Benchmarking small-variant genotyping in polyploids

Daniel P Cooke[1,*], David C Wedge[2], and Gerton Lunter[3,1]

[1]MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [2]Manchester Cancer Research Centre, University of Manchester, Manchester, UK. [3]Department of Epidemiology, University Medical Centre Groningen, Groningen, The Netherlands. [*]Correspondence should be addressed to D.P.C (dcooke@well.ox.ac.uk)

Genotyping from sequencing is the basis of emerging strategies in the molecular breeding of polyploid plants. However, compared with the situation for diploids, where 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 methods on whole-genome tetraploid and hexaploid Illumina and PacBio HiFi datasets created using *in silico* mixtures of diploid Genome In a Bottle samples. We find that genotyping errors are abundant for typical sequencing depths, but Octopus makes 25% less errors than other methods on average. We supplement our benchmarks with concordance analysis in real autotriploid banana datasets.

Polyploidy is common in many plant species, including important agricultural crops such as wheat, potato, oat, coffee, rapeseed, cotton, banana, and sugar cane[1]. In mammals, polyploidization regularly occurs during tumorigenesis, but has also been shown to be a normal part of development in some mouse and humans tissues[2]. Molecular markers have been widely used for decades in artificial polyploid crop breeding to assist 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 selection, and the assembly of high-quality plant reference genomes[3–7], together with developments in resequencing, promise new strategies for quantitative trait analysis with a wider variety of genetic variants and better linkage information than is currently possible[7–10].

Despite these advances, methods for genotyping polyploids from sequencing data has received little scrutiny in comparison to those for diploids[10–13]. Variant calling and genotyping in polyploids is 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, but a sequencing read cannot distinguish identical alleles, in isolation. It therefore becomes harder to determine the copy number of a particular allele for a fixed read depth as the ploidy increases. Moreover, since variant allele observations in the reads are expected to occur proportionally to the copy number divided by the sample ploidy, the ability to distinguish sequencing error from true variation diminishes as copy-number decreases and ploidy increases. 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[14–19], and are becoming more common for somatic mutation calling in tumours[14]. Unfortunately, only a minority are capable of polyploid calling[14–16], and none have been rigorously tested for this purpose. Specialised methods for polyploid genotyping have been developed[20–22], but are only suitable for biallelic SNPs. Crucially, existing benchmarks of polyploid calling methods fall short of the standard demanded for diploid calling[9,13,23,24]. In particular, we are not aware of any that consider indels, genotyping errors in real sequencing data, or representation differences between callers[13]. 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. Our analysis is made available online at `https://github.com/luntergroup/polyploid`.

Results

Synthetic polyploid genomes. We created synthetic tetraploid and hexaploid samples with high quality truth sets by merging Genome In A Bottle (GIAB) v4.2 GRCh38 variants for 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 the GIAB high confidence regions for each sample, resulting in 2.50Gb (86% non-N primary reference) confident tetraploid bases containing $5,095,314$ variants, and 2.49Gb (85% non-N primary reference) confident hexaploid bases containing $5,028,566$ variants. We constructed polyploid Illumina NovaSeq and PacBio HiFi whole-genome test data by mixing reads generated independently for each sample with consistent library preparation and depths (Supplementary Note 2). Each individual sequencing run (both Illumina and PacBio) targeted 35x coverage, resulting in 70x coverage tetraploid samples and 105x coverage hexaploid samples. We confirmed total read counts were similar for each contributing sample, ensuring realistic heterozygous allele frequencies. We then randomly downsampled the full datasets, starting

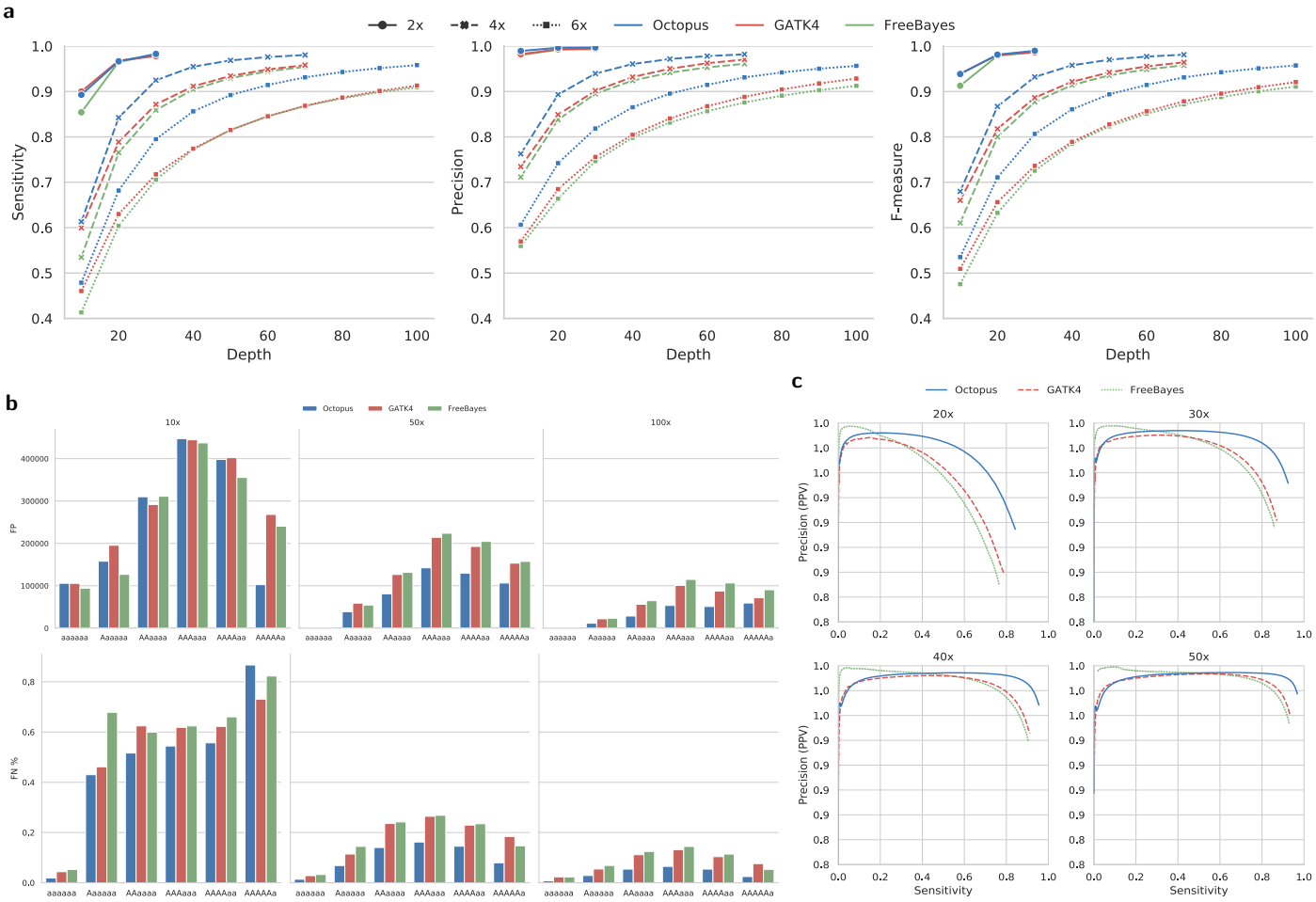

Fig. 1 | Genotyping accuracy in synthetic polyploids. **a** Sensitivity and precision by depth for each caller on real diploid (2x), and synthetic tetraploid (4x) and hexaploid (6x) Illumina datasets. **b** Counts of false positive biallelic genotypes stratified by depth and copy-number (top). Proportion of false negative biallelic genotypes stratified by depth and copy-number (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).

from 10x in 10x intervals to the full coverage, resulting in $2 \times 6 + 2 \times 10 = 32$ polyploid datasets. All reads were mapped to GRCh38, Illumina reads using BWA-MEM and PacBio HiFi reads using pbmm2 (Methods).

Polyploid genotyping accuracy from short-read WGS. We evaluated three popular germline variant callers that support polyploid genotypes: Octopus[14]; GATK4[15]; and FreeBayes[16], on all synthetic polyploid Illumina datasets, and in the diploid HG002 sample to get performance baselines. Other notable germline callers, such as DeepVariant[18], Strelka2[17], and Platypus[19], were not included because they do not support polyploid calling. We also ignored methods that call polyploid SNVs but not indels, such as polyRAD[21]. Other than specifying the ploidy, we used nearly default setting for all callers (Methods). Octopus calls were random forest filtered, GATK4 and FreeBayes calls were hard-filtered using recommended thresholds (Methods). Variants were compared using RTG Tools[25] vcfeval (Methods).

Genotyping accuracy was considerably worse for polyploids compared with diploids (Supplementary Table x). For 30x sequencing depth, on average $1/200$ diploid genotype calls were incorrect, in contrast with $1/11$ for tetraploid and $1/6$ for hexaploid. Sensitivity was similarly affected; there were 8x and 16x more false negatives on average for tetraploid and hexaploid, respectively, compared with diploid, for 30x sequencing. There were also more substantial differences in accuracy between callers for polyploids compared with diploids. Notably, Octopus made 25% less errors than GATK4 and FreeBayes in total, and half the errors on some depths; the largest F-measure difference between callers occurred at moderate sequencing depth: 30x for tetraploid, $50x$ for hexaploid. Accuracy (F-measure) showed a typical logarithmic relationship with sequencing depth for both tetraploid and hexaploid samples, but also 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.

The majority of false positives resulted from incorrect genotype copy-numbers: 87% of false positive biallelic genotype calls (96% of all false positives) were due to incorrect copy number. The most common false positive for all depths were the balanced heterozygotes: AAaa and AAAaaa (Fig. 1b), 92% of which were due to incorrect copy number. A larger fraction of these were made when the true genotype had -1 alternative allele copy rather than $+1$ copy (65% vs 34%; Supplementary Fig. x). The most common biallelic false negatives

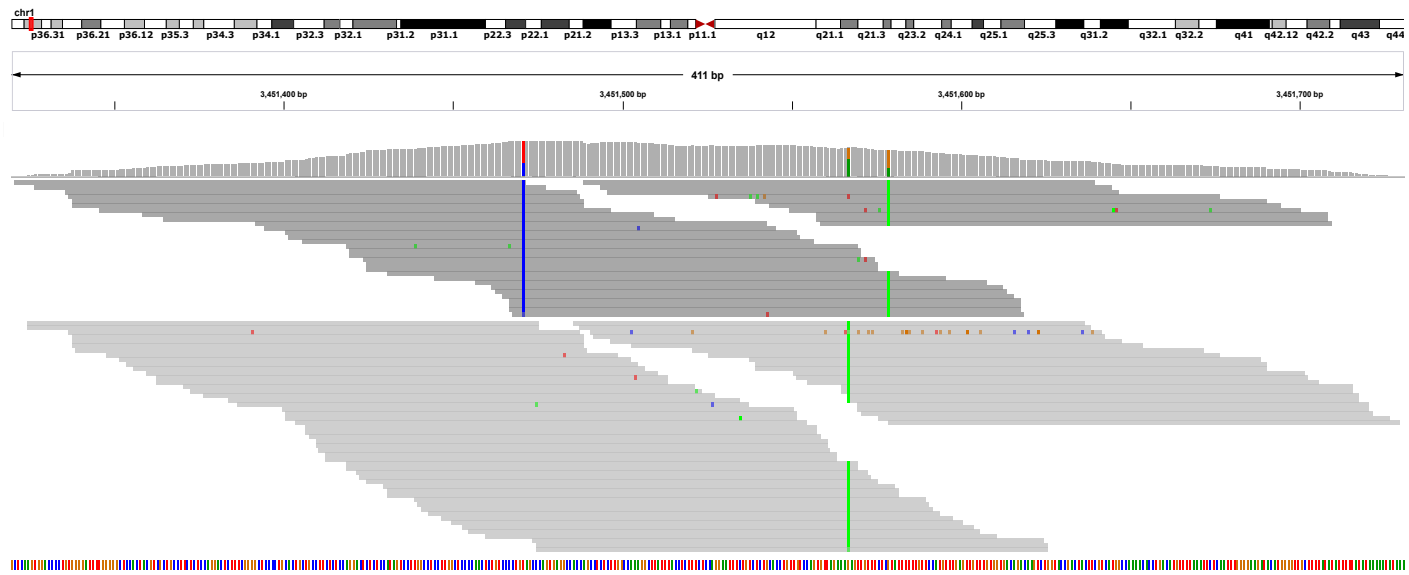

Fig. 2 | Read pileup of HG003-HG004 tetraploid colored and grouped by supported haplotype. The true genotypes for the three SNVs (T>C, G>A, G>A) are AAAa, Aaaa, AAAa. The alternative allele read depths are $30/78$ (38%), $38/64$ (59%), and $20/58$ (34%) respectively. GATK4 and FreeBayes both miscall 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.

in tetraploids were heterozygotes with a single variant copy (simplex), while for hexaploids it was heterozygotes with two variant copies (Supplementary Fig. x). However, normalising by the truth prevalence shows that the most frequent false negative for depths ≥ 30x is the balanced heterozygote for both tetraploid and hexaploid; for depths ≤ 20x the most frequent false negative was simplex (Fig. 1b).

Genotype quality scores were generally well calibrated for all callers (Fig. 1c and Supplementary Fig. x). The average F-measure percentage change for filtered verses unfiltered calls on all tests was -0.1%, -0.2%, and $+3.5\%$, for Octopus, GATK4, and FreeBayes, respectively. Notably, filtering did not always improve F-measure; only Octopus improved F-measure on all tests with filtering, highlighting the advantage of machine-learning based filters compared to hard filters. However, performance differentials between callers were similar for unfiltered calls (Supplementary Table x), showing that most of Octopus's performance advantage come from better genotyping rather than filtering.

Longer haplotypes improve genotyping accuracy. A plausible 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, in particularly with respect to copy number, with larger subsets (and therefore longer haplotypes) since the number of discriminating reads is expected to be proportional to the haplotype length (Fig. 2). To test this, we recalled genotypes in the 30x tetraploid sample using a parametrisation of Octopus designed to generate longer haplotypes than with default settings (Methods). The mean called haplotype length increased from x bases to y bases and the F-measure increased by $x\%$ compared with default settings.

Banana genotyping. Dwarf Cavendish banana (Musa acuminata) is autotriploid consisting of 11 chromosomes with a haploid genome size of around 523Mb, and is an important food source and export-product for many developing countries[4]. 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, NextSeq-500 and HiSeq-1500, to 65x and 55x coverage, respectively[26]. Both datasets were mapped to the DH Pahang v2 reference[4], and genotypes were called with Octopus, GATK4, and FreeBayes. Due to lack of truth data, we relied on less rigorous means to access the quality of the callsets.

We evaluated caller concordance on the two banana datasets using haplotype-aware intersections (Methods). Genotypes called by all callers in both datasets, while substantially the largest intersection set, only accounted for 40% of all distinct genotype calls; 20% of calls were unique to a single callset (Fig. 3). However, there were considerable differences in concordance between the two datasets for each caller: GATK4 had 37% more discordant calls compared with Octopus and 16% more than FreeBayes, despite making 0.5% less calls overall than FreeBayes and only 3% more than Octopus (Table 1). We also found high disconcordance when intersecting by called alleles (Supplementary Fig. x and Supplementary Table x); 58% of distinct alleles were present in all callsets while 11% were unique to a single callset, indicating that that, in comparison to our results on synthetic data, a larger proportion of false calls arise from incorrect variant alleles rather than copy-number errors.

Manual review of discordant calls indicated that a large frac-

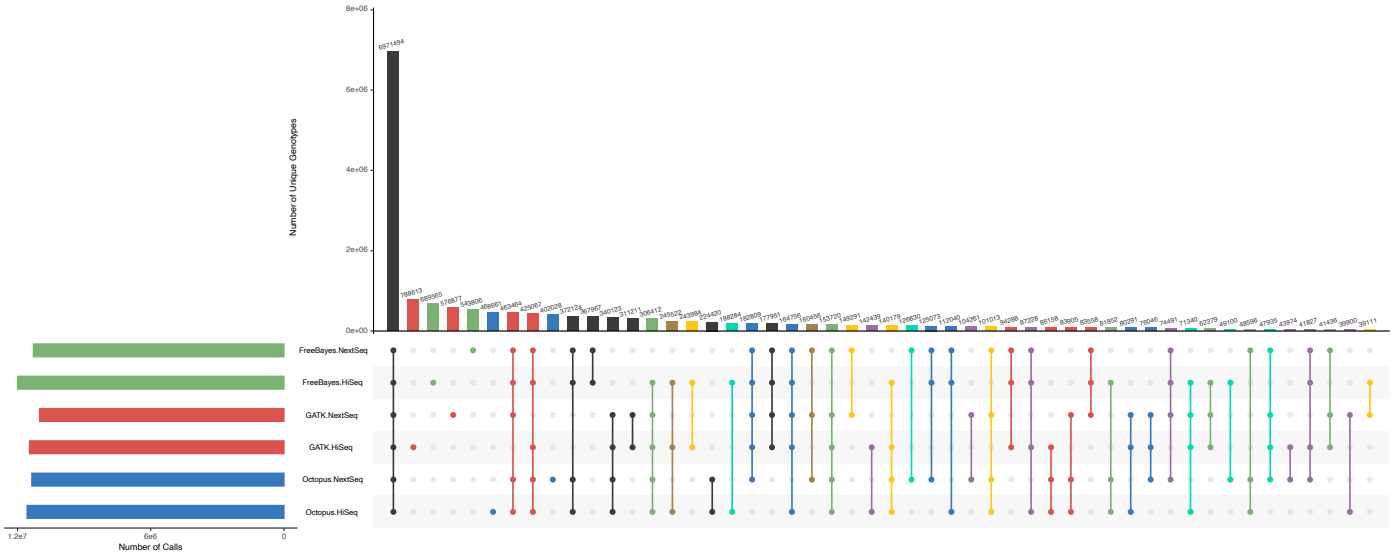

Fig. 3 | Comparison of genotypes called in two Illumina datasets (HiSeq and NextSeq) of banana specimen by Octopus, GATK4, and FreeBayes. 'UpSet' plot shows callset intersections for each caller-dataset pair. The largest $50/63$ intersection sets are shown. Intersections are color coded by caller discordance between the two datasets: No discordances (black), Octopus (blue), GATK4 (red), FreeBayes (green), Octopus & GATK4 (purple), Octopus & FreeBayes (cyan), GATK4 & FreeBayes (yellow), All (brown). The total number of unique genotype calls was 17,277,217.

tion were due to slightly different indels called in each dataset, suggesting failure to discover correct allele(s). To test this, we recalled both datasets with Octopus using the union of the four callsets from Octopus and GATK4 as candidates (Online Methods). The number of called variants increased by $x\%$ and the fraction of concordant calls increased to $x\%$, supporting this hypothesis. Further assessment of a selection of remaining discordant calls using haplotype-tagged and realigned evidence BAMs generated by Octopus (Methods) indicated that major sources of error were: i) Lack of read depth or allele bias; ii) Missmapped reads, possibly due to incomplete or divergent reference; iii) failure to discover a correct allele (in any callset); iv) Probable structural variation.

Discussion

We have shown that genotyping is substantially more difficult in polyploids than in diploids using typical whole-genome sequencing depths, and that there are considerable differences in accuracy between callers. Notably, Octopus produced less than half the total errors of other methods. We believe this is due to Octopus modelling longer haplotypes during genotyping, as this increases statistical confidence in allele copy-number. Re-genotyping using longer haplotypes increased genotyping accuracy, supporting this hypothesis.

Analysis of real autotriploid banana datasets revealed high discordance between callers, and more alarmingly, high discordance for callers on similar datasets of an identical specimen. While 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 clearly higher in real polyploid data. Plausible reasons for this include: i) Greater divergence from the reference genome; ii) Higher levels of repetitive elements in the genome; iii) More structural variation; iv) Less complete reference genome; v) Higher rates of sequencing related errors, such as due to the use of PCR amplification. vi) Callers optimised for human data.

We have only considered single sample polyploid calling in this work, however, multi-sample calling is important for studying population diversity. Population calling in humans is a difficult problem due 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 priors as developed in other methods[20].

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 highly valuable in the development of polyploid calling algorithms, including Octopus. We hope that this work lays the groundwork for future developments.

Table 1 | Concordance in two banana Illumina datasets

Caller	Concordant	Discordant	Total	Concordant %
FreeBayes	9,778,611	3,729,699	13,508,310	72%
GATK4	9,025,920	4,421,440	13,447,360	67%
Octopus	9,854,737	3,219,611	13,074,348	75%

References

1. Song, C. *et al.* Polyploid organisms. *Sci China Life Sci* **55**, 301–11 (2012).
2. Velicky, P. *et al.* Genome amplification and cellular senescence are hallmarks of human placenta development. *PLoS Genet* **14**, e1007698 (2018).
3. Potato Genome Sequencing, C. *et al.* Genome sequence and analysis

of the tuber crop potato. *Nature* **475**, 189–95 (2011).
4. D'Hont, A. *et al.* The banana (musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213–7 (2012).
5. International Wheat Genome Sequencing, C. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361** (2018).
6. Zhuang, W. *et al.* The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet* **51**, 865–876 (2019).
7. Jackson, S. A., Iwata, A., Lee, S. H., Schmutz, J. & Shoemaker, R. Sequencing crop genomes: approaches and applications. *New Phytol* **191**, 915–25 (2011).
8. Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D. & Stromvik, M. V. Current strategies of polyploid plant genome sequence assembly. *Front Plant Sci* **9**, 1660 (2018).
9. Uitdewilligen, J. G. *et al.* A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS One* **8**, e62355 (2013).
10. Bourke, P. M., Voorrips, R. E., Visser, R. G. F. & Maliepaard, C. Tools for genetic studies in experimental populations of polyploids. *Front Plant Sci* **9**, 513 (2018).
11. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* **37**, 561–566 (2019).
12. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**, 595–597 (2018).
13. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**, 555–560 (2019).
14. Cooke, D. P., Wedge, D. C. & Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *bioRxiv* 456103 (2018).
15. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* (2017).
16. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *bioRxiv* (2012).
17. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* (2018).
18. Poplin, R. *et al.* A universal snp and small-indel variant caller using deep neural networks. *Nat Biotechnol* (2018).
19. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**, 912–8 (2014).
20. Blischak, P. D., Kubatko, L. S. & Wolfe, A. D. Snp genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics* **34**, 407–415 (2018).
21. Clark, L. V., Lipka, A. E. & Sacks, E. J. polyrad: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3 (Bethesda)* **9**, 663–673 (2019).
22. Gerard, D., Ferrao, L. F. V., Garcia, A. A. F. & Stephens, M. Genotyping polyploids from messy sequencing data. *Genetics* **210**, 789–807 (2018).
23. Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P. & Jackson, S. A. Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. *Mol Plant* **8**, 831–46 (2015).
24. Yao, Z. *et al.* Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics* **21**, 360 (2020).
25. Cleary, J. G. *et al.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv* (2015).
26. Busche, M., Pucker, B., Viehover, P., Weisshaar, B. & Stracke, R. Genome sequencing of musa acuminata dwarf cavendish reveals a duplication of a large segment of chromosome 2. *G3 (Bethesda)* **10**, 37–42 (2020).

Methods

Synthetic polyploids with real reads. Raw reads (FASTQ) generated for the PrecisionFDA Truth v2 challenge were downloaded from the DNAnexus portal (`https://precision.fda.gov/challenges/10`). Each FASTQ was line counted to ensure realistic haplotype frequencies, before concatenating contributing samples to make the full data polyploid dataset. Downsampling was performed directly on the FASTQ files using *seqtk* with default seed. The sampling fraction was set using: *test depth* / *full depth*, where full depth is $35 \times$ *ploidy*$/2$. Illumina reads were mapped with BWA-mem using default alignment parameters.

Variant calling synthetic 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 Octopus, we also set *--max-genotypes* to 20000 and specified *--disable-early-phase-detection*.

Filtering variant calls. We used HiSeq-2500 data from GIAB for samples HG001, HG006, and HG007 to train Octopus's random forest for polyploids, using the same mixing procedure described or the evaluations. The forest was also trained on HG001. This forest was used to filter all polyploid calls, including the banana datasets. The v0.7.0 germline forest was used to filter diploid calls.

For GATK4, we used filter expressions:

For FreeBayes, we used filter expression:

Identifying copy-number errors. Biallelic sites were identified after intersecting false negative and false positive genotypes previously identified by RTG Tools vcfeval. The intersection was also performed using RTG Tools vcfeval with the *--squash-ploidy* and *--output-mode=annotate* options.

Long haplotypes with Octopus.

Variant calling banana. We tweaked caller parameters for banana calling to account for the high diversity present in banana species. For GATK4, we set *--heterozygosity* to 0.05, *--heterozygosity-stdev* to 0.05, and *--indel-heterozygosity* to 0.01. For FreeBayes, we set *--theta* to 0.05. For Octopus, we set emph--snp-heterozygosity to 0.05, *--snp-heterozygosity-stdev* to 0.05, and *--indel-heterozygosity* to 0.01, *--max-haplotypes* to 300, and *--max-genotypes* to 30000.

Banana concordance analysis. Callsets for the banana datasets were intersected using a custom script (`https://github.com/dancooke/starfish`) that invokes both RTG Tools vcfeval (that only supports 2-way comparisons) and bcftools to achieve multi-sample haplotype-aware comparisons.

Code availability. Octopus source code and documentation is freely available under the MIT licence from `https://github.com/luntergroup/octopus`. Custom Python code used for data analysis is available from `https://github.com/luntergroup/polyploid`.

Data availability. All primary data used for analysis is available from public sources. Links are provided in Supplementary Note x and can also be found, along with automatic download options, in the online code (`https://github.com/luntergroup/polyploid`). Synthetic data can easily be reproduced using the online code.

Acknowledgements. We would like to thank Len Trigg at Real Time Genomics for kindly updating RTG Tools to support polyploid genotypes. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

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

the project.