vs. 34%) (Supplemental Figs. S4, S5; Supplemental Table S4). The most common biallelic false negatives in tetraploids were simplex heterozygotes (those with a single variant allele copy), whereas for hexaploids, it was duplex heterozygotes (Supplemental Figs. S3, S4). However, normalizing by the true prevalence shows that the most frequent false negative for depths of $30\times$ or more is the balanced heterozygote—the point of maximal variance for binomial distributed allele observations—for both tetraploid and hexaploid; for depths of $20\times$ or more, the most frequent false negative was simplex (Fig. 1B). Furthermore, there was a slight tendency to miscall balanced heterozygotes by a -1 variant allele copy rather than a $+1$ copy for all callers (Supplemental Figs. S4, S5).

Genotype quality scores were generally well calibrated for all callers (Fig. 1C; Supplemental Fig. S6). However, filtering did not always improve the F-measure; the average F-measure percentage change for filtered versus unfiltered calls on all tests was -0.1% , -0.2% , and $+3.5\%$ for Octopus, GATK4, and FreeBayes, respectively. Performance differentials between callers were similar for unfiltered calls (Supplemental Table S1), suggesting that most of Octopus' performance advantage comes from better genotyping rather than filtering.

Comparison based on allele matching showed less performance differential between callers, ploidies, and depths, particularly for precision (Supplemental Fig. S7; Supplemental Table S1). However, predominantly because of better sensitivity at low depths, Octopus still made considerably fewer errors in total than GATK (16% fewer) and FreeBayes (36% fewer).

Longer haplotypes improve genotyping accuracy

A possible explanation for Octopus having better genotyping accuracy than GATK4 and FreeBayes is that Octopus considered longer haplotypes—on average—when calculating genotype likelihoods. If the true set of haplotypes including a subset of variants can be confidently determined, then the variance in the genotype posterior probability distribution is expected to decrease, with respect to allele-specific copy number, for larger subsets (and therefore longer haplotypes) because the number of discriminating reads is expected to be proportional to the haplotype length (Fig. 2). To test this, we recalled genotypes in the $30\times$ tetraploid sample using a parametrization of Octopus designed to generate longer haplotypes than with default settings (Methods). The mean called haplotype length increased from 319 bases to 511 bases and the number of raw false positives decreased by 3976, but the number of raw false negatives increased by 3117.

Banana genotyping

Dwarf Cavendish banana (*Musa acuminata*) is autotriploid, consisting of 11 chromosomes with a haploid genome size of ~ 523 Mb and is an important food source and export-product for many developing countries (D'Hont et al. 2012). To support our previous results on real polyploid samples, we called variants (Methods) in a Dwarf Cavendish banana specimen that was previously whole-genome-sequenced with two Illumina technologies, Next-Seq 500 and HiSeq 1500, to $65\times$ and $55\times$ coverage, respectively (Busche et al. 2020). Both data sets were mapped to the DH Pahang v4 reference (D'Hont et al. 2012; Belser et al. 2021) with BWA-MEM, and genotypes were called with Octopus, GATK4, and FreeBayes.

Because of the lack of truth data, we evaluated concordance on the two banana data sets using haplotype-aware intersections (Methods). Genotypes called by all callers in both data sets, although substantially the largest intersection set, only accounted for 38% of all distinct genotype calls; 22% of calls were unique to a single callset (Fig. 3). However, there were considerable

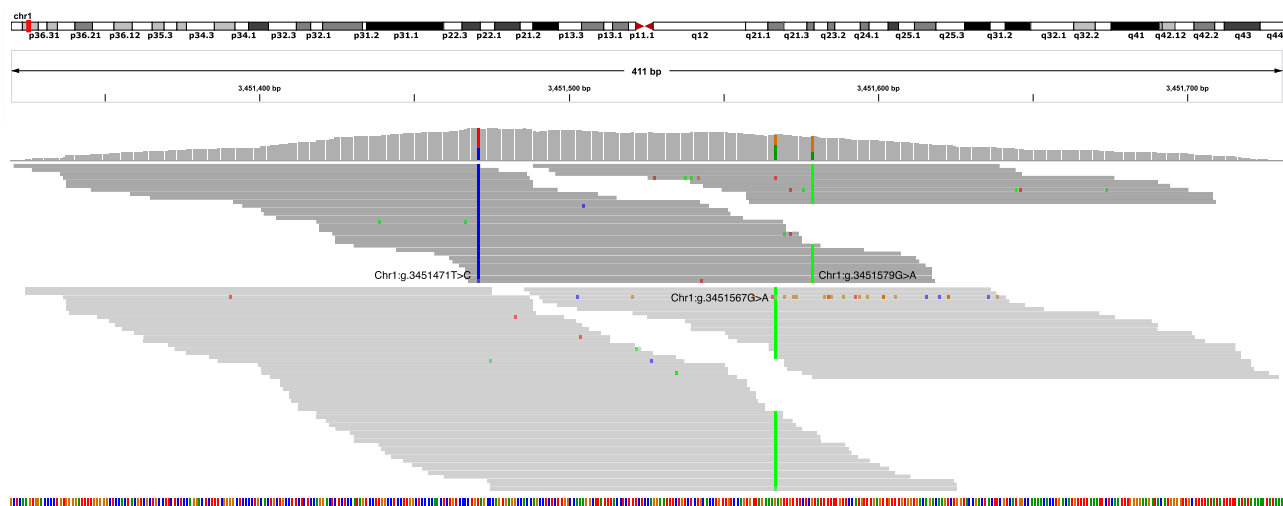


Figure 2. Read pileup of HG003-HG004 tetraploid colored and grouped by supported haplotype. There are two distinct haplotypes (light and dark gray). The true genotypes for the three SNVs (Chr 1:g.3451471T > C, Chr 1:g.3451567G > A, Chr 1:g.3451579G > A) are AAAa, Aaaa, and AAAa. The variant allele read depths are 30/78 (38%), 38/64 (59%), and 20/58 (34%), respectively. GATK4 and FreeBayes both miscall the first two SNVs as AAAa, the most likely genotypes assuming binomially distributed allele observations. Octopus makes the correct calls because it phases all three SNVs, and the first haplotype (including the first and third SNVs) is supported by 74/114 (65%) of reads.

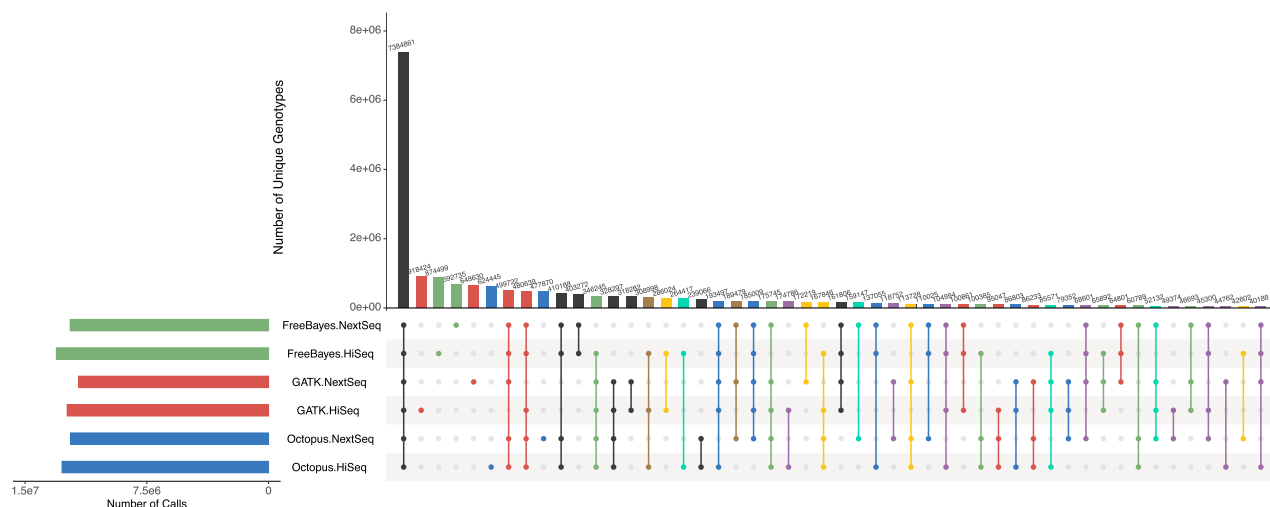


Figure 3. Comparison of genotypes called in two Illumina data sets (HiSeq and NextSeq) of banana specimen by Octopus, GATK4, and FreeBayes. UpSet plot shows callset intersections for each caller–data set pair. The largest 50/63 intersection sets are shown. Intersections are color-coded by caller discordance between the two data sets: no discordances (black), Octopus (blue), GATK4 (red), FreeBayes (green), Octopus & GATK4 (purple), Octopus & FreeBayes (cyan), GATK4 & FreeBayes (yellow), and all (brown). The total number of unique genotype calls was 19,197,247.

differences in concordance between the two data sets for each caller: GATK4 had 31% more discordant calls compared with Octopus and had 11% more than FreeBayes, despite making 2.4% fewer calls overall than FreeBayes and only 1% more than Octopus (Table 1). We also found high discordance when intersecting by alleles (Table 1; Supplemental Fig. S8); only 55% of distinct alleles were present in all callsets, whereas 13% were unique to a single callset, indicating that, in comparison to our results on synthetic data, a larger proportion of false calls arise from incorrect variant alleles rather than genotype errors.

Discussion

We have shown that genotyping is substantially more error-prone in polyploids than in diploids using typical whole-genome sequencing depths, emphasizing that polyploid sequencing studies must be carefully designed to ensure sufficient sequencing depth and caution taken when interpreting polyploid genotype calls. We also found considerable differences in accuracy between callers. Octopus produced fewer than a quarter of the total errors of other methods and half the errors on some data sets. We believe this may be owing to Octopus modeling longer haplotypes, on average, as at loci where variant phasing can be confidently determined, reads assignable to any one allele on a haplotype are implicitly assigned to all other alleles on the same haplotype—even if the read does not overlap them—therefore increasing the effective allele-specific observation count. Regenotyping with lon-

ger haplotypes increased genotyping precision, supporting this hypothesis.

Analysis of real autotriploid banana data sets revealed high discordance between callers, as well as high discordance for all callers on a technical replicate. Although these results were at least consistent with the relative accuracy of callers determined by our benchmarks using synthetic polyploid data (Octopus was the most concordant caller), absolute error rates were evidently higher in real polyploid data. Reasons for this may include greater divergence from the reference genome (Busche et al. 2020); higher levels of repetitive elements in the genome (Jackson et al. 2011); more structural variation (Busche et al. 2020); a less complete reference genome (Jackson et al. 2011); higher rates of sequencing related errors, such as owing to the use of PCR amplification; and bioinformatics algorithms optimized for human data.

We have only considered single-sample polyploid calling in this work; however, multisample calling is important for studying population diversity. Population calling in humans is a difficult problem owing to the computational complexities of joint calling and difficulties in merging independent callsets. Population calling in polyploids will likely be even more challenging and would perhaps benefit from more sophisticated genotype prior models (Blischak et al. 2018).

Moving forward, there is clearly room for improvement in polyploid genotyping from sequencing. The creation of high-quality validation sets with real polyploid samples would be highly valuable in the development of polyploid-calling algorithms,

Table 1. Concordance in two banana Illumina data sets

Match	Caller	Concordant	Discordant	Total	Concordance
GT	FreeBayes	10,390,819	4,516,305	14,907,124	69.70%
GT	GATK4	9,555,918	5,005,372	14,561,290	65.60%
GT	Octopus	10,554,489	3,822,232	14,376,721	73.40%
AL	FreeBayes	11,707,954	3,132,559	14,840,513	79.00%
AL	GATK4	11,114,091	2,445,439	13,559,530	82.00%
AL	Octopus	11,550,003	2,479,828	14,029,831	82.30%

including Octopus. We hope that this work lays the groundwork for future developments.

Methods

Synthetic polyploids with real reads

Raw reads (FASTQ) generated for the PrecisionFDA Truth v2 challenge (Olson et al. 2021) were downloaded from the DNAnexus portal (<https://precision.fda.gov/challenges/10>). Each FASTQ was line-counted to ensure realistic haplotype frequencies, before concatenation of contributing samples to make the full data polyploid data set. Down-sampling was performed directly on the FASTQ files using *seqtk* with the default seed. The sampling fraction was set using *test depth/full depth*, where full depth is $35 \times \text{ploidy}/2$. Reads were mapped with BWA-MEM using default alignment parameters.

Creating polyploid truth genotypes from diploid GIAB samples

Polyploid truth genotypes were generated by concatenating diploid genotypes from GIAB truth VCFs using the BCFtools *merge* and the RTG Tools *vcfilter*, *vcfannotate*, and *vcfsubset* commands. We note that this merge procedure does not resolve variant representation differences between samples. High-confidence BED regions were generated by intersecting GIAB high-confidence BED regions with the BEDTools *multiinter*.

Changes to Octopus for polyploid calling

Although the models that we previously described for Octopus (Cooke et al. 2021) are fully capable of polyploid calls, in practice we found some issues. Runtimes were prohibitive for high ploidies owing to the model always considering every possible genotype for a given set of candidate haplotypes, which is reasonable for diploids but not polyploids. Moreover, sensitivity for simplex variants was not optimal owing to the variant discovery mechanisms not fully accounting for ploidy.

To resolve the runtime issue, we modified the genotype proposal algorithm so that an upper bound on the number of genotypes evaluated can be specified. The algorithm respects this limit by evaluating the full model on the maximum ploidy that results in fewer candidate genotypes than the limit for a given set of haplotypes and then extends a subset of these with greatest posterior probability using each of the candidate haplotypes. The procedure is then applied iteratively, increasing the ploidy by one each iteration until the desired ploidy is reached. We expect this procedure to work well when the number of unique haplotypes present in a region is not substantially greater than the first ploidy considered. We addressed the sensitivity issue by tweaking the pileup and local de novo reassembly candidate variant discovery algorithms to account for the sample ploidy. Octopus source code and documentation are freely available under the MIT license from GitHub (<https://github.com/luntergroup/octopus>).

Variant calling polyploids

For GATK4, we called variants using BAMs with marked duplicates created by GATK4's *MarkDuplicates* tool. Raw BAMs were used for FreeBayes and Octopus. The sample ploidy was specified for all callers: *--organism-ploidy* (Octopus), *--sample-ploidy* (GATK4), and *--ploidy* (FreeBayes). For FreeBayes, we requested genotype qualities with the *-=* option.

Filtering variant calls

For GATK4, we used following filter expressions: `"-filter 'QD < 2.0' --filter-name 'QD2' -filter 'QUAL < 50' --filter-name 'Q50' -filter 'GQ < 5' --filter-name 'GQ5' -filter 'FS > 60.0' --filter-name 'FS60' -filter 'SOR > 3.0' --filter-name 'SOR3' -filter 'MQ < 40.0' --filter-name 'MQ40' -filter 'MQRankSum < -12.5' --filter-name 'MQRankSum-12.5' -filter 'ReadPosRankSum < -8.0' --filter-name 'ReadPosRankSum-8.'" For FreeBayes, we used filter expression "QUAL > 1 & GQ > 1 & SAR > 0 & SAR > 0."`

Genotype and allele comparisons

We used RTG Tools *vcfeval* (v3.12.1) for genotype and allele comparisons, using the *--sample-ploidy* and *--ref-overlap* options. For allele matching, we also used the *--squash-ploidy*, *--XXcom.rtg.vcf.eval.flag-alternates=true*, and *--output-mode="annotate"* options and then determined true and false calls based on the resulting BASE, CALL, BASE_ALTERNATE, and CALL_ALTERNATE annotations.

Identifying genotype errors

Biallelic genotype errors were identified by running RTG Tools *vcfeval* with the *--output-mode="combine"* option and considering biallelic calls with baseline INFO annotations `"BASE=FN_CA"` and `"CALL=FP_CA."`

Long haplotypes with Octopus

To call long haplotypes with Octopus, we provided Octopus with the variant calls it previously produced with default setting as candidates (*--source-candidates*) and disabled de novo variant discovery (*--disable-denovo-variant-discovery*). We also set command line options *--lagging-level=OPTIMISTIC*, *--backtrack-level=AGGRESSIVE*, and *--max-haplotypes=400*.

Banana concordance analysis

Callsets for the banana data sets were intersected using a custom script (<https://github.com/dancooke/starfish>) that invokes both RTG Tools *vcfeval* (that only supports two-way comparisons) and BCFtools to achieve multisample haplotype-aware comparisons. UpSet plots were created with UpSetR (Conway et al. 2017).

Software availability

Custom Snakemake (Köster and Rahmann 2018) and Python code used for data analysis are available from GitHub (<https://github.com/luntergroup/polyploid>) and as Supplemental Code.

Competing interest statement

D.P.C. is a current employee and shareholder of Invitae.

Acknowledgments

We thank Len Trigg at Real Time Genomics for kindly arranging an update to RTG Tools for polyploid genotype support on our request. This work was supported by The Wellcome Trust Genomic Medicine and Statistics PhD Program (grant 203735/Z/16/Z to D.P.C.). The computational aspects of this research were supported by the Wellcome Trust Core Award grant 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Author contributions: D.P.C. formulated and did the analysis and wrote the paper. D.C.W. and G.L. critically reviewed the manuscript and supervised the project.

References

- Belser C, Baurens F-C, Noel B, Martin G, Cruaud C, Istace B, Yahiaoui N, Labadie K, Hřibová E, Doležel J, et al. 2021. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Commun Biol* **4**: 1047. doi:10.1038/s42003-021-02559-3
- Blischak PD, Kubatko LS, Wolfe AD. 2018. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics* **34**: 407–415. doi:10.1093/bioinformatics/btx587
- Bourke PM, Voorrips RE, Visser RGF, Maliepaard C. 2018. Tools for genetic studies in experimental populations of polyploids. *Front Plant Sci* **9**: 513. doi:10.3389/fpls.2018.00513
- Busche M, Pucker B, Viehover P, Weisshaar B, Stracke R. 2020. Genome sequencing of *Musa acuminata* Dwarf Cavendish reveals a duplication of a large segment of chromosome 2. *G3 (Bethesda)* **10**: 37–42. doi:10.1534/g3.119.400847
- Clark LV, Lipka AE, Sacks EJ. 2019. polyRAD: genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3 (Bethesda)* **9**: 663–673. doi:10.1534/g3.118.200913
- Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D, et al. 2015. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. bioRxiv doi:10.1101/023754
- Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA. 2015. Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. *Mol Plant* **8**: 831–846. doi:10.1016/j.molp.2015.02.002
- Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**: 2938–2940. doi:10.1093/bioinformatics/btx364
- Cooke DP, Wedge DC, Lunter G. 2021. A unified haplotype-based method for accurate and comprehensive variant calling. *Nat Biotechnol* **39**: 885–892. doi:10.1038/s41587-021-00861-3
- D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**: 213–217. doi:10.1038/nature11241
- Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN].
- Gerard D, Ferrão LFV, Garcia AAF, Stephens M. 2018. Genotyping polyploids from messy sequencing data. *Genetics* **210**: 789–807. doi:10.1534/genetics.118.301468
- He J, Zhao X, Laroche A, Lu ZX, Liu H, Li Z. 2014. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front Plant Sci* **5**: 484. doi:10.3389/fpls.2014.00484
- Hickey LT, Hafeez AN, Robinson H, Jackson SA, Leal-Bertioli SCM, Tester M, Gao C, Godwin ID, Hayes BJ, Wulff BBH. 2019. Breeding crops to feed 10 billion. *Nat Biotechnol* **37**: 744–754. doi:10.1038/s41587-019-0152-9
- International Wheat Genome Sequencing Consortium (IWGSC), Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, Pozniak CJ, Choulet F, Distelfeld A, et al. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**: 6403. doi:10.1126/science.aar7191
- Jackson SA, Iwata A, Lee SH, Schmutz J, Shoemaker R. 2011. Sequencing crop genomes: approaches and applications. *New Phytol* **191**: 915–925. doi:10.1111/j.1469-8137.2011.03804.x
- Kim C, Guo H, Kong W, Chandnani R, Shuang LS, Paterson AH. 2016. Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci* **242**: 14–22. doi:10.1016/j.plantsci.2015.04.016
- Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, et al. 2018. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**: 591–594. doi:10.1038/s41592-018-0051-x
- Köster J, Rahmann S. 2018. Snakemake: a scalable bioinformatics workflow engine. *Bioinformatics* **34**: 3600. doi:10.1093/bioinformatics/bty350
- Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, et al. 2019. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**: 555–560. doi:10.1038/s41587-019-0054-x
- Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV. 2018. Current strategies of polyploid plant genome sequence assembly. *Front Plant Sci* **9**: 1660. doi:10.3389/fpls.2018.01660
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303-3997 [q-bio.GN].
- Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**: 595–597. doi:10.1038/s41592-018-0054-7
- Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, et al. 2021. precisionFDA truth challenge V2: calling variants from short- and long-reads in difficult-to-map regions. bioRxiv doi:10.1101/2020.11.13.380741
- Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv doi:10.1101/201178
- Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**: 983–987. doi:10.1038/nbt.4235
- Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475**: 189–195. doi:10.1038/nature10158
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. doi:10.1093/bioinformatics/btq033
- Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, WGS500 Consortium, Wilkie AO, McVean G, Lunter G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**: 912–918. doi:10.1038/ng.3036
- Song C, Liu S, Xiao J, He W, Zhou Y, Qin Q, Zhang C, Liu Y. 2012. Polyploid organisms. *Sci China Life Sci* **55**: 301–311. doi:10.1007/s11427-012-4310-2
- Uitendilligen JG, Wolters AM, D'Hoop BB, Borm TJ, Visser RG, van Eck HJ. 2013. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS One* **8**: e62355. doi:10.1371/journal.pone.0062355
- Velicky P, Meinhardt G, Plessl K, Vondra S, Weiss T, Haslinger P, Lendl T, Aumayr K, Mairhofer M, Zhu X, Schütz B, et al. 2018. Genome amplification and cellular senescence are hallmarks of human placenta development. *PLoS Genet* **14**: e1007698. doi:10.1371/journal.pgen.1007698
- Yao Z, You FM, N'Diaye A, Knox RE, McCartney C, Hiebert CW, Pozniak C, Xu W. 2020. Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics* **21**: 360. doi:10.1186/s12859-020-03704-1
- Zhuang W, Chen H, Yang M, Wang J, Pandey MK, Zhang C, Chang W-C, Zhang L, Zhang X, Tang R, et al. 2019. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet* **51**: 865–876. doi:10.1038/s41588-019-0402-2
- Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, et al. 2019. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* **37**: 561–566. doi:10.1038/s41587-019-0074-6

Received March 30, 2021; accepted in revised form December 19, 2021.



Benchmarking small-variant genotyping in polyploids

Daniel P. Cooke, David C. Wedge and Gerton Lunter

Genome Res. published online December 29, 2021
Access the most recent version at doi:[10.1101/gr.275579.121](https://doi.org/10.1101/gr.275579.121)

Supplemental Material <http://genome.cshlp.org/content/suppl/2022/01/18/gr.275579.121.DC1>

P<P Published online December 29, 2021 in advance of the print journal.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Affordable, Accurate
Sequencing.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>
