Evaluating small-variant genotyping errors 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)

The accuracy of polyploid genotyping from sequencing affects several active research areas, notably plant breeding. However, compared with the situation for diploids, where comprehensive ground truth sets and robust benchmarking tools are standard, polyploids have been neglected; there are no benchmarks considering genotype errors for small variants using real data. Genotyping accuracy from sequencing in polyploids is therefore essentially unknown. 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 polyploid Illumina and PacBio HiFi data by *in silico* mixing of diploid Genome In a Bottle samples. We find that polyploid genotyping errors are abundant for typical sequencing depths, but Octopus makes less than half the errors of other methods. We show our results give a credible upper-bound on performance in real polyploids by evaluating autotriploid banana and autotetraploid potato 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 that aim to improve resilience to climate change and disease. More recently, sequencing of polyploids for genotyping and genome assembly has been used, and several high-quality reference genomes have been assembled[3–7]. Despite these advances, methods for genotyping polyploids from sequencing data has received little scrutiny in comparison to those for diploids[8–10].

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 alleles, but sequencing reads are only informative of distinct alleles. It therefore becomes more difficult 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 the power to genotype individual alleles by jointly evaluating combinations of several proximal alleles (haplotypes). They are now standard for diploid calling[11–16] and are becoming more common for somatic mutation calling in tumours[11]. However, despite the success of haplotype-based methods for diploid calling, only a minority are capable of polyploid calling[11–13], and none have been rigorously tested for this purpose. Specialised methods for polyploid genotyping have been developed[17–19], but are only suitable for biallelic SNPs. Moreover, existing benchmarks of polyploid calling methods fall short of the standard demanded for diploid calling[10,20,21]. In particular, we are not aware of any that consider indels, genotyping errors in real sequencing data, or representation differences between callers[10]. Genotyping error rates in polyploids are therefore highly uncertain, potentially misleading downstream analysis.

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 a haplotype-aware comparison tool - RTG Tools[22]. Our aim was to evaluate genotyping accuracy in polyploids from whole-genome sequencing experiments using state-of-the-art methods, including our own - Octopus[11]. Our analyses are made available in online Python code (`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 HG003, HG004, and HG002. 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 read coverage distributions were similar for

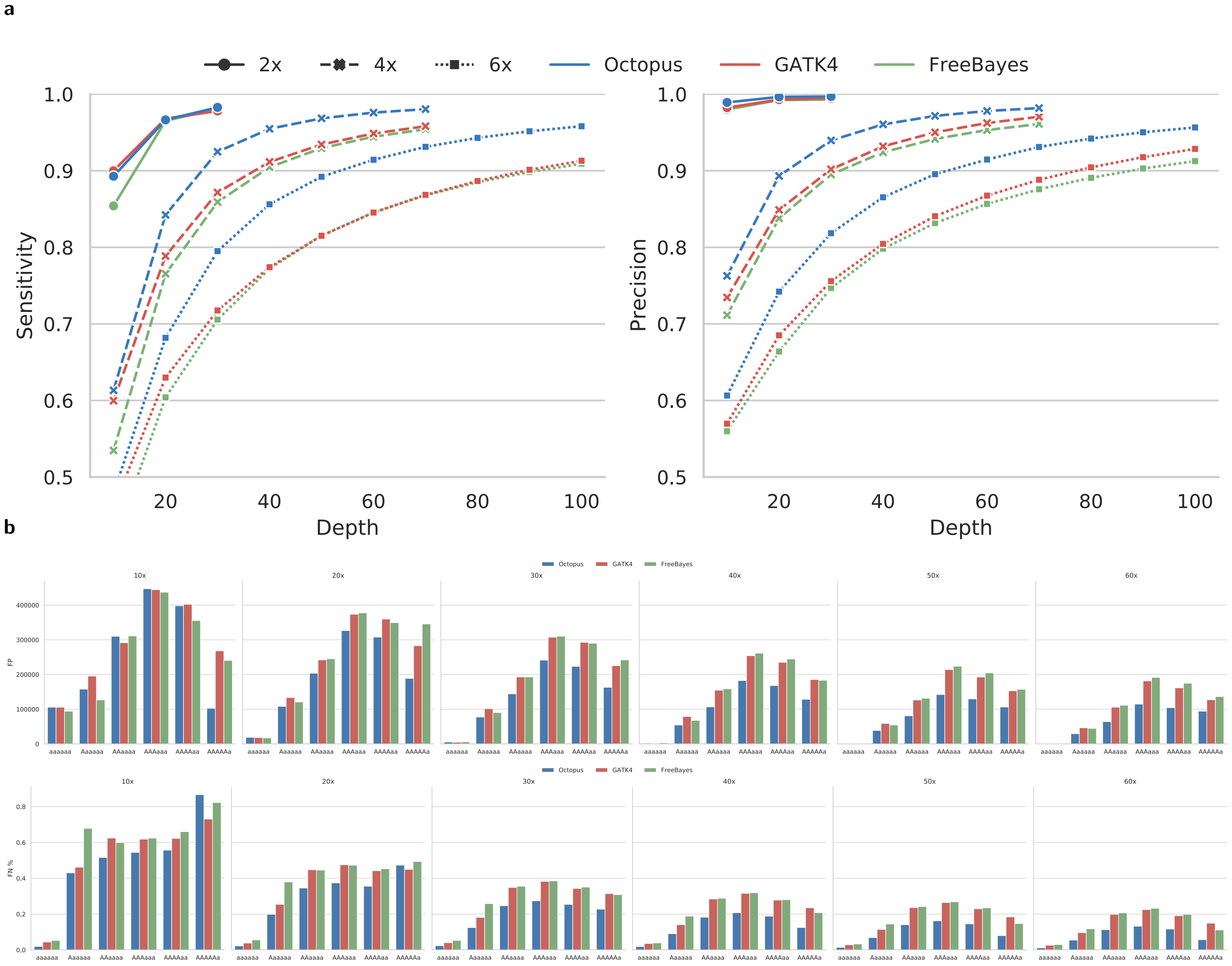

Fig. 1 | Genotyping accuracy in synthetic polyploids. **a** Sensitivity and precision by depth for each caller on real diploid (2x), and synthetic tetraploid (4x) and hexaploid (6x) Illumina datasets. **b** Counts of false positive biallelic genotypes stratified by depth and copy-number (top). Proportion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

each contributing sample (Supplementary Note x and Supp Fig x), ensuring realistic heterozygous allele frequencies. We then randomly downsampled the full datasets, starting 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 (Supplementary Note x), Illumina reads using BWA-MEM and PacBio HiFi reads using pbmm2.

Polyploid genotyping accuracy from short-read WGS. We evaluated three popular germline variant callers that support polyploid genotypes: Octopus[11]; GATK4[12]; and FreeBayes[13], on all synthetic polyploid Illumina datasets, and in the diploid HG002 sample to get performance baselines. Other notable germline callers, such as DeepVariant[15], Strelka2[14], and Platypus[16], were not included because they do not support polyploid calling. We also ignored methods that call polyploid SNVs but not indels, such as polyRAD[18]. Other than specifying the ploidy, we used nearly default setting for all callers (Supplementary Note 3). Octopus calls were random forest filtered, GATK4 and FreeBayes calls were hard-filtered using recommended thresholds (Supplementary Note 3). Variants were compared using RTG Tools[22] vcfeval (Supplementary Note 4).

Genotyping accuracy was considerably lower for polyploids than for diploids. For 30x sequencing depth, on average $1/200$ diploid genotype calls were incorrect, compared 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 on average half the total errors of GATK4 and FreeBayes. There were similar differences for unfiltered calls (Supp Fig x). Sequencing depth showed a typical logarithmic relationship with accuracy (F-measure) for both tetraploid and hexaploid samples. The largest F-measure difference between callers occurred at moderate se-

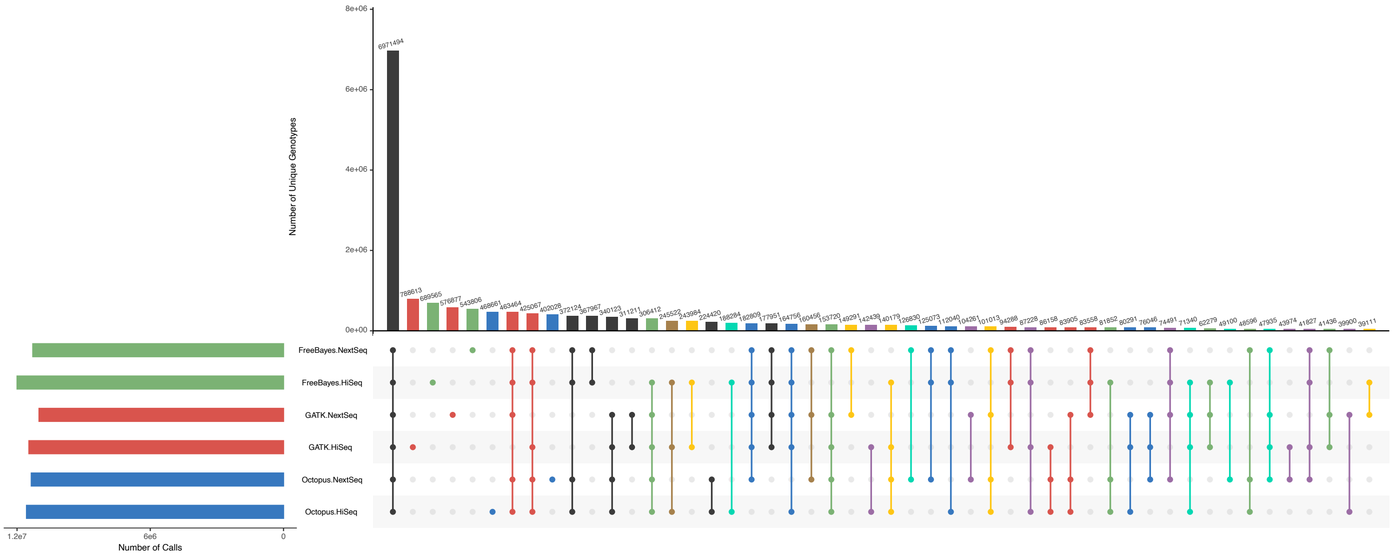

Fig. 2 | Comparison of genotypes called in two Illumina datasets (HiSeq and NextSeq) of banana specimen by Octopus, GATK4, and FreeBayes. 'UpSet' plot shows callset intersections for each caller-dataset pair. The largest $50/63$ intersection sets are shown. Intersections are color coded by caller discordance 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.

quencing depth: 30x for tetraploid, $50x$ for hexaploid. The F-measure also showed a non-linear relationship with ploidy; the performance lost from doubling the ploidy was not recovered by doubling the depth, and the difference increased with depth.

The majority of false positives resulted from incorrect genotype copy-numbers: $x\%$ of false positive biallelic genotype calls ($x\%$ of all false positives) were due to incorrect copy number, although there were notable differences between callers (Fig x). The most common false positive for all depths was the balanced heterozygote: AAaa or AAAaaa (Fig. 1b), of which $x\%$ were due to incorrect copy number. There was no clear bias in the directionality of copy number errors (Supp Fig x). The most common biallelic false negatives in tetraploids were heterozygotes with a single variant copy, while for hexaploids it was heterozygotes with two variant copies (Fig. **??**). However, normalising by the truth prevalence shows that the most frequent false negative for depths ≥ 30x is the balanced heterozygote for all plodies; for depths ≤ 20x the most frequent false negative was the singleton heterozygote.

Genotype quality scores were well calibrated for all callers, although there was a clear advantage to Octopus' machine learning filtering approach (Fig. **??**); Octopus was the only method where filtering improved F-measure on all tests.

Genotyping in real polyploids. To support our previous results on real polyploid samples, we called genotypes in recently sequenced banana[23] and potato[24] specimens. Dwarf Cavendish banana (Musa acuminata) is autotriploid consisting of 11 chromosomes with a haploid genome size of around 523Mb, and is an important export-product for many developing countries[4]. Potato (Solanum tuberosum) is autotetraploid consisting of 12 chromosomes with haploid genome size of around 844Mb[3], and is a vital food source for a large fraction of the world population and is the most cultivated non-grain crop[3]. The banana was whole-genome sequenced twice on different machines: once on an Illumina NextSeq-500 to 65x coverage and once on an Illumina HiSeq-1500 to 55x coverage. Both datasets were mapped to the DH Pahang v2 reference[23]. The potato was whole-genome sequenced on an Illumina to 40x coverage, and we mapped reads to the Solanum tuberosum DM1-3 reference genome[25]. Due to lack of truth data for these samples, we relied on less rigorous means to access the quality of the callsets.

We evaluated the concordance of callers on the two banana datasets with haplotype-aware intersections (Online Methods). The genotype set called by all callers in both datasets, while substantially the largest, only accounted for 40% of all unique genotype calls; 20% of calls were unique to a single callset (Fig. 2). There were also considerable differences between callers: GATK4 had 37% more discordant calls compared with Octopus and 16% more than FreeBayes, despite making less calls overall than FreeBayes and only 3% more than Octopus (Table 1). We found only slightly higher concordance levels when considering allele matching (Supplementary Fig x and Supplementary Table x); 55% of unique alleles were present in all callsets while 12% were unique to a single callset. Since discordant calls imply at least one false positive or false negative, these results suggests that, unlike for our synthetic data, the majority of false calls arise from unique variant allele calls rather than copy-number errors.

Manual review of discordant calls suggested that a large proportion were due to slightly different proximal indels called in each dataset, suggesting failure to discover correct alleles. To test this, we recalled both datasets with Octopus, using variants called in both datasets by Octopus and GATK4 previously as candidates (Online Methods). The number of called variants

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%

line Methods). The number of called variants increased by $\%$ and the fraction of concordant calls increased to 80%, supporting our hypothesis. Further assessment of a selection of remaining discordant calls using haplotagged and realigned evidence BAMs generated by Octopus indicated that major error modes 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.

We accessed the potato calls in a similar manner; x/y ($z\%$) genotypes were made by all three callers, with similar overlaps between callers as for the synthetic and banana results. We manually reviewed a selection of calls by inspecting both the Illumina alignments, and PacBio long-read alignments from the same study.

Discussion

We have shown that polyploids can be reliably variant called for moderate sequencing depths (e.g. ≥ 30x), but accurate genotyping remains challenging for sequencing depths below 60x. We showed that there is considerably less concordance between variant calling methods for polyploids than diploids, and that Octopus substantially outperforms other methods, including GATK4 and FreeBayes, particularly for genotyping.

Our analysis is not without limitations. Our most reliable analysis was on human genomic data, but most polyploid genomes are more challenging to call than human genomes due to higher repetitiveness, less complete references, and higher reference divergence. These challenges likely explain why the upper-bounds on precision that we calculated from the banana datasets were considerably lower than would be expected via extrapolation from our synthetic tetraploid tests. Notably, the divergence of the banana sample from the reference was substantially higher than the human sample. Moreover, our study focuses on small variants, but polyploids, including many plants, are highly susceptible to structural changes, including copy-number changes. Furthermore, without a ground truth it is very difficult to have high confidence in sensitivity. We therefore stress the importance of treating our performance figures on synthetic polyploid data as upper bounds of what can be achieved when applying the same methods to real polyploid 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[17].

In summary, we have conducted the most comprehensive study on genotyping errors in polyploids to date. We found considerable differences in performance between variant calling methods, but

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. 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).
6. International Wheat Genome Sequencing, C. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361** (2018).
7. 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).
8. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* **37**, 561–566 (2019).
9. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**, 595–597 (2018).
10. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**, 555–560 (2019). URL `https://www.ncbi.nlm.nih.gov/pubmed/30858580https://www.nature.com/articles/s41587-019-0054-x.pdf`.
11. Cooke, D. P., Wedge, D. C. & Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *bioRxiv* 456103 (2018).
12. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* (2017).
13. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *bioRxiv* (2012).
14. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* (2018).
15. Poplin, R. *et al.* A universal snp and small-indel variant caller using deep neural networks. *Nat Biotechnol* (2018).
16. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**, 912–8 (2014).
17. 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).
18. 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).
19. Gerard, D., Ferrao, L. F. V., Garcia, A. A. F. & Stephens, M. Genotyping polyploids from messy sequencing data. *Genetics* **210**, 789–807 (2018). URL `https://www.ncbi.nlm.nih.gov/pubmed/30185430https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6218231/pdf/789.pdf`.
20. 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).
21. Yao, Z. *et al.* Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics* **21**, 360 (2020).
22. Cleary, J. G. *et al.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv* (2015).
23. 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).

24. Kyriakidou, M. *et al.* Structural genome analysis in cultivated potato taxa. *Theor Appl Genet* **133**, 951–966 (2020).
25. Hardigan, M. A. *et al.* Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated solanum tuberosum. *Plant Cell* **28**, 388–405 (2016).

Online methods

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.