Variant calling and 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)*

Accurate variant calling and genotyping of polyploids from high-throughput sequencing benefits several important research areas, notably plant breeding. However, compared with the situation for diploid organisms, where comprehensive ground truth sets and robust benchmarking tools are standard, polyploids have been neglected; there are no independent validation samples that include both SNVs and small indels, and only recently was a haplotype-aware comparison tool - RTG Tools - updated with polyploid support. Genotyping error profiles for polyploids are therefore highly uncertain. We previously introduced a variant calling method - Octopus - that was shown to accurately call germline variants in diploid individuals and somatic mutations in cancers. In this paper, we evaluate small variant calling performance in synthetic polyploid Illumina and PacBio HiFi datasets by *in silico* mixing of diploid Genome In a Bottle samples. We find that Octopus substantially outperforms other methods, but genotyping errors are high for typical whole-genome sequencing depths. We show our results reliably demonstrate performance on real polyploids by analysing real autotriploid banana and autotetraploid potato datasets.

Polyploidy is naturally common in many plant species, including important agricultural crops such as wheat, potato, oat, coffee, rapeseed, cotton, banana, and sugar cane[1]. Polyploidization 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 can improve resilience to climate change and disease. More recently, sequencing of polyploids for genotyping and genome assembly has been used, resulting in several high-quality reference genomes[3–7]. Despite these advances, methods that can accurately genotype polyploids from sequencing data have received little attention in comparison to those for diploids, for which several high-quality validation samples exist that have very low genotyping error rates over the majority of the human genome[8,9].

Variant calling and genotyping in polyploids is considerably more difficult than in diploids because the number of possible genotypes at a given loci is exponential in the ploidy, but sequencing reads are only informative of one haplotype. 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 with low copy-number and increased ploidy.

Haplotype-based methods increase the power to genotype individual alleles by jointly evaluating genotypes of several proximal alleles (haplotypes). They are now standard for diploid calling[10–14] and are becoming more common for somatic mutation calling in tumours that provide similar challenges to polyploid calling[10]. The increase in genotyping power is due to some haplotypes being highly implausible under the data likelihood model, and because sequencing errors and variation priors can be more accurately modelled. However, despite the success of haplotype-based methods for diploid calling, only a minority of them are capable of polyploid calling[10–12], and none have been rigorously tested for this purpose. Specialised methods for polyploid calling do exist[15], but most appear to only call SNPs, are not maintained, or do not work with standard bioinformatics formats. Moreover, independent attempts made to benchmark polyploid calling methods fall short of the standard demanded for diploid calling[16,17]. For example, we are not aware of any benchmarks that consider indel variants or account for variant representation differences, likely due to an absence of effective comparison tools until now. Genome-wide genotyping error rates for polyploids are therefore currently highly uncertain, potentially resulting in underutilisation of the sequencing data being generated and misleading inferences.

We sought to address some of these issues by conducting an in-depth analysis of polyploid small variant calling using an independent and comprehensive ground truth and a haplotype-aware comparison tool - RTG Tools[18] - that was recently updated to support polyploid genotypes. Our aim was to evaluate genotyping error rates for polyploid whole-genome sequencing experiments from state-of-the-art methods. We sought to specifically test our own haplotype-based method, Octopus[10], for polyploid calling and make improvements where possible. Octopus is open-source (`https://github.com/luntergroup/octopus`), and all of our results presented here are reproducible and extendible using open-source Python code (`https://github.com/luntergroup/polyploid`).

RESULTS

Optimising Octopus for polyploid calling

While the methods that we previously described for Octopus are fully capable of polyploid calls[10], in practice we found some issues. Runtimes were prohibiting for high ploidies due 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 low copy-number variants was less than ideal due 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 less 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 sample ploidy. The final optimisation we made was to retrain the random forest filtering classifier on polyploid data (Supp Note 1).

Benchmarking synthetic polyploid genomes

We investigated variant calling performance in polyploids by creating synthetic tetraploid and hexaploid sequencing samples with high quality truth sets that were constructed 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. The polyploid Illumina NovaSeq and PacBio HiFi whole-genome test data were made by mixing reads generated by Precision FDA Truth v2 challenge for each sample independently (Supp Note 2). Each individual sequencing run targeted 35x coverage, resulting in 70x and 105x datasets for the tetraploid samples and hexaploid samples, respectively. We then randomly downsampled the full tetraploid datasets to 20x, 30x, 40x, 50x, and 60x, and the full hexaploid datasets to 20x, 30x, 40x, 50x, 60x, 70x, 80x, 90x, and 100x, resulting in $2 \times 6 + 2 \times 10 = 32$ tests. All reads were mapped to GRCh38 (hs38DH), Illumina reads using BWA-MEM and PacBio HiFi reads with pbmm2.

We compared Octopus to GATK4[11] and FreeBayes[12], both of which are popular germline variant callers that support polyploid genotypes. Other notable germline callers, such as DeepVariant and Strelka2, were not included because they do not support polyploid calling. We also ignored methods that call polyploid SNVs but not indels, such as polyRAD. Other than specifying the ploidy, we used default setting for GATK4 and nearly default setting for FreeBayes, and made a reasonable effort to filter calls for both tools (Supp Note 3). Variants were compared using RTG Tools vcfeval (v3.12), which is the only variant comparison tool we are aware of that is both haplotype-aware and supports polyploid genotypes. In particular, we evaluated callsets on both allele match - ignoring genotype errors - and genotype match (Supp Note 4).

Genotyping accuracy for typical short read WGS

We found that Octopus had highest F-measure in every test.

Long reads improve genotyping accuracy

Genotyping in real polyploids

To support our previous results on real polyploid samples, we called variants and genotypes in recently sequenced banana[19] and potato[20] 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 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[19]. The potato was sequenced on an Illumina to 40x coverage, and we mapped reads to the Solanum tuberosum DM1-3 reference genome[21]. Due to lack of truth data for these samples, we relied on less rigorous means to access the quality of the callsets.

First, we evaluated the concordance of each caller between the two independent datasets of the banana specimen. In total there were k unique genotypes called, x (y%) of which were called by all three callers in both callsets, which we considered likely to be true positives. Octopus had the highest concordance between the two callsets, with x/y (z%) of genotypes being consistent (called in both callsets), compared with x/y (z%) for GATK4 and x/y (z%) for FreeBayes. Since any inconsistent genotypes calls necessarily imply at least one false positive, or false negative, the concordance gives a lower bound on false positive rate. We accessed relative sensitivity by manually reviewing a selection consistent calls made by only one caller, and further accessed specificity by manually reviewing cases where two callers made a consistent call and the other made a consistent no call. We found that x/y (z%) Octopus-only calls were likely real, compared with x/y (z%) and x/y (z%) for GATK4 and FreeBayes, respectively. Most Octopus-only calls were due to Reassuringly, we found that overlaps for consistent calls between callers were in similar proportions to the synthetic tetraploid and hexaploid callsets.

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. $\geq$ 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 often more difficult to sequence and map than human genomes due to being more repetitive. 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 advise that the error rates indicated by our analysis are taken as lower bounds. Ultimately, a resource similar to GIAB but with real polyploid specimens will be required to provide a better error rate estimates. We hope these results will be useful in heading towards that goal.

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 different genotype prior models than those typically used for diploid analysis[22].

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. Cooke, D. P., Wedge, D. C. & Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *bioRxiv* 456103 (2018).
11. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* (2017).
12. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *bioRxiv* (2012).
13. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* (2018).
14. Poplin, R. *et al.* A universal snp and small-indel variant caller using deep neural networks. *Nat Biotechnol* (2018).
15. 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).
16. 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).
17. Yao, Z. *et al.* Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics* **21**, 360 (2020).
18. Cleary, J. G. *et al.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv* (2015).
19. 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).
20. Kyriakidou, M. *et al.* Structural genome analysis in cultivated potato taxa. *Theor Appl Genet* **133**, 951–966 (2020).
21. 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).
22. 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).

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 Supp 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.

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.