Title: Variant calling in polyploids for population and quantitative genetics

Short title: Phillips - Variant calling in polyploids

Authors: Alyssa R. Phillips[1*]

Affiliations: [1]Department of Evolution and Ecology, University of California, Davis, Davis CA 95616

***Corresponding author email:** arphillips@ucdavis.edu

Abstract

Advancements in genome assembly and sequencing technology have made whole genome sequence (WGS) data and reference genomes accessible to study polyploid species. The genome-wide coverage and greater marker density provided by WGS data, compared to popular reduced-representation sequencing approaches, can greatly improve our understanding of polyploid species and polyploid biology. However, biological features that make polyploid species interesting also pose challenges in read mapping, variant identification, and genotype estimation. Accounting for characteristics, like allelic dosage uncertainty, homology between subgenomes, and variance in chromosome inheritance mode, in variant calling can reduce errors. Here, I discuss the challenges of variant calling in polyploid WGS data and discuss where potential solutions can be integrated into a standard variant calling pipeline.

Keywords: polyploidy, variant calling, whole genome sequence, population genetics, quantitative genetics, mixed-ploidy

1. Introduction

Recent progress in genome assembly and sequencing technology has increased accessibility to study the genomics of polyploids, or organisms that have experienced whole genome duplication and have more than two sets of chromosomes (Formenti et al., 2022; Gladman et al., 2023). Notably, improvements in long-read sequencing and the accuracy of scaffolding technology have enabled the assembly of highly heterozygous and polyploid reference genomes at a chromosome-scale (Kyriakidou et al., 2018; Hotaling et al., 2023). In parallel, the cost of short-read sequencing has continued to decline causing whole genome resequencing of polyploid populations to become increasingly feasible (Fuentes-Pardo and Ruzzante, 2017). As polyploidy is a critical character of cancer cells, common in fish, amphibians, and insects, and ubiquitous in the plant kingdom, including many economically important crops, the extension of modern genomics technologies to polyploid systems is important for our broader understanding of medicine, and biodiversity, agriculture (Udall and Wendel, 2006; Wood et al., 2009; Zack et al., 2013; One Thousand Plant Transcriptomes Initiative, 2019; Román-Palacios et al., 2021; David, 2022). These advances have already begun to improve our understanding of the origins of polyploid species (Bertioli et al., 2019; Edger et al., 2019; Goeckeritz et al., 2023), genome reorganization and stabilization after polyploidization (Chen et al., 2020; Bohutínská et al., 2021; Wang et al., 2022; Session and Rokhsar, 2023), and the role of polyploidy in adaptation of wild and domesticated species (Hollister et al., 2012; Chen et al., 2021; Lovell et al., 2021; Ebadi et al., 2023; Hämälä et al., 2023). Nevertheless, these studies have only scratched the surface of polyploid biology.

Population and quantitative genetics particularly benefit from the availability of reference genomes and whole genome sequence (WGS) data. These fields use variable loci, loci with two or more alleles segregating in a population, to study the genetic composition of populations and complex traits over space and time in response to selection, genetic drift, mutation, and migration. WGS data in combination with a reference genome offers genome-wide coverage and the ability to identify variable loci, also referred to as variants, at a higher density than reduced representation sequencing (RRS) approaches. RRS approaches, such as genotype-by-sequencing (GBS) and restriction site-associated DNA sequencing (RADseq), are currently used in the majority of polyploid population and quantitative genetics studies due to their comparatively low cost and the growing number of user-friendly software packages for analysis (Poland and Rife, 2012). RRS approaches are useful for sampling a portion of the genome to, for example, characterize population structure or complete quantitative trait locus (QTL) analysis. However, RRS does not have high enough marker density for genome-wide analyses central to studying patterns of selection, identifying the genetic basis of adaptive traits, and genomic prediction (Tiffin and Ross-Ibarra, 2014; Lowry et al., 2017; but see de Bem Oliveira et al., 2020). Additionally, WGS data improves the detection of structural variants (SVs) and transposable elements (TEs), although both are still challenging even in diploid systems (Ewing, 2015; Baduel et al., 2019; Mahmoud et al., 2019; Cooke et al., 2022; Ramakrishnan et al., 2022). Detection and inclusion of SVs and TEs are important because they affect gene expression and function and are signatures of the stabilization and reorganization of the genome post-polyploidization (Lisch, 2013; Kosugi et al., 2019).

The improvement in variant detection offered by WGS data is useful only when variants can be confidently called and genotypes accurately estimated. Typical sources of error in diploid variant calling include sequencing errors, misalignment of reads to the reference genome, misassembly of the reference genome, and natural structural variation (Li, 2014; Mahmoud et al., 2019; Lou and Therkildsen, 2022). Polyploidy exacerbates these sources of error and introduces additional challenges due to the associated characteristics like large haploid genome sizes, homology between subgenomes, genome fractionation, and elevated polymorphism (Bennett and Leitch, 2011; Page and Udall, 2015; Blischak et al., 2018). As a result, there may be higher variant calling errors in polyploids. Errors in the variant calling pipeline will subsequently be carried into all downstream analyses leading to misestimation of metrics like allele frequencies, heterozygosity, and linkage.

Universal solutions to reduce errors in variant calling are challenging to identify as polyploids are not a uniform group. Polyploids are generally categorized as allopolyploids, which form through hybridization of two or more species, or autopolyploids, which derive from genome doubling of a single species. Further, they can be described by their chromosome inheritance patterns. Allopolyploids have disomic inheritance, like diploids where chiasma for between only homologous chromosomes, and autopolyploids have polysomic chromosome inheritance, where there is no preferential pairing among chromosomes and chiasmata may form between more than two homologous chromosomes (Stift et al., 2008). However, the rate of preferential pairing and chromosome inheritance mode may vary across the genome in allo- and autopolyploids depending on the relatedness amongst subgenomes and the time since polyploidization (Stebbins, 1947; Mason and Wendel, 2020). This distinction between inheritance modes is important because even low rates of recombination between subgenomes can bias allele frequencies to be more homozygous than expected (Meirmans and Van Tienderen, 2013). Polyploids may additionally vary in haploid genome size, mating system, repeat content, and degree of diploidization, all of which may impact variant calling and genotype estimation.

In this review, I identify significant challenges of variant calling in polyploid WGS data and, where available, propose potential solutions that can be integrated into standard variant calling pipelines (Figure 1; Appendix S1, see Supporting Information with this article; reviewed in Van der Auwera et al., 2013; De Summa et al., 2017; Fuentes-Pardo and Ruzzante, 2017; Therkildsen and Palumbi, 2017; O'Leary et al., 2018; Lou et al., 2021). The scope of this discussion is limited to WGS data aligned to the study species' reference genome, although aspects of this discussion may apply to RRS and reference-free approaches. Additionally, I focus on the identification of single nucleotide variants (SNVs) as well as small SVs (< 50 bp) that can be identified by some polyploid variant calling software (Cooke et al., 2022). As the genomics of polyploids is a rapidly growing area of research, established best practices are limited. By highlighting barriers in variant calling, I aim to raise readers' awareness of potential sources of error and motivate the innovation of new and effective solutions.

2. Challenges to variant calling in polyploid systems

2.1 Resource requirements scale with genome size

The foremost barrier to polyploid genomics remains the cost of sequencing and high-performance computing (HPC) resources for analysis. Sequencing cost increases with both haploid genome size and ploidy level while computational costs primarily scale with haploid genome size. Sequencing large genomes is expensive as more sequencing runs are required to reach a target coverage, or the genome-wide average number of reads sequenced for a given site. For example, Chen et al. (2024) have found sequencing the allohexaploid bread wheat genome to 5X coverage currently costs 473 times that of diploid rice and 21 times that of maize, a diploidized paleotetraploid (Gaut and Doebley, 1997). This disparity in sequencing cost at low coverage is increased by many existing polyploid genotyping algorithms requiring high coverage to overcome allelic dosage uncertainty, which is the ambiguity in the number of alternate allele copies in polyploid genotypes (Gerard et al., 2018; Clark et al., 2019; Cooke et al., 2022). The minimum coverage requirement to obtain high-confidence genotypes may range from 10 to over 50X depending on the ploidy level and genotyping software, whereas diploids need only 8X coverage (Cooke et al., 2022; Jighly, 2022). After sequencing has been accomplished, access to HPC is needed for data storage and analysis because the size of sequence alignment files (BAMs) and variant call files (VCFs) produced in the variant calling pipeline scale with genome size and sample size (Muir et al., 2016; Weiß et al., 2018). Failing to sequence to sufficient coverage or limiting sample size to meet budget constraints may result in insufficient sampling of alleles and rare variants, the misestimation of allele frequencies, and low power in analyses like admixture analysis and genome wide association (Jighly, 2022).

2.2 Genome-wide redundancy and elevated polymorphism increase errors in read mapping

Aligning reads to polyploid genomes is challenging because polyploids have an elevated level of polymorphism and multiple occurrences of related sequences (Otto and Whitton, 2000; Page and Udall, 2015). Both of these biological features violate assumptions of read mapping algorithms that assume divergence among loci is larger than divergence among alleles at a single locus (Musich et al., 2021); polymorphism creates an excess of divergence while repeated sequences are too similar. Violation of this assumption results in the incorrect and failed mapping of reads. I will briefly describe how these two biological features may create genotyping errors.

As the density of SNVs and SVs in a locus increases, sequence similarity among alleles declines and reads containing alternate alleles are less likely to align (Nielsen et al., 2011; Brandt et al., 2015). This is an issue in polyploids as they are expected to have higher diversity than their diploid progenitors due to functional redundancy between subgenomes enabling the accumulation of mutations. Additionally, the post-polyploidization process of fractionation, which is gene loss leading to stabilization of the polyploid genome or diploidization, increases structural variation (Haldane, 1933; Otto and Whitton, 2000; Ma and Gustafson, 2005; Emery et al., 2018; Beric et al., 2021). As an example in the 1000 Genomes Project (*Homo sapiens*), 18.6% of SNV calls in highly polymorphic *HLA* genes were incorrect due to failed mapping of the alternate allele creating bias towards the reference allele, known as allele bias (Brandt et al., 2015). Alternate reads may also fail to align to inversions due to disagreement at the inversion boundaries, and reads mapping to presence-absence variants (PAVs) will fail to align if the reference contains the ‘absence’ variant (Sun et al., 2018; Gui et al., 2022). As a result, the reference genotype selected for read mapping and time since whole genome duplication will determine the extent of allele bias and the variants detected. Allele bias will be highest in autopolyploids, where reads are aligned to only one copy of the duplicated genome (see Section 2.4). Allele bias is likely an issue genome-wide, although the effect of increased polymorphism on read mapping has yet to be quantified in a polyploid system.

Analogously, genomic features like loci of common ancestry, repetitive elements, and copy number variants (CNVs) promote mismapping because there are multiple occurrences of similar sequences across the genome. In autopolyploids, whole genome duplication produces duplicate loci between subgenomes that are indistinguishable immediately after duplication. Whereas in allopolyploids, loci of common ancestry are brought back together by hybridization. Both diploids and polyploids contain repeat dense regions and CNVs caused by small-scale duplications and retrotransposons (Brandt et al., 2015). As a result, reads may have equal similarities to multiple positions in the reference genome causing reads to equally map to multiple loci (i.e. multiply mapping reads) or improperly align to a closely related locus (Li et al., 2008). The extent of error in read mapping due to these redundant genomic features is dependent on the divergence among the loci of common ancestry, known as homologous loci, the age of the polyploidization event, the divergence between parental genomes, mutation rate, and strength of selection on a given locus. Given these factors, read mapping will be most challenging where loci of common ancestry have not accumulated mutations, such as immediately after whole genome duplication or in genes under purifying selection. Additionally, read mapping may be challenging in recently formed polyploids if purifying selection is relaxed genome-wide post-polyploidization allowing rapid TE expansion (McClintock, 1984).

If the errors in read mapping discussed here are not resolved, failed alignment of reads may lead to the undercalling of variants, overestimation of homozygosity, and underestimation of population alternative allele frequencies. The mismapping of reads further exacerbates these issues in addition to creating false variants which could create false signals of allele sharing and alter patterns of genome-wide heterozygosity. This can significantly increase downstream errors in the estimation of population divergence, gene flow, genome-wide diversity, and identification of causal variants in GWAS and selection scans.

2.3 Incomplete or misassembled polyploid reference genomes increase genotyping error

Undetected errors in the assembly of polyploid genomes create genotyping errors similar to homologous loci and SVs. For instance, chimeric subgenome assemblies, where scaffolds from one subgenome are misassembled into another subgenome, cause reads to fail to map at misassembled scaffold junctions. This leads to genotyping errors at scaffold junctions and incorrect variant positions that impact analyses using linkage information, such as genome scan approaches and estimating runs of homozygosity. In an incomplete reference genome, reads belonging to missing regions will either not align or map to homologous loci (Fig. 2). Reads that successfully map to a homolog are likely to be biased toward the reference allele. However, if reads with the alternative allele do align to a homolog, false heterozygotes may be called (Fig. 2A). Comprehensively addressing the challenge of poor read mapping caused by low reference genome quality will require continued improvement of the reference genome. As comprehensive reviews on genome assembly are available elsewhere (Zhang et al., 2019; Zhou et al., 2022; Gladman et al., 2023), I later discuss practical solutions to mitigate these issues and enhance the accuracy of genotyping when using existing genome assemblies.

2.4 Allele dosage cannot be determined if ploidy and inheritance mode are unknown

Determining the allele dosage**,** the number of reference and alternate alleles, present at each sequenced site for a given individual is imperative for accurate genotyping. In diploids, the reference genome is ideally phased, meaning the maternal and paternal copy of each chromosome is assembled so each chromosome in the assembly has two 'haplotypes' (Gladman et al., 2023). All reads are aligned to only one of the two haplotypes and, as a result, the possible genotype values at a site are 0, 1, and 2 corresponding to the number of alternate alleles. The range of potential genotypes for a polyploid is less clear as there are multiple factors to consider: ploidy level, chromosome inheritance mode, and the reference genome quality. This is because autopolyploids and allopolyploids have distinct reference genome structures (Kihara and Ono, 1926; Kyriakidou et al., 2018; Zhang et al., 2019). Ideally, autopolyploid assemblies are phased so all copies (i.e. haplotypes) of the genome are assembled. Assuming the autopolyploid has no preferential pairing amongst chromosomes (i.e. complete polysomic inheritance), all reads should be aligned to only one haplotype, similar to diploids, and the maximum allele dosage would be equal to the ploidy (Fig. 3B). In allopolyploids, the paternal and maternal haplotypes of each ancestral subgenome are assembled and reads are aligned to one haplotype of each subgenome simultaneously (Fig. 3A). Here, the maximum allele dosage would be the ploidy divided by the number of subgenomes. As an example, consider the allotetraploid switchgrass (*Panicum virgatum*) reference genome, which contains two phased subgenomes (Napier et al., 2022). Switchgrass is a mixed-ploidy species composed of tetraploids ($2n = 4x$) and octoploids ($2n = 8x$). As both subgenomes were successfully assembled, Napier et al. (2022) concurrently aligned reads to one haplotype of each subgenome and called genotypes for the tetraploid and octoploid samples as diploid (0, 1, 2) and tetraploid genotype values (0, 1, 2, 3, 4), respectively. If the switchgrass reference genome was not phased, the ploidy of each sample was unknown, or if it was unclear whether the species is allo- or autopolyploid, the correct allele dosage could not be determined. Unknown or incorrect allele dosage can result in the misestimation of allele frequencies and heterozygosity, similar to co-dominant markers like AFLPs (Dufresne et al., 2014).

2.5 Existing tools cannot account for further biological complexity

The reach of polyploid population and quantitative genetics is limited by further biological complexities. Commonly, populations may be mixed-ploidy, meaning they contain genotypes of varying ploidy levels (Kolář et al., 2017). Additionally, inheritance mode may vary along the genome (Allendorf et al., 2015). Variance in inheritance mode occurs because, following whole genome duplication, it is likely that all homologs pair together, and thus experience polysomic inheritance. However, over time, sequence divergence among homologous chromosomes may lead to preferential pairing and allow the return of disomic inheritance in some regions of the genome (Allendorf et al., 2015). In addition to mixed ploidy and inheritance mode, polyploid species may have multiple origins (Holloway et al., 2006; Soltis et al., 2009) and often hybridize (Alix et al., 2017), which makes population and quantitative genetics challenging. It is difficult to develop a variant calling pipeline that considers this complexity in a meaningful way while also producing genotypes that can be used in existing downstream tools. For example, existing software packages that estimate genotypes for mixed-ploidy populations require separate estimations for each ploidy (Blischak et al., 2018; Gerard et al., 2018; Clark et al., 2019; Van der Auwera and O'Connor, 2020; Cooke et al., 2021). In multi-sample variant calling, which incorporates information from multiple samples to improve genotype estimates, the separation of samples by ploidy reduces the utility and power of this approach (Liu et al., 2013). The mismapping of reads further exacerbates these issues in addition to creating false variants which could create false signals of allele sharing and alter patterns of genome-wide heterozygosity. Alternative approaches such as estimating genotypes at the same allele dosage for all cytotypes will result in underestimating heterozygous genotypes for higher ploidy levels and inaccurate allele frequency estimations.

3. Proposed solutions to incorporate polyploid complexity in variant calling

3.1 Balancing sequencing depth and precision may reduce sequencing costs

Careful experimental design, consideration of downstream analysis, and alternative genotyping approaches can be leveraged to reduce the cost of working with polyploid WGS data. Although a certain level of sequencing coverage is required to overcome allelic dosage uncertainty, high sequencing depth is not required for all analyses. Jighly (2022) argues that sequencing depth should be selected depending on the research question and analysis plan, in conjunction with the ploidy level, as sequencing depth has diminishing returns. Analyses that require the detection of low-frequency and rare variants, such as inferring novel alleles, will require a higher depth. In contrast, studies examining population structure and differentiation, which rely on common alleles to differentiate groups, may accommodate a lower sequencing depth. Therefore, considering the research question and analysis plan when determining the target coverage will prevent over-sequencing and extend a budget.

The increased allele dosage uncertainty that comes from low sequencing depth (<10X) can be partially mitigated by the use of genotype likelihoods (GLs) or continuous genotypes in place of categorical genotypes. A GL is the probability of the sequencing data given the possible genotypes. GLs can be directly used in some software or they can be used to infer genotypes. Polyploid-capable software such as

GATK, EBG, Updog, and polyRAD (Blischak et al., 2018; Gerard et al., 2018; Clark et al., 2019; Van der Auwera and O'Connor, 2020), infer categorical genotypes from GLs. Updog and polyRAD can also estimate continuous genotypes, which are continuous values of the likely allele count (Gerard et al., 2018; Clark et al., 2019; Njuguna et al., 2023). The combination of low-coverage data and GLs or continuous genotypes is becoming increasingly popular in large-scale studies due to its affordability (Korneliussen et al., 2014; Grandke et al., 2016; Batista et al., 2022). Further, GLs and continuous genotypes reduce allelic dosage uncertainty by incorporating genotyping certainty and may be beneficial in moderate or high-coverage sequence data. These alternative genotypes have been shown to provide more accurate estimates than categorical genotypes in numerous population and quantitative genetics analyses (Korneliussen et al., 2014; Grandke et al., 2016; Gerard, 2021b; Shastry et al., 2021; Batista et al., 2022; Rasmussen et al., 2024). Continuous genotypes can be easily integrated into existing software, however, software for downstream population and quantitative genetic analysis with polyploid GLs is still limited.

3.2 Alternative read alignment approaches, genotype callers, and variant filters may reduce errors caused by poor read mapping

Several strategies can be applied to reduce read mapping errors caused by homology, high polymorphism, or low reference genome quality throughout the variant calling pipeline. First, alternative alignment approaches could be applied to improve read mapping and assignment to subgenomes. For example, iterative read mapping is a promising strategy. Here, all reads are mapped to the reference genome but only reads that map to exactly one place in the genome (i.e. uniquely mapped reads) are retained. Then, a pseudo-reference genome is generated by replacing variable sites with the alternate alleles from the uniquely mapping reads, reads are re-mapped to the pseudo-reference, and, again, only uniquely mapped reads are retained (Rozowsky et al., 2011; Xu et al., 2020). When applied to maize whole-genome bisulfite sequencing data to reduce mapping bias, this approach was found to increase the detection of methylated cytosines by 5% (Xu et al., 2020). Alternatively, the software WASP alters the mapped reads, instead of the reference genome, to have the opposite allele. The altered reads are remapped and only kept if they map in the same location (van de Geijn et al., 2015). Both iterative read mapping approaches are particularly useful for reducing the number of multiply mapping reads and reducing false heterozygotes. Other alternative read mapping solutions have been developed specifically to identify subgenome differences in allopolyploids by either comparing polymorphisms to modern diploid progenitors (Mithani et al., 2013; Page et al., 2013; Peralta et al., 2013; Khan et al., 2016) or competitively mapping reads between subgenomes (Page and Udall, 2015). The former approach requires knowledge of the diploid progenitors and the ladder approach has limited benefits if both subgenomes of the allopolyploid are assembled. As a result, iterative read mapping is currently the most promising solution for improving read mapping.

Second, a genotype caller that considers allele bias and read-mapping errors could be used in addition to iterative read mapping to reduce the extent of false heterozygous or homozygous calls. The popular polyploid genotype caller Updog estimates the degree of allele bias simultaneously with genotype estimation (Gerard et al., 2018). No other polyploid genotype callers, to my knowledge, account for allele bias. Emerging solutions to reducing genotyping error from poor read mapping include the modification of variant calling algorithms developed for CNVs (Layer et al., 2014; Prodanov and Bansal, 2022) or ancient DNA (Günther and Nettelblad, 2019). For example, the software ancient DNA software, snpAD (Prüfer, 2018), iteratively estimates genotype probabilities and *r*, the frequency at which the sequences are sampled from the reference allele at heterozygous sites, to account for reference bias. Although snpAD is not currently able to estimate polyploid GLs, algorithms such as this have the potential to improve uncertainty in polyploid genotyping caused by poor read mapping.

Third, variant filters may be applied to exclude any remaining false-positive variants and genotyping errors caused by mismapped reads. Filters that have been used for this purpose discriminate variants by mapping quality, maximum coverage, and local linkage disequilibrium (Fig. 1E). I will briefly review these filters. To begin, mapping quality is a commonly applied 'hard' filter (Appendix S1) and is estimated as the phred-scaled probability a read is aligned to the wrong position. It is determined by the number of mismatches in the alignment while considering the quality of all other possible alignments (Li et al., 2008). Reads that map equally to multiple homologs (i.e. multiply mapping reads; Figure 2C) will have a mapping quality of zero and be removed in standard variant filtering pipelines. Typically, a mapping quality is applied to remove reads below a quality of 10 to 40 (Van der Auwera et al., 2013; Korneliussen et al., 2014; Puritz et al., 2014), which is equivalent to removing sites with greater than 0.01-10% probability of alignment error.

Exclusion of mismapped reads could also be accomplished using a maximum coverage filter. If reads improperly map to a given site, the site would have higher coverage than expected given the average genome-wide coverage (Fig. 2A). Applying this logic, maximum depth filters are commonly used to exclude false heterozygotes in repetitive regions of the genome (Li, 2014), but these are generally set too high to exclude reads mismapping in non-repetitive regions. In polyploid systems, this approach has been adopted to set a low per-site maximum depth threshold using models of expected read depth (Bohutínská et al., 2021; Korani et al., 2021; Phillips et al., 2023; Yu et al., 2023), although the efficacy of this filter and the best read depth model has not been determined.

A promising novel approach to exclude false-positive variants is to leverage the expectation that two true neighboring variants may have correlated allele frequencies within a population, known as local linkage disequilibrium (LD) (Bukowski et al., 2018). Variants in low LD with nearby variants would be excluded. This approach may also be useful in resolving the alignment of multiply-mapping reads by measuring local LD at each site the read is aligned to determine the most likely position, although this is likely computationally time-consuming and is yet to be tested in diploids or polyploids. LD estimates are biased by genotype uncertainty, which is exaggerated in polyploid genotypes, but this can be remedied with the recently developed R package ldsep that provides computationally efficient methods to estimate LD from diploid and polyploid GLs (Gerard, 2021a, b).

Other variant filters, such as the removal of loci with excess heterozygosity or departure from Hardy-Weinberg equilibrium (HWE), have also been explored for removing false-positive variants. If the mismapped reads carry the alternate allele, these filters may be able to remove false heterozygous sites (Keller et al., 2013; McKinney et al., 2017; Ahrens et al., 2020; Clark et al., 2022; Bohutínská et al., 2023). Researchers should exercise caution in applying filters that assume populations are at HWE because many biological factors, such as a non-panmictic population structure, small population sizes, and genetic drift, cause deviations from HWE (Pearman et al., 2022). Polyploidy itself deviates from diploid HWE therefore methods developed in Gerard (2022b) and Gerard (2023) should be used to properly account for unknown rates of double reduction (Gerard, 2022a).

3.3 Information on ploidy, chromosome inheritance mode, and reference quality can be integrated to determine allele dosage

Investment in the determination of ploidy level and inheritance mode of the reference genotype and sequenced genotypes towards the beginning of an experiment, although potentially time-intensive, is strongly recommended to identify the correct allele dosage. Traditionally, ploidy and inheritance mode have been determined using chromosome squashes (Goldblatt and Lowry, 2011), flow cytometry (Bennett and Leitch, 2011; Pellicer and Leitch, 2020) and fluorescence *in situ* hybridization (FISH), where fluorescent probes are used to label specific DNA sequences to identify and track chromosome pairings (Szadkowski et al., 2010; Chester et al., 2013; Parra-Nunez et al., 2020). Unfortunately, these approaches are time-intensive, require specialized equipment, and are an uncommon skill set. With the advent of next-generation sequencing, there has been a large research effort to determine ploidy from allele frequency distributions (Margarido and Heckerman, 2015; Augusto Corrêa Dos Santos et al., 2017; Weiß et al., 2018; Ranallo-Benavidez et al., 2020; Soraggi et al., 2022; Sun et al., 2023; Viruel et al., 2023; Gaynor et al., 2024). Sequence-based approaches have also begun to be explored for determining inheritance mode. One approach proposed by Scott et al. (2023) compares estimated allelic depth distributions to those expected under disomic and tetrasomic inheritance, although this approach is sensitive to demography. Other approaches include leveraging divergence among genes duplicated during whole genome duplication to detect windows of disomic or tetrasomic inheritance along the genome (Campbell et al., 2019; Scott et al., 2023) and the joint inference of inheritance mode and demography (Blischak et al., 2023; Roux et al., 2023) or genotypes (discussed in Section 3.4; Gerard et al., 2018; Clark et al., 2019). Sequence-based approaches are exceptionally promising for determining ploidy and inheritance mode in systems where flow cytometry and FISH are especially difficult or impossible, such as succulents and herbarium samples.

In cases where allele dosage cannot be determined because the ploidy and inheritance mode of the reference genotype is unknown, the reference scaffolds could be filtered to only one copy of syntenic scaffolds for read mapping. If the scaffolds can be assigned into subgenomes, such as in an allopolyploid, scaffolds would be filtered within each subgenome. This is a strategy applied in many systems with contig assemblies (Hellsten et al., 2013; Neale et al., 2022; Phillips et al., 2023). The risk of aligning to only a subset of scaffolds is that a large proportion of reads may not align and variants could be underdetected.

3.4 Current accepted practices for navigating polyploid data with additional biological complexity

Existing tools are limited in their ability to incorporate complexity such as mixed ploidy and inheritance mode, but variant calling pipelines have the potential to accommodate this additional axis of diversity in several ways. For datasets with mixed ploidy, the current best practice is to call genotypes separately for each cytotype, if using a joint genotyping approach (Napier et al., 2022; Bohutínská et al., 2023; De Luca et al., 2023). In cases where the secondary cytotype is rare or undersampled, it is advisable to exclude the minority cytotypes from the study because variability in downstream analyses attributable to cytotype differences may not be detectable with small sample sizes. If multiple cytotypes are included in the study, it should be noted that polyploid genotypes have inherently different expected variations in allele frequencies which can significantly impact downstream analyses (Faske, 2023). Similarly to mixed-ploidy analyses, allele dosage should be specified per-site in species with mixed inheritance modes. If the regions of the genome with polysomic inheritance are known, the per-site specification can be accomplished with any polyploid genotype caller, although this has rarely been applied outside of the Salmonids (Campbell et al., 2019). Alternatively, if polysomic regions are known, sites could be filtered to include only disomic or polysomic regions (Bourret et al., 2013). In the majority of cases, the rate of preferential pairing or the regions undergoing polysomic inheritance will be unknown. Here, the genotype calling software Updog (Gerard et al., 2018) and polyRAD (Clark et al., 2019) may be useful as their approaches determine inheritance mode during genotype estimation. Updog accomplishes this by simultaneously estimating genotypes and the rate of preferential pairing in a population, assuming bivalent pairing only. Comparatively, polyRAD determines inheritance mode by estimating genotypes for all possible user-specified genotypes and then uses a χ^2 statistic to determine the best genotype at each site. The polyRAD approach is particularly useful as it allows both ploidy and inheritance mode to vary among genotypes. There is no current best practice for mixed inheritance mode among these approaches, but they should be considered as even low rates of polysomic inheritance can affect allele frequencies across subgenomes (Meirmans and Van Tienderen, 2013). Consequently, careful consideration is required when analyzing populations with biological complexity beyond polyploidy.

4. Conclusions

Complex polyploid biology may produce errors in read mapping, variant calling, and genotyping. The extent of error often depends on the quality of the reference genome and biological reasons like the age of the polyploidization event, extent of fractionation, divergence between parental genomes, and strength of selection at a given locus. As such, bioinformatic solutions can be selectively applied to resolve sources of error prevalent in a given polyploid system. In Figure 1, I summarize where existing solutions can be integrated into a standard variant calling pipeline. The study of polyploid genomes is a growing field and, as such, there may be additional solutions in active development.

Further improvements to variant calling in polyploids will require focused research in three primary areas: evaluation of variant filters, development of downstream software that incorporates genotype uncertainty, and high-throughput estimation of ploidy and inheritance mode. First, empirical studies evaluating the efficacy of variant filters are needed to understand when their application is appropriate and which thresholds are effective. It is equally as important to set a threshold that excludes low-quality variants while also not over-filtering the data, as variant classes important in downstream analyses may be unintentionally excluded (Linck and Battey, 2019; Pearman et al., 2022). Second, continued development of population and quantitative genetics software that utilize GLs is needed (Korneliussen et al., 2014; Grandke et al., 2016; Gerard, 2021b; Shastry et al., 2021; Batista et al., 2022; Rasmussen et al., 2024). The adoption of GLs to reduce sequencing costs is likely to be limited until more user-friendly software becomes available. Theory and tools are also lacking for the analysis of mixed-ploidy and mixed-inheritance mode datasets. Third, continued development of methods for high throughput estimation of ploidy and inheritance mode is greatly needed. While there has been substantial development in this area (see Section 3.3), the majority of approaches still necessitate ample ground truthing (Gaynor et al., 2024).

Emerging technologies may have the potential to improve variant detection. Long-read sequencing data overcomes many read mapping challenges as the extended read length increases the information available to determine the best alignment (Chen et al., 2024). Similar to short-read sequencing, long-read sequencing is increasingly cost-effective and accurate (De Coster et al., 2021; Kim et al., 2024). Additionally, pan-genomic approaches, such as haplotype graphs and sequence variation groups, have recently been applied in polyploid systems to detect a diversity of SVs as well as multiallelic sites (Gordon et al., 2020; Bayer et al., 2021; Della Coletta et al., 2021; Lovell et al., 2021; Wang et al., 2022). The adoption of the variant calling practices reviewed here, continued investment in the assembly of polyploid reference genomes, and early adoption of novel genomic tools will enhance contemporary population and quantitative genetics studies in polyploids.

Author contributions

The author was solely responsible for the conceptualization, research, and writing of the entire manuscript.

Acknowledgments

The author would like to thank Jeffrey Ross-Ibarra, Elli Cryan, Natasha Dhamrait, Regina Fairbanks, Samantha Snodgrass, Tyler Kent, Elisabeth Forrestel, Jennifer Gremer, Michelle Gaynor, and Trevor Faske for their feedback on this manuscript. The ForBio "Population genetics of polyploids, from theory to practice" workshop provided a useful space to discuss these ideas. This project was funded by the National Science Foundation (NSF) grant numbers 1822330 and 1934384.

Data availability statement

No datasets were generated or analyzed for this study.

Supporting information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Appendix 1 - A brief overview of variant calling

References

Ahrens, C. W., E. A. James, A. D. Miller, F. Scott, N. C. Aitken, A. W. Jones, P. Lu-Irving, et al. 2020. Spatial, climate and ploidy factors drive genomic diversity and resilience in the widespread grass *Themeda triandra*. *Molecular Ecology* 29: 3872–3888.

Alix, K., P. R. Gérard, T. Schwarzacher, and J. S. P. Heslop-Harrison. 2017. Polyploidy and interspecific hybridization: Partners for adaptation, speciation and evolution in plants. *Annals of Botany* 120: 183–194.

Allendorf, F. W., S. Bassham, W. A. Cresko, M. T. Limborg, L. W. Seeb, and J. E. Seeb. 2015. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *The Journal of Heredity* 106: 217–227.

Augusto Corrêa Dos Santos, R., G. H. Goldman, and D. M. Riaño-Pachón. 2017. ploidyNGS: Visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics* 33: 2575–2576.

Baduel, P., L. Quadrana, B. Hunter, K. Bomblies, and V. Colot. 2019. Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. *Nature Communications* 10: 5818.

Batista, L. G., V. H. Mello, A. P. Souza, and G. R. A. Margarido. 2022. Genomic prediction with allele dosage information in highly polyploid species. *Theoretical and Applied Genetics* 135: 723–739.

Bayer, P. E., A. Scheben, A. A. Golicz, Y. Yuan, S. Faure, H. Lee, H. S. Chawla, et al. 2021. Modelling of gene loss propensity in the pangenomes of three *Brassica* species suggests different mechanisms between polyploids and diploids. *Plant Biotechnology Journal* 19: 2488–2500.

de Bem Oliveira, I., R. R. Amadeu, L. F. V. Ferrão, and P. R. Muñoz. 2020. Optimizing whole-genomic prediction for autotetraploid blueberry breeding. *Heredity* 125: 437–448.

Bennett, M. D., and I. J. Leitch. 2011. Nuclear DNA amounts in angiosperms: Targets, trends and tomorrow. *Annals of Botany* 107: 467–590.

Beric, A., M. E. Mabry, A. E. Harkess, J. Brose, M. E. Schranz, G. C. Conant, P. P. Edger, et al. 2021. Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales). *G3* 11.

Bertioli, D. J., J. Jenkins, J. Clevenger, O. Dudchenko, D. Gao, G. Seijo, S. C. M. Leal-Bertioli, et al. 2019. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nature Genetics* 51: 877–884.

Blischak, P. D., L. S. Kubatko, and A. D. Wolfe. 2018. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics* 34: 407–415.

Blischak, P. D., M. Sajan, M. S. Barker, and R. N. Gutenkunst. 2023. Demographic history inference and the polyploid continuum. *Genetics* 224.

Bohutínská, M., M. Alston, P. Monnahan, T. Mandáková, S. Bray, P. Paajanen, F. Kolář, and L. Yant. 2021. Novelty and convergence in adaptation to whole genome duplication. *Molecular Biology and Evolution* 38: 3910–3924.

Bohutínská, M., J. Vlček, P. Monnahan, and F. Kolář. 2023. Population genomic analysis of diploid-autopolyploid species. *Methods in Molecular Biology* 2545: 297–324.

Bourret, V., M. P. Kent, C. R. Primmer, A. Vasemägi, S. Karlsson, K. Hindar, P. McGinnity, et al. 2013. SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology* 22: 532–551.

Brandt, D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, and D. Meyer. 2015. Mapping bias overestimates reference allele frequencies at the *HLA* genes in the 1000 Genomes Project Phase I data. *G3* 5: 931–941.

Bukowski, R., X. Guo, Y. Lu, C. Zou, B. He, Z. Rong, B. Wang, et al. 2018. Construction of the third-generation *Zea mays* haplotype map. *GigaScience* 7: 1–12.

Campbell, M. A., M. C. Hale, G. J. McKinney, K. M. Nichols, and D. E. Pearse. 2019. Long-term conservation of ohnologs through partial tetrasomy following whole-genome duplication in Salmonidae. *G3* 9: 2017–2028.

Chen, X., C. Tong, X. Zhang, A. Song, M. Hu, W. Dong, F. Chen, et al. 2021. A high-quality *Brassica napus* genome reveals expansion of transposable elements, subgenome evolution and disease resistance. *Plant Biotechnology Journal* 19: 615–630.

Chen, Y., W. Wang, Z. Yang, H. Peng, Z. Ni, Q. Sun, and W. Guo. 2024. Innovative computational tools provide new insights into the polyploid wheat genome. *aBIOTECH*.

Chen, Z. J., A. Sreedasyam, A. Ando, Q. Song, L. M. De Santiago, A. M. Hulse-Kemp, M. Ding, et al. 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nature Genetics* 52: 525–533.

Chester, M., M. J. Lipman, J. P. Gallagher, P. S. Soltis, and D. E. Soltis. 2013. An assessment of karyotype restructuring in the neoallotetraploid *Tragopogon miscellus* (Asteraceae). *Chromosome Research* 21: 75–85.

Clark, L. V., A. E. Lipka, and E. J. Sacks. 2019. polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3* 9: 663–673.

Clark, L. V., W. Mays, A. E. Lipka, and E. J. Sacks. 2022. A population-level statistic for assessing Mendelian behavior of genotyping-by-sequencing data from highly duplicated genomes. *BMC Bioinformatics* 23: 101.

Cooke, D. P., D. C. Wedge, and G. Lunter. 2021. A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology* 39: 885–892.

Cooke, D. P., D. C. Wedge, and G. Lunter. 2022. Benchmarking small-variant genotyping in polyploids. *Genome Research* 32: 403–408.

David, K. T. 2022. Global gradients in the distribution of animal polyploids. *Proceedings of the National Academy of Sciences of the United States of America* 119: e2214070119.

De Coster, W., M. H. Weissensteiner, and F. J. Sedlazeck. 2021. Towards population-scale long-read sequencing. *Nature Reviews Genetics* 22: 572–587.

Della Coletta, R., Y. Qiu, S. Ou, M. B. Hufford, and C. N. Hirsch. 2021. How the pan-genome is changing crop genomics and improvement. *Genome Biology* 22: 3.

De Luca, D., E. Del Guacchio, P. Cennamo, L. Paino, and P. Caputo. 2023. Genotyping-by-sequencing provides new genetic and taxonomic insights in the critical group of *Centaurea tenorei*. *Frontiers in Plant Science* 14: 1130889.

De Summa, S., G. Malerba, R. Pinto, A. Mori, V. Mijatovic, and S. Tommasi. 2017. GATK hard filtering: Tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinformatics* 18: 119.

Dufresne, F., M. Stift, R. Vergilino, and B. K. Mable. 2014. Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology* 23: 40–69.

Ebadi, M., Q. Bafort, E. Mizrachi, P. Audenaert, P. Simoens, M. Van Montagu, D. Bonte, and Y. Van de Peer. 2023. The duplication of genomes and genetic networks and its potential for evolutionary adaptation and survival during environmental turmoil. *Proceedings of the National Academy of Sciences of the United States of America* 120: e2307289120.

Edger, P. P., T. J. Poorten, R. VanBuren, M. A. Hardigan, M. Colle, M. R. McKain, R. D. Smith, et al. 2019. Origin and evolution of the octoploid strawberry genome. *Nature Genetics* 51: 541–547.

Emery, M., M. M. S. Willis, Y. Hao, K. Barry, K. Oakgrove, Y. Peng, J. Schmutz, et al. 2018. Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS Genetics* 14: e1007267.

Ewing, A. D. 2015. Transposable element detection from whole genome sequence data. *Mobile DNA* 6: 24.

Faske, T. 2023. 2 does not equal 4: Variance dissimilarities in mixed-ploidy genomic data cause irregular patterns in PCA and other clustering analyses. *In Abstracts***,** Botany, Boise, ID.

Formenti, G., K. Theissinger, C. Fernandes, I. Bista, A. Bombarely, C. Bleidorn, C. Ciofi, et al. 2022. The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution* 37: 197–202.

Fuentes-Pardo, A. P., and D. E. Ruzzante. 2017. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology* 26: 5369–5406.

Gaut, B. S., and J. F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences of the United States of America* 94: 6809–6814.

Gaynor, M. L., J. B. Landis, T. K. O'Connor, R. G. Laport, J. J. Doyle, D. E. Soltis, J. M. Ponciano, and P. S. Soltis. 2024. nQuack: An R package for predicting ploidal level from sequence data using site-based heterozygosity. *bioRxiv*: 2024.02.12.579894.

van de Geijn, B., G. McVicker, Y. Gilad, and J. K. Pritchard. 2015. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods* 12: 1061–1063.

Gerard, D. 2023. Bayesian tests for random mating in polyploids. *Molecular Ecology Resources* 23: 1812–1822.

Gerard, D. 2022a. Comment on three papers about Hardy-Weinberg equilibrium tests in autopolyploids. *Frontiers in Genetics* 13: 1027209.

Gerard, D. 2022b. Double reduction estimation and equilibrium tests in natural autopolyploid populations. *Biometrics*.

Gerard, D. 2021a. Pairwise linkage disequilibrium estimation for polyploids. *Molecular Ecology Resources* 21: 1230–1242.

Gerard, D. 2021b. Scalable bias-corrected linkage disequilibrium estimation under genotype uncertainty. *Heredity* 127: 357–362.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens. 2018. Genotyping polyploids from messy sequencing data. *Genetics* 210: 789–807.

Gladman, N., S. Goodwin, K. Chougule, W. Richard McCombie, and D. Ware. 2023. Era of gapless plant genomes: Innovations in sequencing and mapping technologies revolutionize genomics and breeding. *Current Opinion in Biotechnology* 79: 102886.

Goeckeritz, C. Z., K. E. Rhoades, K. L. Childs, A. F. Iezzoni, R. VanBuren, and C. A. Hollender. 2023. Genome of tetraploid sour cherry (*Prunus cerasus* L.) 'Montmorency' identifies three distinct ancestral *Prunus* genomes. *Horticulture Research* 10: uhad097.

Goldblatt, P., and P. P. Lowry. 2011. The Index to Plant Chromosome Numbers (IPCN): Three decades of publication by the Missouri Botanical Garden come to an end. *Annals of the Missouri Botanical Garden* 98: 226–227.

Gordon, S. P., B. Contreras-Moreira, J. J. Levy, A. Djamei, A. Czedik-Eysenberg, V. S. Tartaglio, A. Session, et al. 2020. Gradual polyploid genome evolution revealed by pan-genomic analysis of *Brachypodium hybridum* and its diploid progenitors. *Nature Communications* 11: 3670.

Grandke, F., P. Singh, H. C. M. Heuven, J. R. de Haan, and D. Metzler. 2016. Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: A comparative study in hexaploid *Chrysanthemum*. *BMC Genomics* 17: 672.

Gui, S., W. Wei, C. Jiang, J. Luo, L. Chen, S. Wu, W. Li, et al. 2022. A pan-*Zea* genome map for enhancing maize improvement. *Genome Biology* 23: 178.

Günther, T., and C. Nettelblad. 2019. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genetics* 15: e1008302.

Haldane, J. B. S. 1933. The part played by recurrent mutation in evolution. *The American Naturalist* 67: 5–19.

Hämälä, T., C. Moore, L. Cowan, M. Carlile, D. Gopaulchan, M. K. Brandrud, S. Birkeland, et al. 2023. Impact of whole-genome duplications on structural variant evolution in the plant genus *Cochlearia*. *bioRxiv*: 2023.09.29.560073.

Hellsten, U., K. M. Wright, J. Jenkins, S. Shu, Y. Yuan, S. R. Wessler, J. Schmutz, et al. 2013. Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 110: 19478–19482.

Hollister, J. D., B. J. Arnold, E. Svedin, K. S. Xue, B. P. Dilkes, and K. Bomblies. 2012. Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. *PLoS Genetics* 8: e1003093.

Holloway, A. K., D. C. Cannatella, H. C. Gerhardt, and D. M. Hillis. 2006. Polyploids with different origins and ancestors form a single sexual polyploid species. *The American Naturalist* 167: E88–101.

Hotaling, S., E. R. Wilcox, J. Heckenhauer, R. J. Stewart, and P. B. Frandsen. 2023. Highly accurate long reads are crucial for realizing the potential of biodiversity genomics. *BMC Genomics* 24: 117.

Jighly, A. 2022. When do autopolyploids need poly-sequencing data? *Molecular Ecology* 31: 1021–1027.

Keller, I., C. E. Wagner, L. Greuter, S. Mwaiko, O. M. Selz, A. Sivasundar, S. Wittwer, and O. Seehausen. 2013. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology* 22: 2848–2863.

Khan, A., E. J. Belfield, N. P. Harberd, and A. Mithani. 2016. HANDS2: Accurate assignment of homoeallelic base-identity in allopolyploids despite missing data. *Scientific Reports* 6: 29234.

Kihara, H., and T. Ono. 1926. Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. *Zeitschrift für Zellforschung und Mikroskopische Anatomie* 4: 475–481.

Kim, C., M. Pongpanich, and T. Porntaveetus. 2024. Unraveling metagenomics through long-read sequencing: A comprehensive review. *Journal of Translational Medicine* 22: 111.

Kolář, F., M. Čertner, J. Suda, P. Schönswetter, and B. C. Husband. 2017. Mixed-ploidy species: Progress and opportunities in polyploid research. *Trends in Plant Science* 22: 1041–1055.

Korani, W., D. O'Connor, Y. Chu, C. Chavarro, C. Ballen, B. Guo, P. Ozias-Akins, et al. 2021. De novo QTL-seq identifies loci linked to blanchability in peanut (*Arachis hypogaea*) and refines previously identified QTL with low coverage sequence. *Agronomy* 11: 2201.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15: 356.

Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* 20: 117.

Kyriakidou, M., H. H. Tai, N. L. Anglin, D. Ellis, and M. V. Strömvik. 2018. Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science* 9: 1660.

Layer, R. M., C. Chiang, A. R. Quinlan, and I. M. Hall. 2014. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology* 15: R84.

Li, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30: 2843–2851.

Li, H., J. Ruan, and R. Durbin. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18: 1851–1858.

Linck, E., and C. J. Battey. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources* 19: 639–647.

Lisch, D. 2013. How important are transposons for plant evolution? *Nature Reviews Genetics* 14: 49–61.

Liu, X., S. Han, Z. Wang, J. Gelernter, and B.-Z. Yang. 2013. Variant callers for next-generation sequencing data: A comparison study. *PloS One* 8: e75619.

Lou, R. N., A. Jacobs, A. P. Wilder, and N. O. Therkildsen. 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology* 30: 5966–5993.

Lou, R. N., and N. O. Therkildsen. 2022. Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation. *Molecular Ecology Resources* 22: 1678–1692.

Lovell, J. T., A. H. MacQueen, S. Mamidi, J. Bonnette, J. Jenkins, J. D. Napier, A. Sreedasyam, et al. 2021. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* 590: 438–444.

Lowry, D. B., S. Hoban, J. L. Kelley, K. E. Lotterhos, L. K. Reed, M. F. Antolin, and A. Storfer. 2017. Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources* 17: 142–152.

Mahmoud, M., N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck. 2019. Structural variant calling: The long and the short of it. *Genome Biology* 20: 246.

Margarido, G. R. A., and D. Heckerman. 2015. ConPADE: Genome assembly ploidy estimation from next-generation sequencing data. *PLoS Computational Biology* 11: e1004229.

Mason, A. S., and J. F. Wendel. 2020. Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Frontiers in Genetics* 11: 1014.

Ma, X.-F., and J. P. Gustafson. 2005. Genome evolution of allopolyploids: A process of cytological and genetic diploidization. *Cytogenetic and Genome Research* 109: 236–249.

McClintock, B. 1984. The significance of responses of the genome to challenge. *Science* 226: 792–801.

McKinney, G. J., R. K. Waples, L. W. Seeb, and J. E. Seeb. 2017. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources* 17: 656–669.

Meirmans, P. G., and P. H. Van Tienderen. 2013. The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity* 110: 131–137.

Mithani, A., E. J. Belfield, C. Brown, C. Jiang, L. J. Leach, and N. P. Harberd. 2013. HANDS: A tool for genome-wide discovery of subgenome-specific base-identity in polyploids. *BMC Genomics* 14: 653.

Muir, P., S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, et al. 2016. The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biology* 17: 53.

Musich, R., L. Cadle-Davidson, and M. V. Osier. 2021. Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Frontiers in Plant Science* 12: 657240.

Napier, J. D., P. P. Grabowski, J. T. Lovell, J. Bonnette, S. Mamidi, M. J. Gomez-Hughes, A. VanWallendael, et al. 2022. A generalist-specialist trade-off between switchgrass cytotypes impacts climate adaptation and geographic range. *Proceedings of the National Academy of Sciences of the United States of America* 119: e2118879119.

Neale, D. B., A. V. Zimin, S. Zaman, A. D. Scott, B. Shrestha, R. E. Workman, D. Puiu, et al. 2022. Assembled and annotated 26.5 Gbp coast redwood genome: A resource for estimating evolutionary adaptive potential and investigating hexaploid origin. *G3* 12.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* 12: 443–451.

Njuguna, J. N., L. V. Clark, A. E. Lipka, K. G. Anzoua, L. Bagmet, P. Chebukin, M. S. Dwiyanti, et al. 2023. Impact of genotype-calling methodologies on genome-wide association and genomic prediction in polyploids. *The Plant Genome* 16: e20401.

O'Leary, S. J., J. B. Puritz, S. C. Willis, C. M. Hollenbeck, and D. S. Portnoy. 2018. These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.

Otto, S. P., and J. Whitton. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* 34: 401–437.

Page, J. T., A. R. Gingle, and J. A. Udall. 2013. PolyCat: A resource for genome categorization of sequencing reads from allopolyploid organisms. *G3* 3: 517–525.

Page, J. T., and J. A. Udall. 2015. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. *BMC Genetics* 16 Suppl 2: S4.

Parra-Nunez, P., M. Pradillo, and J. L. Santos. 2020. How to perform an accurate analysis of metaphase I chromosome configurations in autopolyploids of *Arabidopsis thaliana*. *In* M. Pradillo, and S. Heckmann [eds.], Plant Meiosis: Methods and Protocols, 25–36. Springer New York, New York, NY.

Pearman, W. S., L. Urban, and A. Alexander. 2022. Commonly used Hardy-Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. *Molecular Ecology Resources* 22: 2599–2613.

Pellicer, J., and I. J. Leitch. 2020. The Plant DNA C-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. *The New Phytologist* 226: 301–305.

Peralta, M., M.-C. Combes, A. Cenci, P. Lashermes, and A. Dereeper. 2013. SNiPloid: A Utility to Exploit High-Throughput SNP Data Derived from RNA-Seq in Allopolyploid Species. *International Journal of Plant Genomics* 2013: 890123.

Phillips, A. R., A. S. Seetharam, P. S. Albert, T. AuBuchon-Elder, J. A. Birchler, E. S. Buckler, L. J. Gillespie, et al. 2023. A happy accident: A novel turfgrass reference genome. *G3* 13.

Poland, J. A., and T. W. Rife. 2012. Genotyping−by−sequencing for plant breeding and genetics. *The Plant Genome* 5: 92–102.

Prodanov, T., and V. Bansal. 2022. Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing. *Nature Communications* 13: 3221.

Prüfer, K. 2018. snpAD: An ancient DNA genotype caller. *Bioinformatics* 34: 4165–4171.

Puritz, J. B., C. M. Hollenbeck, and J. R. Gold. 2014. dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* 2: e431.

Ramakrishnan, M., L. Satish, A. Sharma, K. Kurungara Vinod, A. Emamverdian, M. Zhou, and Q. Wei. 2022. Transposable elements in plants: Recent advancements, tools and prospects. *Plant Molecular Biology Reporter* 40: 628–645.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11: 1432.

Rasmussen, M. S., C. Wiuf, and A. Albrechtsen. 2024. Inferring drift, genetic differentiation, and admixture graphs from low-depth sequencing data. *bioRxiv*: 2024.01.29.577762.

Román-Palacios, C., C. A. Medina, S. H. Zhan, and M. S. Barker. 2021. Animal chromosome counts reveal a similar range of chromosome numbers but with less polyploidy in animals compared to flowering plants. *Journal of Evolutionary Biology* 34: 1333–1339.

Roux, C., X. Vekemans, and J. Pannell. 2023. Inferring the demographic history and inheritance mode of tetraploid species using ABC. *In* Y. Van de Peer [ed.], Polyploidy: Methods and Protocols, 325–348. Springer US, New York, NY.

Rozowsky, J., A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, et al. 2011. AlleleSeq: Analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology* 7: 522.

Scott, A. D., J. D. Van de Velde, and P. Y. Novikova. 2023. Inference of polyploid origin and inheritance mode from population genomic data. *In* Y. Van de Peer [ed.], Polyploidy: Methods and Protocols, 279–295. Springer US, New York, NY.

Session, A. M., and D. S. Rokhsar. 2023. Transposon signatures of allopolyploid genome evolution. *Nature Communications* 14: 3180.

Shastry, V., P. E. Adams, D. Lindtke, E. G. Mandeville, T. L. Parchman, Z. Gompert, and C. A. Buerkle. 2021. Model-based genotype and ancestry estimation for potential hybrids with mixed-ploidy. *Molecular Ecology Resources* 21: 1434–1451.

Soltis, D. E., R. J. A. Buggs, W. B. Barbazuk, P. S. Schnable, and P. S. Soltis. 2009. On the origins of species: Does evolution repeat itself in polyploid populations of independent origin? *Cold Spring Harbor Symposia on Quantitative Biology* 74: 215–223.

Soraggi, S., J. Rhodes, I. Altinkaya, O. Tarrant, F. Balloux, M. C. Fisher, and M. Fumagalli. 2022. HMMploidy: Inference of ploidy levels from short-read sequencing data. *Peer Community Journal* 2.

Stebbins, G. L., Jr. 1947. Types of polyploids; their classification and significance. *Advances in Genetics* 1: 403–429.

Stift, M., C. Berenos, P. Kuperus, and P. H. van Tienderen. 2008. Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: A general procedure applied to *Rorippa* (yellow cress) microsatellite data. *Genetics* 179: 2113–2123.

Sun, M., E. Pang, W.-N. Bai, D.-Y. Zhang, and K. Lin. 2023. ploidyfrost: Reference-free estimation of ploidy level from whole genome sequencing data based on de Bruijn graphs. *Molecular Ecology Resources* 23: 499–510.

Sun, S., Y. Zhou, J. Chen, J. Shi, H. Zhao, H. Zhao, W. Song, et al. 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nature Genetics* 50: 1289–1295.

Szadkowski, E., F. Eber, V. Huteau, M. Lodé, C. Huneau, H. Belcram, O. Coriton, et al. 2010. The first meiosis of resynthesized *Brassica napus*, a genome blender. *New Phytologist* 186: 102–112.

Therkildsen, N. O., and S. R. Palumbi. 2017. Practical low-coverage genome wide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources* 17: 194–208.

Tiffin, P., and J. Ross-Ibarra. 2014. Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution* 29: 673–680.

Udall, J. A., and J. F. Wendel. 2006. Polyploidy and crop improvement. *Crop Science* 46: S–3–S–14.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, et al. 2013. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* 43: 11.10.1–11.10.33.

Van der Auwera, G. A., and B. D. O'Connor. 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, Incorporated.

Viruel, J., O. Hidalgo, L. Pokorny, F. Forest, B. Gravendeel, P. Wilkin, and I. J. Leitch. 2023. A bioinformatic pipeline to estimate ploidy level from target capture sequence data obtained from herbarium specimens. *Methods in Molecular Biology* 2672: 115–126.

Wang, M., J. Li, Z. Qi, Y. Long, L. Pei, X. Huang, C. E. Grover, et al. 2022. Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. *Nature Genetics* 54: 1959–1971.

Weiß, C. L., M. Pais, L. M. Cano, S. Kamoun, and H. A. Burbano. 2018. nQuire: A statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* 19: 122.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America* 106: 13875–13879.

Xu, G., J. Lyu, Q. Li, H. Liu, D. Wang, M. Zhang, N. M. Springer, et al. 2020. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nature Communications* 11: 5539.

Yu, R.-M., N. Zhang, B.-W. Zhang, Y. Liang, X.-X. Pang, L. Cao, Y.-D. Chen, et al. 2023. Genomic insights into biased allele loss and increased gene numbers after genome duplication in autotetraploid *Cyclocarya paliurus*. *BMC Biology* 21: 168.

Zack, T. I., S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, et al. 2013. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics* 45: 1134–1140.

Zhang, X., S. Zhang, Q. Zhao, R. Ming, and H. Tang. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* 5: 833–845.

Zhou, Y., J. Zhang, X. Xiong, Z.-M. Cheng, and F. Chen. 2022. *De novo* assembly of plant complete genomes. *Tropical Plants* 1: 1–8.

Figures

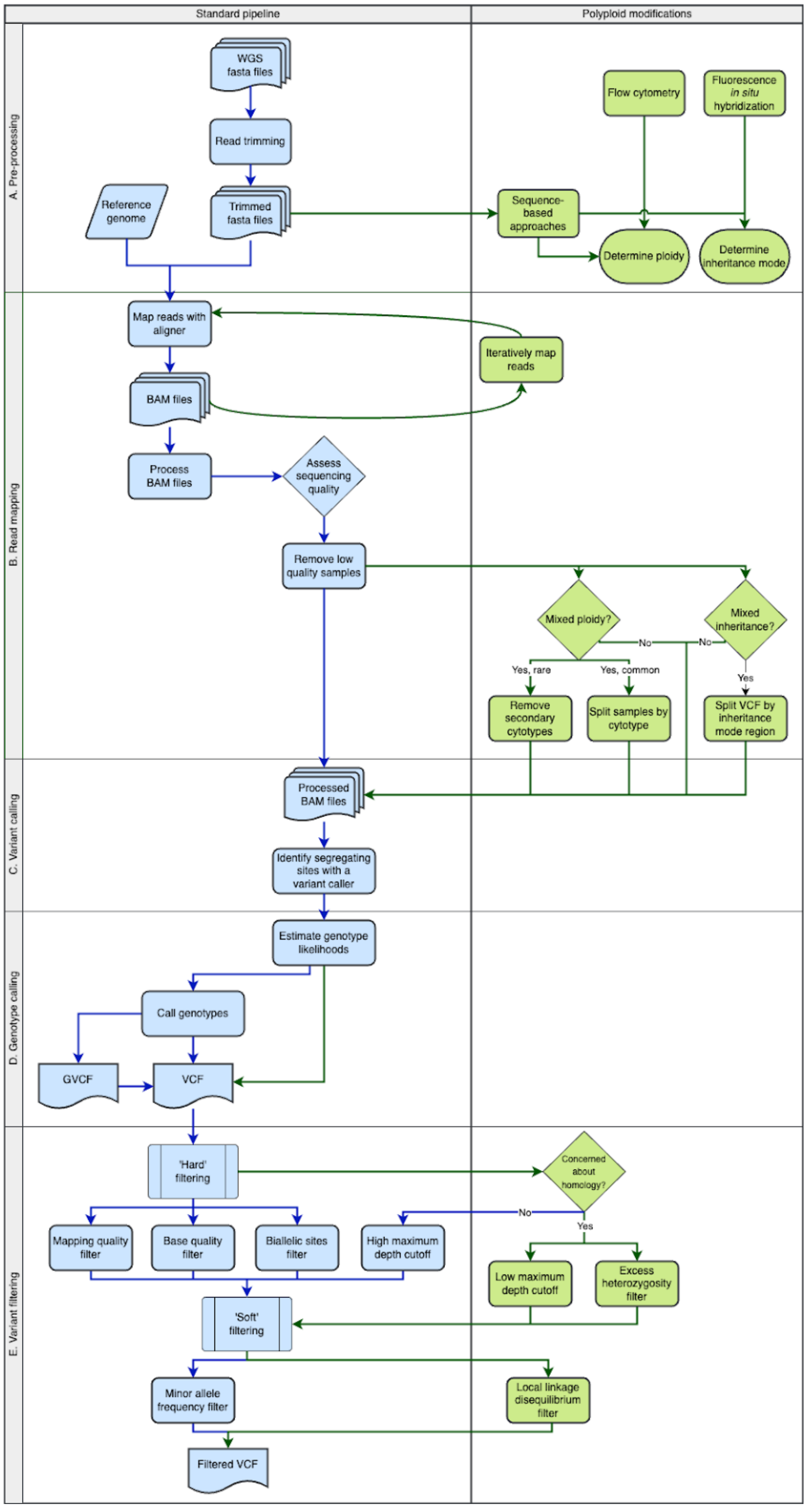
Standard pipeline
Polyploid modifications
A. Pre-processing
WGS fasta files
Read trimming
Reference genome
Trimmed fasta files
Flow cytometry
Fluorescence *in situ* hybridization
Sequence-based approaches
Determine ploidy
Determine inheritance mode
B. Read mapping
Map reads with aligner
Iteratively map reads
BAM files
Process BAM files
Assess sequencing quality
Remove low quality samples
Mixed ploidy?
Mixed inheritance?
No
No
Yes, rare
Yes, common
Yes
Remove secondary cytotypes
Split samples by cytotype
Split VCF by inheritance mode region
C. Variant calling
Processed BAM files
Identify segregating sites with a variant caller
D. Genotype calling
Estimate genotype likelihoods
Call genotypes
GVCF
VCF
E. Variant filtering
'Hard' filtering
Concerned about homology?
No
Yes
Mapping quality filter
Base quality filter
Biallelic sites filter
High maximum depth cutoff
Low maximum depth cutoff
Excess heterozygosity filter
'Soft' filtering
Minor allele frequency filter
Local linkage disequilibrium filter
Filtered VCF

Figure 1. A standard variant calling pipeline (blue) can be adapted for polyploid systems (modifications in green). (A) Before beginning variant calling, raw sequence data may need trimming to remove adapters and low-quality bases. An effort should be made to determine the ploidy and chromosome inheritance mode of the sequenced genotypes, as this information will be incorporated later in the pipeline. Multiple approaches can be used to determine ploidy and inheritance mode depending on the researcher's skillset. (B) Reads are mapped to the reference genome using an aligner. Binary alignment maps (BAMs) are output from the aligners and processed by adding read groups, removing duplicate reads, and then sorting. Sequencing and alignment quality are assessed so low-quality samples may be identified and removed before variant calling. Samples should be split by ploidy and regions by inheritance mode, if necessary, at this stage. (C) Variants are called (D) and then genotype likelihoods and genotypes are estimated. Variant calling and genotyping are often completed using the same software but can be run separately. Genotype calling can be skipped if genotype likelihoods will be used downstream. A variant call file (VCF) is output if invariant sites are discarded, otherwise the output is a genomic variant call file (GVCF). (E) Variants are filtered first by removing low-quality sites (i.e. hard filtering). Then, variants are filtered to prioritize variants specific to downstream analyses (i.e. soft filtering). A more detailed description of the standard pipeline, including useful polyploid aligners and genotype calling software, is provided in Appendix S1.

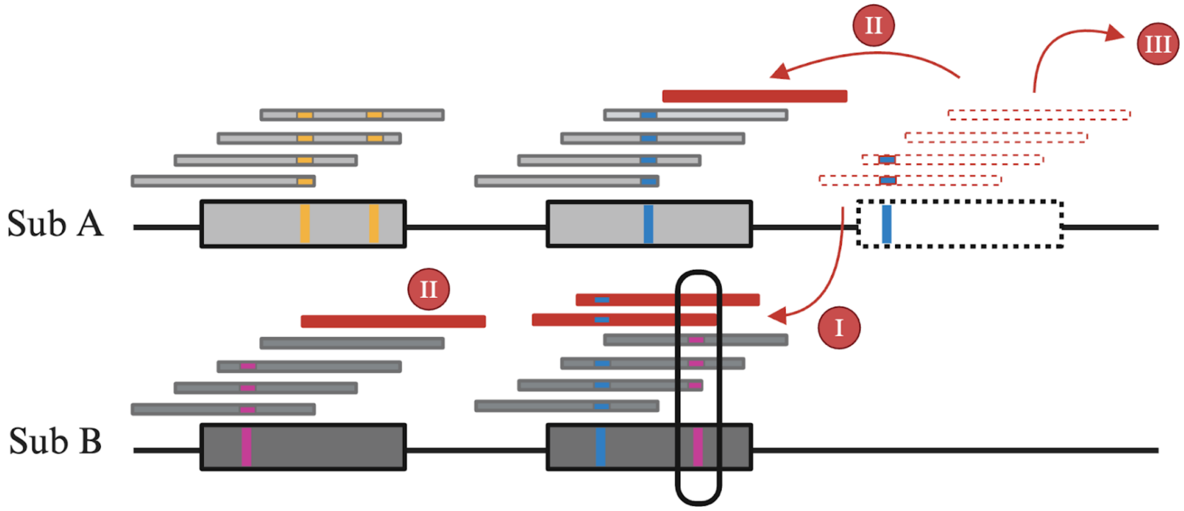

Figure 2. A syntenic block between subgenome A and subgenome B in an allotetraploid is depicted. This region in subgenome A contains three genes (light gray) while subgenome B (dark gray) contains two. The genes contain one or two segregating sites, with alleles depicted as yellow, pink, and blue. The assembly of subgenome A is incomplete, missing the farthest right gene (dashed line). Reads that should have aligned to the missing gene (red reads) instead may **(I)** align to a homolog in subgenome B resulting in a false heterozygote call, **(II)** map equally to other homologs within or across subgenomes, or **(III)** fail to align. This figure was created with BioRender.com.

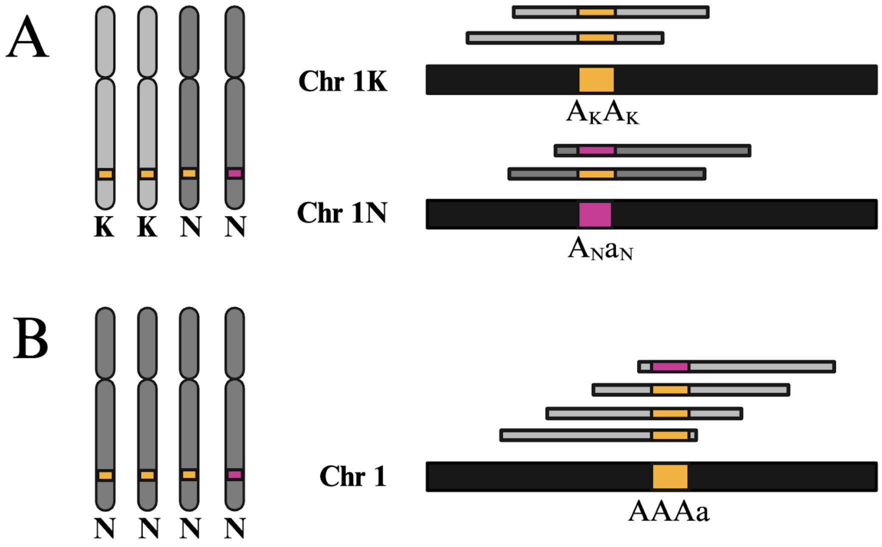

Figure 3. Read mapping and the called allele dosage in allo- and autopolyploids differs due to the structure of the reference genome. Reads (gray) are shown aligning the reference genome (black) with alleles for the focal variant in pink or yellow. **(A)** In an allotetraploid with two subgenomes (subgenome K in light gray and subgenome N in dark gray), reads are mapped to one haplotype of each parental subgenome, and diploid genotypes are called. **(B)** In an autotetraploid with no preferential pairing, all reads are mapped to a single haplotype. Here, reads are aligned to a haplotype carrying the yellow A allele at the focal variant.

Appendix S1

A brief overview of variant calling

In diploid and polyploid systems, variant calling involves a series of qualitative decisions that depend on the biology of the study system and data quality. A variant calling pipeline, as described here, includes the alignment of reads to the reference genome, variant calling, genotype estimation, and variant filtering. Consideration of ploidy in downstream analyses has been well-reviewed elsewhere (Dufresne et al., 2014; Meirmans et al., 2018; Ackiss and Balao, 2020; Bohutínská et al., 2023). Here, I aim to provide an overview of a general variant calling pipeline to support discussions of where this pipeline may be improved for polyploid systems. I provide citations for commonly used software where relevant.

To begin, reads are mapped to a reference genome using a short-read aligner to generate the sequence alignment maps (SAMs) or binary alignment maps (BAMs). The aligner is selected depending on the read length, sequencing method, and divergence of the sequenced sample from the reference genome (Altmann et al., 2012; Bąk et al., 2021; Musich et al., 2021). The Burrow-Wheeler aligner (`BWA-MEM` and `BWA-MEM2`) is a highly popular short-read aligner (Li, 2013; Md et al., 2019). Additionally, the best practice is to use a reference genome closely related to your samples of interest, but how closely related your reference genome needs to be to your samples will depend on the divergence between species and amongst populations. (Günther and Nettelblad, 2019). For example, in a *Zea mays* RNA-seq study, as much as one-half of alleles with increased gene expression were not detected when reads from the inbred line, B73, were mapped to the reference of a second inbred line, Mo17, because *Z. mays* has high nucleotide diversity and structural variation (Zhan et al., 2021).

The SAMs or BAMs are processed to remove duplicate reads and add read groups, which provide an improved evaluation of sequencing and alignment quality but have limited effect on variant detection (Ebbert et al., 2016). `SAMtools` (Danecek et al., 2021) and `GATK` (De Summa et al., 2017; Van der

Auwera and O'Connor, 2020) provide useful guidelines and pipelines for effectively processing the alignment files. The sequencing and alignment quality should be evaluated for attributes such as mapping quality, the percent of reads mapping, and coverage before variant calling (Nielsen et al., 2011). Although this can be accomplished with custom scripts, software like `Qualimap` provides a user-friendly evaluation of sequence quality (García-Alcalde et al., 2012; Okonechnikov et al., 2016). If the quality is poor, reads may need to be trimmed to remove adapters or low-quality bases and re-mapped (Sewe et al., 2022). `Trimmomatic` (Sewe et al., 2022) and `fastp` (Chen et al., 2018; Chen, 2023) efficiently detect and trim a wide variety of adaptor sequences.

Variants are then identified using a variant caller, which determines whether a particular site in a sequenced sample is different from the reference genome. Many variant callers, such as `GATK` (Van der Auwera and O'Connor, 2020), were developed for human genomes and have been adopted for use with highly repetitive plant genomes. Before genotype calling, sites that are fixed across sequenced samples, known as invariant sites, are often excluded to improve computational efficiency. It should be noted that the inclusion of invariant sites is important for many population and quantitative genetics analyses, such as the estimation of nucleotide diversity and demographic history, and they can be added back into the pipeline after variant calling. Genotypes are subsequently called where the most likely genotype is estimated based on the number of references and alternate reads that are mapped to a given site (Nielsen et al., 2011).

The same software is often used for both variant calling and genotyping. Importantly, the genotype caller selected should be able to estimate polyploid genotypes. Polyploid genotype callers have been sufficiently compared and reviewed elsewhere (Grandke et al., 2016; Blischak et al., 2018; Gerard et al., 2018; Clark et al., 2019; Cooke et al., 2022). Briefly, polyploid variant and genotype callers that can be applied to whole genome sequence data include `GATK`, `freebayes(Garrison and Marth, 2012),EBG`

(Blischak et al., 2018), Updog (Gerard et al., 2018), polyRAD (Clark et al., 2019), and Octopus (Cooke et al., 2021). Additionally, GATK, freebayes, and Octopus can identify small structural variants under 50 bp (Cooke et al., 2022). Each polyploid genotype considers different aspects of polyploid biology in their estimation, and as such, researchers should select the caller that fits the biology of their study system the best. For example, Updog considers allele bias (see in Section 3.2) and preferential pairing in genotype estimation, while polyRAD considers per-site variance in inheritance mode (see in Section 3.4) (Gerard et al., 2018; Clark et al., 2019). Notably, Updog, polyRAD, and Octopus support binomial priors, which are considered 'informative' priors because they assume genotypes follow HWE, unlike GATK which usesuniform that assume genotypes have equal probabilities (McKenna et al., 2010; Gerard et al., 2018; Clark et al., 2019; Cooke et al., 2021). Additionally, polyRAD offers additional informative priors that consider population structure and mapping populations (Clark et al., 2019). Genotype callers and priors should be carefully selected as genotypes will be heavily influenced by the priors at low sequencing coverage (Clark et al., 2019).

Finally, variants are filtered to remove sites with false-positive variants and low-confidence genotypes. This is often accomplished using custom scripts, GATK, VCFtools (Danecek et al., 2011), or several other packages. Variant filtering is often grouped into two parts: 'hard' and 'soft' filtering (De Summa et al., 2017). In hard filtering, sites that fail to pass a set of quality controls are removed to reduce the likelihood of falsely identifying them as polymorphic. The quality controls may include mapping quality, base quality, depth, and strand bias (defined in Van der Auwera and O'Connor, 2020). Biallelic sites are typically selected when hard filtering, regardless of ploidy, as most empirical and theoretical population and quantitative genetics assume only two alleles (but see Karlin, 1990; Balding and Nichols, 1995; Ferretti et al., 2018; Broman et al., 2019 for examples of multi-allelic approaches). After hard filtering, soft filters are applied to prioritize variants specific to downstream analyses, often ad-hoc. For example, a minor allele frequency filter is a soft filter often applied to exclude sites with rare variants. Thresholds for hard and soft filtering are user-defined and formal testing of the significance of a given threshold is uncommon. Researchers often derive thresholds from those previously applied within their study system, review articles (Van der Auwera et al., 2013; Clevenger et al., 2015), or, less commonly, those tested in an empirical study (Linck and Battey, 2019; Pearman et al., 2022). Importantly, researchers should take care not to over-filter their datasets as many population and quantitative genetics analyses can be biased by datasets where particular variant classes were excluded (Linck and Battey, 2019; Pearman et al., 2022).

References

Ackiss, A. S., and F. Balao. 2020. Diving in uncharted waters: An updated genetics toolkit highlights the challenges of polyploidy in landscape genomics analyses. *Molecular Ecology Resources* 20: 841–843.

Altmann, A., P. Weber, D. Bader, M. Preuss, E. B. Binder, and B. Müller-Myhsok. 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics* 131: 1541–1554.

Bąk, A., D. Bodziony, G. Migdałek, C. S. Pareek, and K. Żukowski. 2021. Evaluation of analytical protocols of alignment mapping tools using high throughput next-generation genome sequencing data. *Translational Research in Veterinary Science* 3: 61.

Balding, D. J., and R. A. Nichols. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96: 3–12.

Blischak, P. D., L. S. Kubatko, and A. D. Wolfe. 2018. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics* 34: 407–415.

Bohutínská, M., J. Vlček, P. Monnahan, and F. Kolář. 2023. Population genomic analysis of diploid-autopolyploid species. *Methods in Molecular Biology* 2545: 297–324.

Broman, K. W., D. M. Gatti, P. Simecek, N. A. Furlotte, P. Prins, Ś. Sen, B. S. Yandell, and G. A. Churchill. 2019. R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations. *Genetics* 211: 495–502.

Chen, S. 2023. Ultrafast one−pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* 2.

Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34: i884–i890.

Clark, L. V., A. E. Lipka, and E. J. Sacks. 2019. polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3* 9: 663–673.

Clevenger, J., C. Chavarro, S. A. Pearl, P. Ozias-Akins, and S. A. Jackson. 2015. Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. *Molecular Plant* 8: 831–846.

Cooke, D. P., D. C. Wedge, and G. Lunter. 2021. A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology* 39: 885–892.

Cooke, D. P., D. C. Wedge, and G. Lunter. 2022. Benchmarking small-variant genotyping in polyploids. *Genome Research* 32: 403–408.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10.

De Summa, S., G. Malerba, R. Pinto, A. Mori, V. Mijatovic, and S. Tommasi. 2017. GATK hard filtering: Tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinformatics* 18: 119.

Dufresne, F., M. Stift, R. Vergilino, and B. K. Mable. 2014. Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology* 23: 40–69.

Ebbert, M. T. W., M. E. Wadsworth, L. A. Staley, K. L. Hoyt, B. Pickett, J. Miller, J. Duce, et al. 2016. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* 17 Suppl 7: 239.

Ferretti, L., A. Klassmann, E. Raineri, S. E. Ramos-Onsins, T. Wiehe, and G. Achaz. 2018. The neutral frequency spectrum of linked sites. *Theoretical Population Biology* 123: 70–79.

García-Alcalde, F., K. Okonechnikov, J. Carbonell, L. M. Cruz, S. Götz, S. Tarazona, J. Dopazo, et al. 2012. Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics* 28: 2678–2679.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv*.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens. 2018. Genotyping polyploids from messy sequencing data. *Genetics* 210: 789–807.

Grandke, F., P. Singh, H. C. M. Heuven, J. R. de Haan, and D. Metzler. 2016. Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: A comparative study in hexaploid *Chrysanthemum*. *BMC Genomics* 17: 672.

Günther, T., and C. Nettelblad. 2019. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genetics* 15: e1008302.

Karlin, S. 1990. Levels of multiallelic overdominance fitness, heterozygote excess and heterozygote deficiency. *Theoretical Population Biology* 37: 129–149.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.

Linck, E., and C. J. Battey. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources* 19: 639–647.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20: 1297–1303.

Md, V., S. Misra, H. Li, and S. Aluru. 2019. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. *arXiv*: arXiv:1907.12931.

Meirmans, P. G., S. Liu, and P. H. van Tienderen. 2018. The analysis of polyploid genetic data. *The Journal of Heredity* 109: 283–296.

Musich, R., L. Cadle-Davidson, and M. V. Osier. 2021. Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Frontiers in Plant Science* 12: 657240.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* 12: 443–451.

Okonechnikov, K., A. Conesa, and F. García-Alcalde. 2016. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32: 292–294.

Pearman, W. S., L. Urban, and A. Alexander. 2022. Commonly used Hardy-Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. *Molecular Ecology Resources* 22: 2599–2613.

Sewe, S. O., G. Silva, P. Sicat, S. E. Seal, and P. Visendi. 2022. Trimming and Validation of Illumina Short Reads Using Trimmomatic, Trinity Assembly, and Assessment of RNA-Seq Data. *In* D. Edwards [ed.], Plant Bioinformatics: Methods and Protocols, 211–232. Springer US, New York, NY.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, et al. 2013. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* 43: 11.10.1–11.10.33.

Van der Auwera, G. A., and B. D. O'Connor. 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, Incorporated.

Zhan, S., C. Griswold, and L. Lukens. 2021. *Zea mays* RNA-seq estimated transcript abundances are strongly affected by read mapping bias. *BMC Genomics* 22: 285.