Sancho *et al. Plant Methods* (2025) 21:134
https://doi.org/10.1186/s13007-025-01458-6

Plant Methods

SOFTWARE Open Access

AlloSHP: deconvoluting single homeologous polymorphism for phylogenetic analysis of allopolyploids

R. Sancho[1,2*], P. Catalán[1,3], J. P. Vogel[4] and B. Contreras-Moreira[3,5*]

Abstract

Background The genomic and evolutionary study of allopolyploid organisms involves multiple copies of homeologous chromosomes, making their assembly, annotation, and phylogenetic analysis challenging. Bioinformatics tools and protocols have been developed to study polyploid genomes, but sometimes require the assembly of their genomes, or at least the genes, limiting their use.

Results We have developed AlloSHP, a command-line tool for detecting and extracting single homeologous polymorphisms (SHPs) from the subgenomes of allopolyploid species. This tool integrates three main algorithms, WGA, VCF2ALIGNMENT and VCF2SYNTENY, and allows the detection of SHPs for the study of diploid-polyploid complexes with available diploid progenitor genomes, without assembling and annotating the genomes of the allopolyploids under study. AlloSHP has been validated on three diploid-polyploid plant complexes, *Brachypodium*, *Brassica*, and *Triticum-Aegilops*, and a set of synthetic hybrid yeasts and their progenitors of the genus *Saccharomyces*. The results and congruent phylogenies obtained from the four datasets demonstrate the potential of AlloSHP for the evolutionary analysis of allopolyploids with a wide range of ploidy and genome sizes.

Conclusions AlloSHP combines the strategies of simultaneous mapping against multiple reference genomes and syntenic alignment of these genomes to call SHPs, using as input data a single VCF file and the reference genomes of the known or closest extant diploid progenitor species. This novel approach provides a valuable tool for the evolutionary study of allopolyploid species, both at the interspecific and intraspecific levels, allowing the simultaneous analysis of a large number of accessions and avoiding the complex process of assembling polyploid genomes.

Keywords Allopolyploids, Homeologous chromosomes, Single homeologous polymorphisms-SHPs, Subgenome, Synteny

*Correspondence:
R. Sancho
rsancho@eead.csic.es
B. Contreras-Moreira
bcontreras@eead.csic.es
[1]Departamento de Ciencias Agrarias y del Medio Natural, Escuela Politécnica Superior de Huesca, Universidad de Zaragoza, Huesca, Spain
[2]Present address: Laboratory of Genetics and Plant Breeding, Department of Genetics and Plant Production, Estación Experimental de Aula Dei-Consejo Superior de Investigaciones Científicas, Avenida Montañana 1005, Zaragoza E50059, Spain
[3]Grupo de Bioquímica, Biofísica y Biología Computacional, BIFI-UNIZAR, Unidad Asociada al CSIC por EEAD, Zaragoza, Spain
[4]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[5]Laboratory of Computational and Structural Biology, Department of Genetics and Plant Production, Estación Experimental de Aula Dei-Consejo Superior de Investigaciones Científicas, Avenida de Montañana 1005, Zaragoza E50059, Spain

Background

A polyploid is an organism with three or more complete sets of chromosomes. Two types of polyploids can be defined according to their origin; while autopolyploids result from non-reduced gametic crosses within or between populations of the same species, allopolyploids derive from crosses between different species, i.e. hybrids [1]. It is estimated that 30–70% of all angiosperms are polyploids [2–5]. This range varies depending on the methodology and the taxonomic circumscription of the plants used to estimate it [6]. There is consensus that polyploidization is one of the key mechanisms of speciation and is ubiquitous among angiosperms [1, 7]. Indeed, it is accepted that all seed plants have undergone at least one round of whole genome duplication (WGD) in their evolutionary history and are considered to have a paleopolyploid ancestry [8, 9]. Although polyploidization events have played a key role in plant evolution, they have also occurred to a lesser extent in other organisms such as animals and fungi, resulting in polyploid species of fish, amphibians, reptiles, and, although extremely rare, also birds and mammals. Polyploid species can be found among invertebrates such as crustaceans, insects, mollusks, annelids or nematodes, among others [10, 11]. In the case of fungi, some exist as stable haploid, diploid, or polyploid (heteroploid) cells or organisms, while others change ploidy under certain conditions [12, 13].

Rapid advances in both sequencing technologies and bioinformatics make it possible to generate and analyze vast amounts of sequencing data from a wide variety of organisms. This is also leading to a remarkable and continuous increase in the number of available genomes, allowing the analysis of species and populations at the pan-genomic level, i.e. using multiple reference genomes simultaneously (*Brachypodium distachyon* [14]; barley [15]; *Arabidopsis* [16, 17]; wheat [18]; *Brassica oleracea* [19]; among other species reviewed in [20, 21]). Although advances in sequencing technologies, especially with the development of long reads (PacBio/ONT), and bioinformatics tools for genome assembly have facilitated the availability of allopolyploid reference genomes, their assembly still presents additional difficulties compared to that of diploid genomes, such as the difficulty of allocating reads to highly similar subgenomes or the greater numbers of gene copies and repetitive elements. Technological advances increasingly minimize these technical difficulties; however, sequencing costs and computational requirements remain major obstacles to obtaining high-quality allopolyploid reference genomes, especially in large-scale studies requiring a large number of species or individuals (pan-genomes, population studies, etc.). These aspects, together with the additional complexity of allopolyploids due to their multiple sets of chromosomes derived from different progenitor genomes, make evolutionary studies of these organisms challenging [22–27].

Several approaches and tools have been developed for the analysis of sequences, genes, and polymorphisms focused on phylogenetic studies of polyploid species (Table 1). However, many of them focus on coding regions (genes), which on the one hand require prior assembly and annotation of their genes, and on the other hand leave out the intergenic regions, which contain many informative loci that can help infer the evolutionary relationships of populations of diploid-polyploid complexes. Thus, Bombarely et al. [28] used the consensus diploid transcriptome to identify homeologous SNPs in the genus *Glycine* and then built a progenitor reference set for each polyploid species joining the progenitors' diploid transcriptome sets. These references were used to separate reads according to their preferential mapping to one or the other progenitor genome. Oxelman et al. [29] reviewed the explicit species network methods, such as the permutation approach (e.g., PhyloNet [30, 31]) and simultaneous gene tree and species network inference (AlloppNet [32]), used to infer the allotetraploid origins of multiple genera. Marcussen et al. [33] constructed a dated allopolyploid network from individual gene trees of the genus *Viola* using three low-copy nuclear genes (GPI, NRPD2a, and SDH). Kamneva et al. [34] studied the phylogeny of several diploid and polyploid species of the genus *Fragaria* using a large number of multilabeled gene trees, and Sancho et al. [35] developed a protocol (PhyloSD) to infer the homeologous subgenomes in *Brachypodium* allopolyploids and reconstruct their evolution, even when their diploid ancestors are unknown (orphan subgenomes).

Other approaches have been developed to analyze interspecific hybrids and allopolyploids using SNPs or synteny and microsynteny-based approaches. However, these two strategies have not been combined to reconcile syntenic homeologous SNPs in a single alignment that allows phylogenetic inference and studies of population structure at the subgenomic level. Regarding SNP discovery, some approaches use the genomes of their extant closest available progenitor genomes to infer the SNPs from each of the homeologous subgenomes of the allopolyploid, defined as homeoSNPs. For example, Page et al. [36] implemented the PolyCat pipeline to map and categorize the genomic data generated from allopolyploids and it was tested in cotton. Peralta et al. developed SNiPloid [37], a web tool focused on SNP analysis of RNA-Seq data obtained from allotetraploids. Mithani et al. and Khan et al. implemented HANDS [38] and HANDS2 [39], a method to characterize homeolog-specific polymorphisms (HSPs) in polyploid genomes, tested in the allopolyploids bread wheat and *Brassica* genus. Kulkarni et al. developed the Comprehensive Allopolyploid

Table 1 Summary of tools for polyploid subgenome studies

Tool and Ref.	Foundation	Strengths	Limitations
AlloSHP (This study, 2025)	A command-line tool designed to detect and extract single homeologous polymorphisms (SHPs) in allopolyploid species by integrating simultaneous mapping to multiple reference genomes with syntenic genome alignment.	Assembly of the allopolyploid genome or genes is not required. Instead, for each allopolyploid, reads are mapped once against the concatenated reference genomes of its diploid parents, while preserving SNP positional traceability relative to the parental references. For output, both FASTA and VCF formats are available.	Recovered SNPs are restricted to syntenic regions between the parental diploid genomes. Heterozygous sites are excluded to minimize false positives in the final set of SHPs. Reference genomes from diploid progenitors or extant relatives are required.
CAPG (Kulkarni et al. 2023) [40]	A tool that defines an explicit likelihood to weight read alignments against both subgenomic references and to genotype individual allopolyploids from whole-genome resequencing data.	Variant calls are reported in VCF format, incorporating measures such as genotype likelihoods to assess statistical support. For each individual, sites are classified as homoeologous SNPs, allelic SNPs within the subgenome, or invariant. Heterozygous positions can be dealt with.	The approach was evaluated using allotetraploid species (peanut and cotton) as well as simulated datasets. It requires reference sequences from both subgenomes with known alignments in homologous regions.
PhyloSD (Sancho et al. 2022) [35]	A pipeline that integrates three sequential algorithms—Nearest Diploid Species Node, Bootstrapping Refinement, and Subgenome Assignment—which involves computational filtering, homeolog labeling, and homeolog allocation to subgenomes.	Reference genomes are not required, and subgenomes can be identified even when one or more diploid progenitors are missing, including both known and 'ghost/orphan' subgenomes.	Gene and/or CDS assemblies from both diploid and polyploid species are necessary to infer gene trees. The approach was evaluated using allotetraploids and allohexaploids of *Brachypodium* and *Triticum-Aegilops* complexes.
GRAMPA (Thomas et al. 2017) [46]	A gene-tree reconciliation algorithm that extends traditional reconciliation approaches to multi-labeled trees, enabling the explicit modeling of polyploid evolution.	This method infers polyploidy type (allo- vs. auto-), identifies whole-genome duplication events, and quantifies gene duplications and losses associated with polyploidy.	Gene trees and species tree topologies are required.
HANDS2 (Khan et al. 2016) [39]	A next-generation sequencing-based tool, an updated version of the HANDS tool (38), enabling highly accurate genome-wide identification of homeolog-specific bases in allopolyploids, even without a diploid progenitor.	It supports up to ten diploid progenitors and works in the absence of a diploid progenitor.	Accurate base assignment requires high-quality RNA-seq data with sufficient coverage. The method struggles to detect gene silencing in one or more subgenomes, potentially leading to incorrect base calls. Additionally, gene conversion or homeologous exchanges can introduce bias in base assignments.
PolyDog (Page et al. 2014; Page and Udall 2015) [41, 47]	A read categorization tool based on a dual-reference approach using data from two reference sequences, one for each genome of an allotetraploid.	Provides higher resolution than GSNAP (used in PolyCat) by handling multiple mappings, without requiring assembly of the allopolyploid genome or genes.	For each allopolyploid, reads must be mapped separately to the reference genomes of both diploid parents, increasing computational time and resources. This approach was tested in allotetraploid cotton and requires reference genomes of the diploid progenitors or their extant relatives.
PolyCat (Page et al. 2013; 2014) [36, 41]	A pipeline for mapping and categorizing all types of NGS data from allopolyploid organisms, using GSNAP's single-nucleotide polymorphism (SNP)-tolerant mapping to minimize the mapping efficiency bias caused by SNPs between genomes.	Assembly of the allopolyploid genome is not required and only a single diploid reference genome is needed.	The approach is limited by the genomic density of homoeo-SNPs, and the reads can only be classified if they overlap at least one SNP. This method was tested in allotetraploid cotton and requires reference genomes from at least one diploid progenitor or extant relative.
SNiPloid (Peralta et al. 2013) [37]	A web tool based on the coassembly of homoeologs, comparing either putative SNPs detected from an allopolyploid to those obtained in its parental genomes, or putative SNPs derived from two allopolyploid accessions to search for polymorphism.	No allopolyploid genome or gene assembly is required. SNPs can be classified into categories based on hypothesized evolutionary patterns using a web-based tool.	Evaluated using a diploid reference transcriptome and RNA-seq data from the allotetraploid *Coffea arabica*. Bias may occur due to greater similarity between the reference diploid transcriptome and one of the two subgenomes of the allopolyploid.

Table 1 (continued)

Tool and Ref.	Foundation	Strengths	Limitations
AlloppMUL/ AlloppNET (Jones et al. 2013) [32]	Statistical inference models for allopolyploid networks: AlloppMUL infers the multilabeled species tree directly, allowing topology, node times, and branch population sizes to vary freely, treating the diploid genomes within allotetraploids as if they belonged to separate species. In contrast, AlloppNET explicitly models hybridization as a node in the species network, requiring the diploid genomes within allotetraploids to share population sizes and speciation events.	Both models are implemented in the BEAST framework.	Gene assembly is required. Tested on a limited genes of tetraploids from the genera *Pachycladon* (Brassicaceae) and *Silene* (Caryophyllaceae), as well as on simulated datasets. Multiple priors need to be established. Heterozygosity can result in multiple alleles per locus, complicating the distinction between homeologous and heterozygous alleles.

Genotyper (CAPG) method [40], which uses a likelihood to weight read alignments against both subgenomic references and then genotype individual allopolyploids from whole-genome resequencing data. Page & Udall , Phillips and Session have published reviews of the various methods for mapping and categorizing reads, as well as the variant-calling approaches required to study polyploid organisms [27, 41] and to identify the ancestors of allopolyploid subgenomes [42].

Although synteny and microsynteny-based approaches have been widely used in phylogenetic inference, they have mainly studied diploid species and have focused on the synteny of coding regions, i.e., the collinearity of genes (Brassica [43]; rosids [44]; angiosperms [45]), without exploiting the phylogenetic information of the rest of the genome.

Similarly, available approaches to reveal the homeologous subgenomes of allopolyploid species have also focused on coding regions. These require genomes to be previously assembled and annotated, and sometimes prior information on the diploid species of particular group. Furthermore, these protocols can sometimes be very complex for users without prior bioinformatics knowledge.

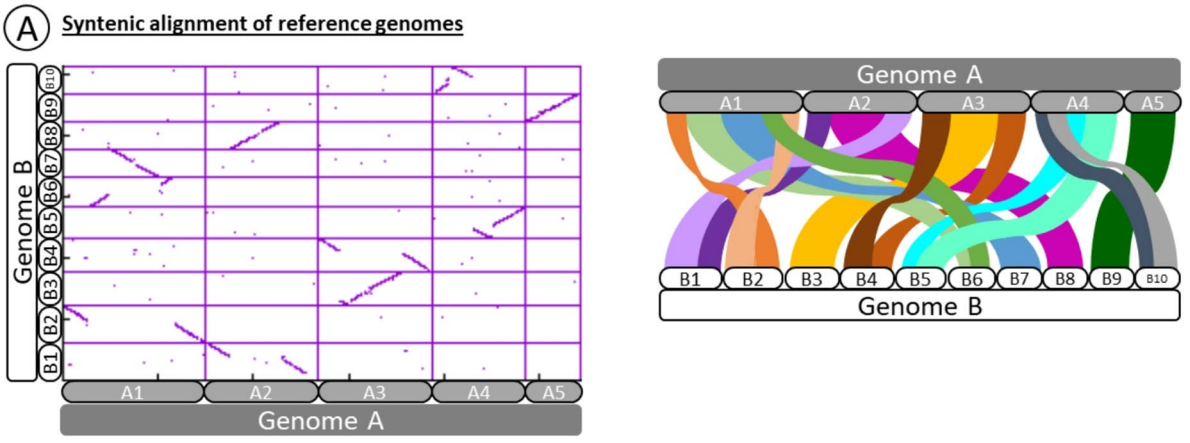
To simplify the bioinformatics protocols used in the study of allopolyploids, as well as the type of sequences and prior information required, we have developed an integrative whole-genome synteny-based phylogenetic inference approach to detect syntenic homeoSNPs, defined here as Single Homeologous Polymorphisms (SHP), for the study of diploid-polyploid complexes for which diploid progenitor genomes are available. The AlloSHP pipeline does not require assembling nor annotating the genomes of the allopolyploids under study. Our approach allows phylo(sub)genomic analysis at both the inter- and intra-specific levels, providing insight into which lineages may be involved in the origin and evolution of the different polyploid populations that are part of diploid-polyploid complexes.

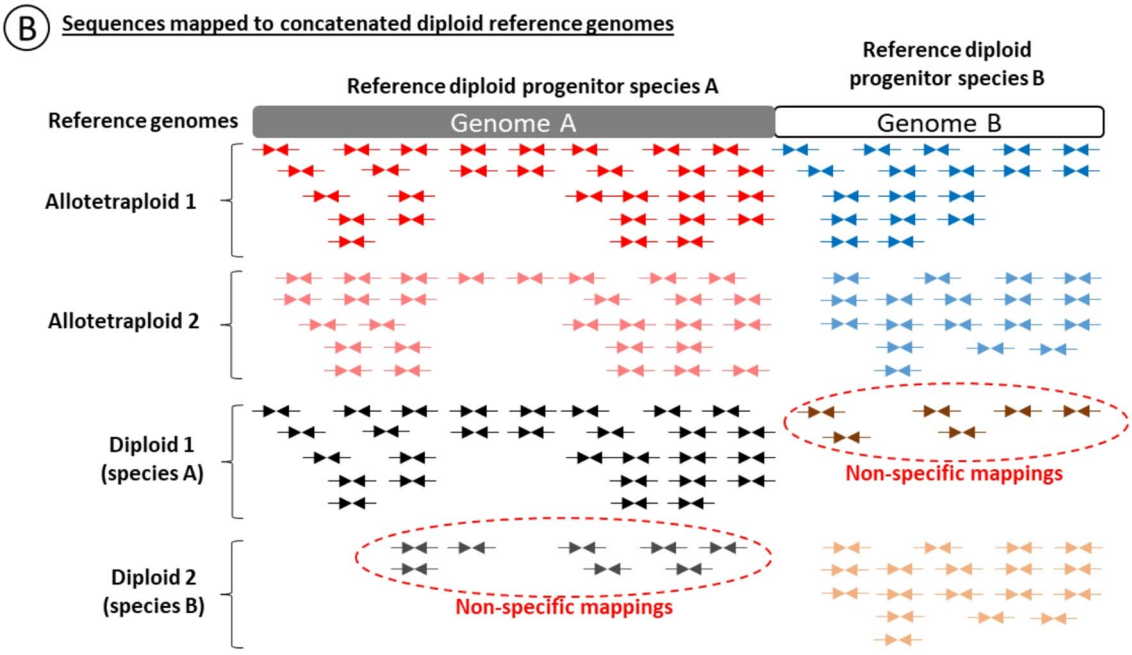
Implementation

The input files required to use this protocol are (i) the reference genomes of the diploid progenitor species whose syntenic positions are going to be determined (FASTA format, Fig. 1A) and (ii) a VCF summarizing the mapping of reads of all the polyploid samples to be studied (Fig. 1B). The VCF is obtained by merging BAM files, one per sample, and must contain the DP (total read depth) field. The step-by-step instructions and the external software used are detailed in the following sections and supplementary information.

The AlloSHP detection pipeline has three core algorithms, included as scripts in the repository https://github.com/eead-csic-compbio/AlloSHP: WGA (Whole Genome Alignment), VCF2ALIGNMENT, and VCF2SYNTENY. Each algorithm plays a key role in extracting the syntenic positions between the reference genomes, determining the most confident SNPs according to sequencing depth and missing sample thresholds, and finally combining both types of information to deconvolute and extract SHPs (Figs. 1 and 2).

The *Whole Genome Alignment (WGA)* algorithm by default calls the aligner CGaln [48] to perform the alignment and detect the syntenic segments between pairs of reference genomes after soft-masking repeated sequences (Figs. 1A and 2A). One progenitor reference genome is defined as "primary or master (genome A)" and the others as "secondary (genomes B, C, .)". There can be any number of "secondary" progenitor reference genomes according to the ploidy of the allopolyploids under study, but they must always be aligned against the same master reference genome [e.g. the study of an allohexaploid includes three reference genomes from its diploid progenitors. One genome is established as the "master (A)," while the other two are established as "secondary (B and C)"]. The syntenic positions are extracted and used downstream. Three main output files are generated: (i) a 0-based BED list of syntenic positions indicating chromosome, position, strand, nucleotide, CGaln syntenic block, and SNP presence between primary and secondary reference genomes; (ii) a Log file containing the parameters, thresholds, and additional information





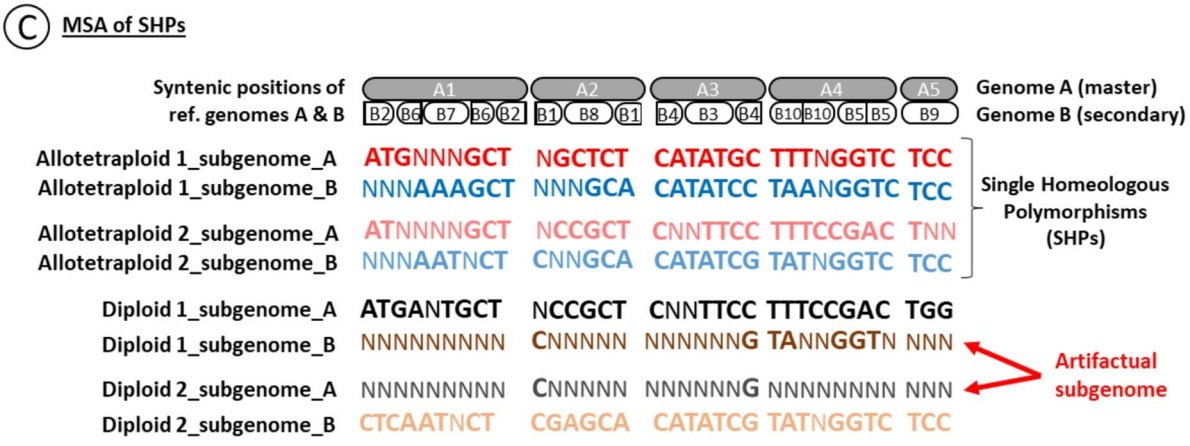


Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Simplified illustration of the main pipeline processes. **A** Alignment of syntenic regions between reference genomes A (chromosomes A1-A5) and B (chromosomes B1-B10) using CGaln (left) and displayed in a riparian plot (right). **B** Mapping of reads against the concatenated reference genomes A and B. Reads from allopolyploid samples (sample 1 and sample 2) are represented by arrows (forward and reverse). Red and blue colors indicate reads mapped against genome A or B, respectively. **C** Multiple sequence alignment of Single Homeologous Polymorphisms (SHPs). Each letter corresponds to an SHP mapped against the reference genomes A or B. Each sample (Allotetraploid 1 and 2, and Diploid (species A) and Diploid (species B)) has as many draft subgenomes as reference genomes used in the mapping. Artifactual subgenomes are those derived from unspecific mappings and are defined by setting a %SHP threshold using cross-validation against the non-self mappings of diploid samples. These are then eliminated downstream

extracted from the CGaln output; and (iii) a PDF dot plot file showing the syntenic regions between reference genomes. This needs to be inspected to assess the quality of the alignment in terms of matched regions and noise.

The *VCF2ALIGNMENT* algorithm parses an input VCF (variant call format) file and produces a list of valid homozygous positions (valid loci), taking into account the minimum read depth (-d) and maximum missing samples (-m) thresholds. Heterozygous sites are handled as missing data to avoid detecting allelic SNPs as homeologous SNPs [40]. Optionally, the -H flag can be used to maintain heterozygous sites in downstream analysis. The result is a LOG file with a list of positions (valid loci) that have passed the thresholds. This list indicates the chromosome and position with respect to the master reference genome, as well as the number of missing samples and the called nucleotide (ATGC). In addition, at the end of the file, the total number of valid loci and polymorphic loci is shown, as well as for each target sample and per reference chromosome. Optionally, multiple sequence alignments (MSA) can be calculated, but on this protocol, this only really makes sense on the next step (Figs. 1B and 2B).

The *VCF2SYNTENY* algorithm parses the (i) VCF file obtained from reads mapped to multiple concatenated reference genomes, (ii) the LOG file of valid loci computed by VCF2ALIGNMENT, and (iii) the synteny-based equivalent coordinates (BED file) computed by WGA to align the polymorphic loci referenced by syntenic positions, separating them on each reference genome, and defining them as SHPs. The resulting MSA will have as many subgenomes as reference genomes were used (Figs. 1C and 2C). Some of these subgenomes might be artifacts resulting from residual, unspecific reads mapped against one of the reference genomes (e.g., non-self-mapping in diploid samples, or mapping from non-progenitor species in allopolyploids). These "false subgenomes", the so-called "artifactual subgenomes", must be eliminated from the final multiple sequence alignment (MSA). If the ploidy, and therefore the number of subgenomes expected to be recovered, is unknown, we propose using a cross-validation criterion based on the percentage of SHPs recovered from non-self mappings (not the same species) in diploid samples. In diploid samples, only one predominant genome/subgenome is expected to be recovered. For each subgenome, the highest SHP percentage obtained in the diploid samples from the non-specific mappings will be set as a threshold (see Figs. 1B and C). If multiple accessions of the same diploid species are included, the artifactual subgenome threshold can be set as either the average non-specific %SHP value among the accessions or the highest value. All subgenomes with values equal to or lower than this threshold will be identified as artifactual subgenomes and removed from the final MSA (see Results section).

Outgroup species sequences can optionally be added into the SHP alignment by indicating in the configsynt file the BED file of the syntenic positions between the master genome and the outgroup species genome computed with WGA (see details in *Brachypodium* case below).

AlloSHP is available for Linux and MacOS operating systems. It can be installed via a local compilation or conda (https://anaconda.org/bioconda/alloshp).

Evaluation of the protocol: species, reference genomes and target samples

The protocol was tested and validated using the genome sequences of four diploid-polyploid complexes from three plant groups (*Brachypodium distachyon* complex, *Triticum-Aegilops* complex, and *Brassica* complex) and species from one yeast genus (*Saccharomyces* haploid, diploidized species and synthetic hybrids). The datasets include sequences from the allopolyploids under study and their closest extant diploid progenitor species. Accessions from the diploid species were also included as control samples. The genomes of the diploid species were used as references to conduct the mapping of the reads from the allopolyploid and diploid species of each complex to reconstruct their phylogeny at the subgenome level (Suppl. Figure 1; Suppl. Table 1).

The *Brachypodium distachyon* complex analysis included the diploid *B. distachyon* [49] and *B. stacei* (JGI Genome Portal: https://genome.jgi.doe.gov/portal/ [50]; Catalan et al. (unpublished)) reference genomes and two diploid *B. distachyon* (ABR2 and Bd21), two diploid *B. stacei* (ABR114 and T.E4.3), and two allotetraploid *B. hybridum* (Bhyb26 and ABR113) ecotypes as samples (JGI Genome Portal; [14, 51, 52]). *Oryza sativa* Japonica group cv. Nipponbare (NCBI: https://www.ncbi.nlm.nih.gov/; [53]) was included as outgroup species genome (Suppl. Figure 1 A; Suppl. Table 1).

The *Brassica* complex analysis included the diploid *Brassica oleracea* [54] and diploid *Br. rapa* [55] reference

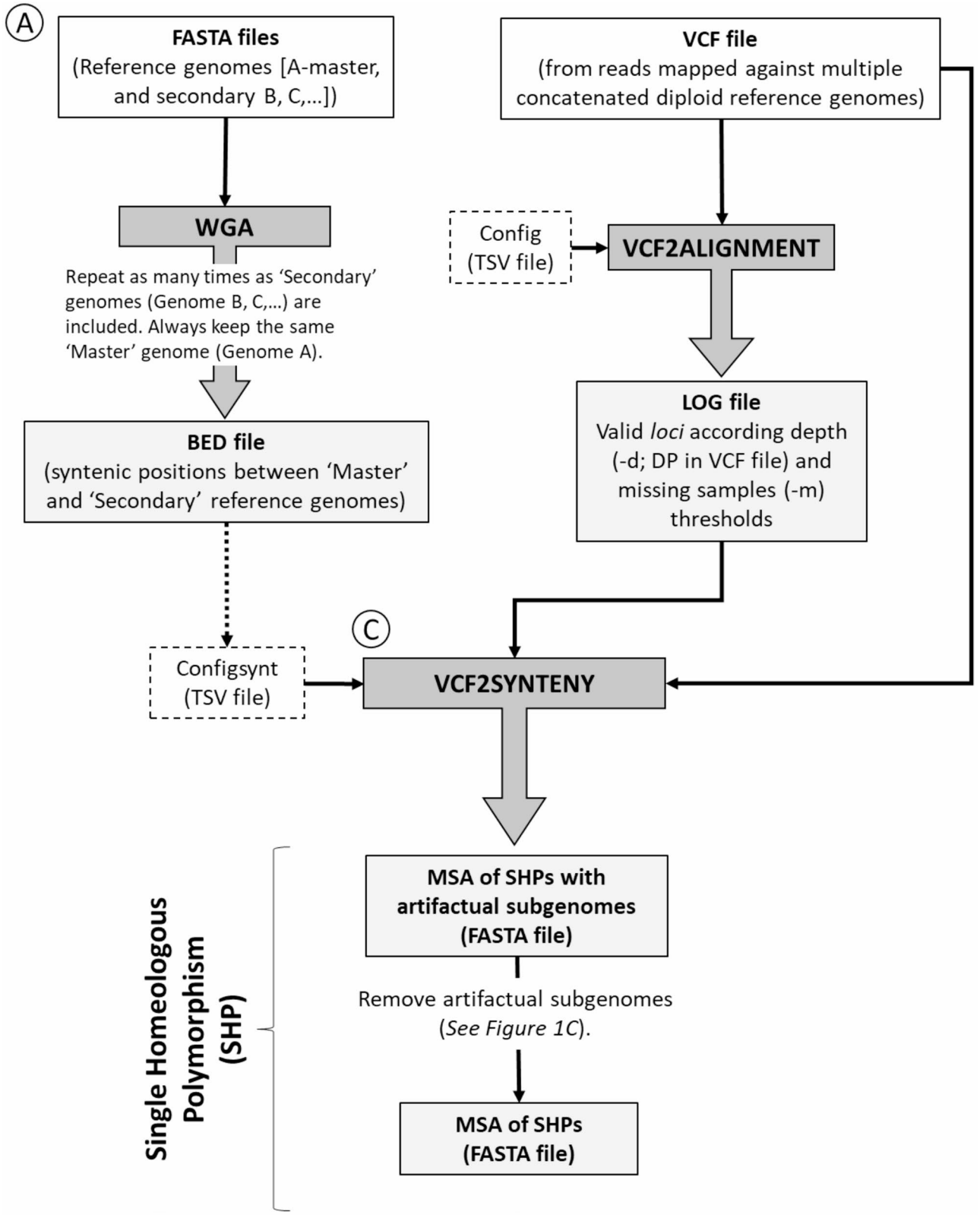


Fig. 2 Flowchart of the main tasks and deliverables of the AlloSHP pipeline. The gray squares indicate the three main algorithms (WGA (**A**), VCF2ALIGNMENT (**B**), and VCF2SYNTENY (**C**)) that make up the pipeline. The white squares indicate the input files required for each algorithm and the resulting output files. The dashed boxes indicate the two configuration files needed for the VCF2ALIGNMENT and VCF2SYNTENY algorithms

genomes and two diploid *Br. oleracea*, two diploid *Br. rapa* and five allotetraploid *Br. napus* cultivars as samples [54] (Suppl. Figure 1B; Suppl. Table 1).

The *Triticum-Aegilops* complex analysis included the diploid *Triticum urartu* [56], diploid *Aegilops tauschii* [57] and diploid *Ae. speltoides* [58] reference genomes and the diploids *T. urartu*, *Ae. tauschii*, *Ae. speltoides* (two samples, Y2032 and Tivon), the allotetraploid *T. turgidum*, and the allohexaploid *T. aestivum* accessions as samples [56, 59, 60] (Suppl. Figure 1 C; Suppl. Table 1).

Saccharomyces yeast analysis included the haploid *Saccharomyces cerevisiae* [61], *S. kudriavzevii*, *S. mikatae*, *S. paradoxus* [62, 63], and *S. uvarum* reference genomes, and a diploidized *S. cerevisiae* and six other *Saccharomyces* synthetic hybrids [yHRWh24 (*S. cerevisiae x S. kudriavzevii x S. mikatae x S. uvarum*), yHRWh4 (*S. mikatae x S. kudriavzevii*), yHRW134 (Diploidized *S. cerevisiae*), yHRWh13 (S. *mikatae x S. uvarum*), yHRWh10 (*S. cerevisiae x S. uvarum*), yHRWh18 (*S. kudriavzevii x S. paradoxus*), yHRWh51 (*S. uvarum x S. mikatae x S. kudriavzevii*) [59, 64] as samples (Suppl. Figure 1D; Suppl. Table 1).

Evaluation of the protocol: pre-processing and integrative pipeline application

The same pre-processing of the sequence reads was applied to the four data sets. For each of the samples to be analyzed, the quality check (QC) report was generated using FastQC 0.11.9 software [65] before and after filtering the sequences using Trimmomatic 0.39 [66]. The reads that passed QC were mapped against the concatenated reference genome (including the diploid reference genomes of the progenitor species of the polyploid species under study) using minimap2.1-r1122 [67, 68]. The mapped reads (SAM format) were converted to BAM format and sorted using samtools 1.16.1 [69] software (--sort and --view functions). Samtools was also used to obtain the statistics of the mapped reads, both in general (--flagstats) and per chromosome of each reference genome (--index and --idxstats). Variant calling was performed using the bcftools 1.16 software [69] (--mpleup with parameters -a DP and -call. The resulting VCF files for each sample were compressed using bgzip 1.19 [70], and indexed and merged using bcftools (--index and --merge) to create the multi-sample VCF file. This VCF file was used as the input file for the VCF2ALIGNMENT and VCF2SYNTENY algorithms.

Syntenic positions between reference genomes were then calculated using WGA, which requires the *utils/mapcoords.pl* script. By default, this algorithm uses the CGaln software [48] to compute syntenic blocks and positions. The aligner GSAlign [71] was also integrated and tested; however, the analysis performed on the *Brachypodium* reference genomes showed poorer results in terms of recovered syntenic regions and this test was not further extended.

Repetitive regions of the reference genomes had to be masked out, so WGA applied a repetitive region detection and masking procedure by default using Red [72] and Red2Ensembl.py [73]. The resulting 0-based BED file contained the list of syntenic positions and was one of the input files used by the VCF2SYNTENY algorithm. In addition, WGA also generated a PDF plot of resulting syntenic blocks by calling the gnuplot 6.0 program [74], which helped to visually control the quality of any produced alignments (Suppl. Figure 2). The parameters used for the four data sets analyzed are listed in Suppl. Table 2. Bedtools intersect v2.31.1 [75] was used to estimate the proportion of SHPs sitting in annotated genes and non-coding regions of the master genome, as annotated in the respective GFF file.

The VCF2ALIGNMENT algorithm was used to filter the positions in the VCF file that passed the filtering step with respect to the read depth (-d) and number of missing samples (-m) thresholds. In addition, indels were filtered out from the VCF to eliminate downstream inconsistencies. The imposed thresholds were $d \geq 5$ (i.e. DP ≥ 5 reads in VCF file) for all data set analyzed, and $m \leq 3$ for *Brachypodium*, $m \leq 5$ for *Brassica*, $m \leq 4$ for *Triticum-Aegilops*, and $m \leq 10$ for *Saccharomyces* for missing data. The resulting LOG file with information about the positions that passed the filters was used as the input file for VCF2SYNTENY.

The VCF2SYNTENY algorithm reconciled the information of the syntenic positions obtained by WGA (BED file) and the valid positions obtained by VCF2ALIGNMENT (LOG file), effectively deconvoluting SHPs, which were assigned to the corresponding subgenomes. It is important to note that it is necessary to specify which reference genome will be established as the master genome, since this will be the one used for referencing the positions in the generated multiple sequence alignment (MSA). When using two reference genomes, either one can be selected as the master genome. When three or more genomes are used for evolutionary close species, the longest genome is generally selected as the master genome. However, the results of syntenic blocks obtained with CGaln should always be considered when making this decision.

Comparing synteny-based SNPs with polymorphisms between single-copy orthologues

In order to measure how reliable WGA-based SNPs are, they were compared to SNPs observed between single-copy orthologous genes annotated in Ensembl Plants release 62. This was done for two pairs of genomes analyzed in this work, *Brassica oleracea* vs. *Brassica rapa*, and *Triticum urartu* vs. *Aegilops tauschii*. In order to

match the data in Ensembl, assembly GCA_000309985.1 was used for *B. rapa*, and assemblies GCA_003073215.1 and GCA_002575655.1 were used for *T. urartu* and *Ae. tauschii*, respectively. These were aligned with WGA using the same parameters as before. As for single-copy orthologues, pairwise alignments were obtained with *ens_single-copy_core_genes.pl* from https://github.com/Ensembl/plant-scripts [76]. As Ensembl Compara ort hologues are computed using protein sequences, recipe R8 from https://github.com/Ensembl/plant-scripts was modified to convert CDS to genomic coordinates. Finally, SNPs from both sources were intersected with bedtools intersect v2.27.1 [75].

Phylogenetic analysis using single homeologous polymorphisms (SHP) datasets

Each of the four SHPs datasets obtained using our protocol was processed by removing artifactual subgenomes with low percentages of mappings and recovered syntenic SNPs. They were then filtered using the snp-sites v.2.5.1 tool [77] to retain only the variable positions. Phylogenetics trees were inferred using IQtree v.2.2.2.6 software [78] with parameters -m MFP + ASC -AICc -alrt 1000 -B 1000 -T AUTO. Phylograms were rooted at the midpoint, except for the *Brachypodium* phylogram, which was rooted using the *O. sativa* outgroup, and visualized using the FigTree v.1.4.4 software [79]. The same alignment used to infer the phylogeny was used to calculate the genome-wide average nucleotide identity (gwANI) using the pANIto software [80].

Results

Proportion of reads mapped against diploid reference genomes

The proportions of reads mapped using minimap2 against each concatenated diploid progenitor reference genome were as expected. Thus, the reads of the diploid species were mostly mapped against their corresponding diploid species reference genome. The ecotypes of the diploid species *B. stacei* and *B. distachyon* mapped between 92% and 98% against their respective genomes (Suppl. Table 3 A). The cultivars of the diploid species *Br. oleracea* and *Br. rapa* mapped between 78% and 88% of reads (Suppl. Table 3B). Between 87% and 97% of reads from diploids of the *Triticum-Aegilops* complex, *Ae. speltoides*, *Ae. tauschii* and *T. urartu*, mapped to their respective genomes (Suppl. Table 3 C). In yeast, diploidized *S. cerevisiae* mapped 96% of reads to its reference genome (Suppl. Table 3D).

Regarding the polyploid samples analyzed, the number of mappings against each reference genome varied depending on the species and ploidy. Thus, the mapping ratio of the allotetraploid ecotypes of *B. hybridum* (DDSS) was approximately 50:50 (subgenome D: subgenome S), but with some differences between ecotypes. The ABR113 ecotype showed 50% of the reads mapped against the *B. distachyon* reference genome compared to 54% of the Bhyb26 ecotype (subgenome D) and the respective 49% and 46% against the *B. stacei* reference genome (subgenome S) (Suppl. Table 3 A). The accessions of the allotetraploid *Brassica napus* (ArArCoCo) also showed a 50:50 ratio of mappings to their reference genomes, with slightly more mappings against the *Br. rapa* (subgenome Ar) reference genome (51–52%) than against the *Br. oleracea* reference genome (subgenome Co) (48–49% of mappings) (Suppl. Table 3B). In the case of the *Triticum-Aegilops* complex, the allotetraploid *T. turgidum* (AABB) mapped predominantly against its progenitors, *T. urartu* (46%; subgenome A) and *Aegilops speltoides* (40%; subgenome B), while 14% of the reads that mapped against *Ae. tauschii* were considered artefacts (see Threshold of artifactual subgenomes section) and eliminated from downstream analyses. The allohexaploid *T. aestivum* (AABBDD) mapped 34% against *T. urartu* (subgenome A), 29% against *Ae. speltoides* (subgenome B) and 37% against *Ae. tauschii* (subgenome D) (Suppl. Table 3 C).

The six *Saccharomyces* synthetic hybrids analyzed mapped predominantly against their parents, but with high variability in their proportions (Suppl. Table 3D). The synthetic hybrid Sce x Sku x Smi x Suv mapped predominantly against two of its parents, *S. mikatae* (Smi; 40%) and *S. kudriavzevii* (Sku; 38%), and to a lesser extent against its two other parents, *S. cerevisiae* (Sce; 12%) and *S. uvarum* (Suv; 11%). A very small percentage of reads (Spa; 0.1%) mapped against the reference genome *S. paradoxus*, which is not a parent of this hybrid. The synthetic hybrid Smi x Sku mapped 51% and 48% of reads against its two reference genomes, *S. mikatae* (Smi) and *S. kudriavzevii* (Sku), respectively. Less than 1% of the reads mapped against the other three non-parental reference genomes of the synthetic hybrid. The Smi x Suv hybrid mapped 60% and 68% against its parental *S. mikatae* (Smi) and *S. uvarum* (Suv) reference genomes, respectively. Less than 2% of the remaining reads mapped to the genomes of non-progenitor species. The Sce x Suv hybrid mapped 59% of reads against its progenitor genome *S. cerevisiae* (Sce) and 38% against that of *S. uvarum* (Suv). The remaining mappings (3%) were likely unspecific. The synthetic hybrid Sku x Spa, resulting from the cross between *S. kudriavzevii* x *S. paradoxus*, showed a balanced proportion of 48% mapping to both parental genomes. Finally, the Suv x Smi x Sku synthetic hybrid from the cross between *S. uvarum*, *S. mikatae* and *S. kudriavzevii* mapped predominantly against its parental genomes *S. mikatae* (Smi; 38%) and *S. kudriavzevii* (Sku; 37%), and less to that of *S. uvarum* (Suv; 24%) (Suppl. Table 3D).

Syntenic positions recovered among diploid progenitor reference genomes by WGA

The computation of syntenic positions between diploid reference genomes, master versus secondary, using WGA and its main dependency CGaln, is a fundamental step for the detection of SHPs from subgenomes of the polyploids under study. The number of syntenic positions recovered in each data set varied depending on the homology between the different chromosomes of the species. Thus, the reference genomes used in the *Brachypodium* group presented the highest percentage of synteny among tested plants. The *Brachypodium stacei* reference genome (ABR114 ecotype) had syntenic positions covering 29% of the *B. distachyon* reference genome (Bd21 ecotype) determined as a master reference genome (Suppl. Figure 2 A; Suppl. Table 4 A). As expected, the syntenic regions between *B. distachyon* (master genome) and *Oryza sativa* (outgroup genome) were notably reduced to 2.5% of the *B. distachyon* genome. Most of these regions (98.6%) were located within coding regions (Suppl. Table 4 A). In the case of *Brassica*, this percentage was significantly reduced, with syntenic positions between the *Br. rapa* and *Br. oleracea* reference genomes covering only the 5% of the *Br. oleracea* master reference genome (Suppl. Figure 2B; Suppl. Table 4B). Similarly, synteny between the reference genomes of *Aegilops* and *T. urartu* was approximately 3% of the *T. urartu* master reference genome (Suppl. Figure 2 C; Suppl. Table 4 C). In the yeast example, the compared species showed synteny ranging from 32% (*S. uvarum*) to 55% (*S. paradoxus*) of the master reference genome *S. cerevisiae* (Suppl. Figure 2D; Suppl. Table 4D).

The percentage of syntenic positions detected in coding regions (genes) of the master genome varied significantly among the species analyzed. In *Brachypodium* (Suppl. Table 4 A), 73.5% of detected syntenic positions were located in genes, compared to only 48 – 43% in *Brassica* (Suppl. Table 4B) and *Triticum-Aegilops* (Suppl. Table 4 C). In yeast, 86.8–96.2% of the syntenic positions were located in coding regions (Suppl. Table 4D).

As a separate benchmark, we measured to what extent WGA-based SNPs match SNPs observed between single-copy orthologous resulting from protein alignments and phylogenetic inference within the Ensembl Plants genome browser. The results in Suppl. Table 5 suggest that the majority of SNPs called by these two orthogonal approaches are actually the same (63% for *Brassica oleracea* and 71.5% for *Triticum urartu*). These numbers improve when the distance among matched nucleotides increases (from 10 to 1000 bp), revealing that a fraction of syntenic blocks don't match pairwise alignments based on amino acid sequences.

Single homeologous polymorphisms (SHPs) recovered by AlloSHP pipeline

The percentages of SHPs recovered for each subgenome (Suppl. Table 6 A-D) generally correlate with those obtained in the mappings (Suppl. Table 3 A-D). However, in some polyploids, changes in the number of predominant SHPs were observed. For example, the same proportion of SHPs was recovered in the ecotypes ABR113 and Bhyb26 of the allotetraploid *B. hybridum*, with slightly more SHPs obtained from the S-subgenome (Bsta; 50.5%) than from the D-subgenome (Bdis; 49.5%) (Suppl. Table 6 A), while these proportions were inverse in the mappings (Suppl. Table 3 A).

The accessions of the allotetraploid *Brassica napus* showed the same proportion of SHPs (Suppl. Table 6B) as the mappings (Suppl. Table 3B), with percentages of 51–52% of SNPs from the A subgenome (Brr; *Br. rapa*) and 48–49% of SHPs from the C subgenome (Bro; *Br. oleracea*) (Suppl. Table 6B).

However, in the *Triticum-Aegilops* group, some variations between mapping (Suppl. Table 3 C) and SHPs recovered for each subgenome were detected (Suppl. Table 6 C). As expected, in the *Ae. tauschii* and *T. urartu* diploid samples, 96% and 99% of the recovered SHPs (Suppl. Table 6 C) came from mappings against their own reference genomes. However, in the case of the diploid *Ae. speltoides* Y2032 and Tivon samples analyzed, there was a disparity between the mappings (Suppl. Table 3 C) and %SHP (Suppl. Table 6 C) recovered between both samples, with those of sample Y2032 (sel-mappings: 86%; self-SHP: 69%) being much less specific compared to Tivon (sel-mappings: 90%; self-SHP: 87%). This variation between samples may be influenced by the quality and number of reads, as well as the phylogenetic distance of the sample from the reference genome used. In the allotetraploid *T. turgidum*, 54% and 34% of the recovered SHPs correspond to subgenome A (Tur; *T. urartu*) and subgenome B (Aes; *Ae. speltoides*), respectively (Suppl. Table 6 C). SHPs recovered from the allohexaploid *T. aestivum* are distributed among its three subgenomes A, B and D with percentages of about 39%, 19% and 42%, respectively (Suppl. Table 6 C).

The greatest variation in the percentage of mappings (Suppl. Table 3 A-C) and SHPs detected for each subgenome (Suppl. Table 6 A-C) may be influenced by the percentage of masked genome sequences, i.e., the number and size of repetitive regions. In the three reference genomes of *Triticum-Aegilops* complex species, 80–86% of the genomes were masked. In contrast, in the reference genomes of *Brachypodium* and *Brassica*, the percentage was 31–35%.

The proportions of SHPs recovered in the synthetic yeast hybrids (Suppl. Table 6D) differed from previously obtained mappings (Suppl. Table 3D), especially in those

recovered from the parent *S. uvarum*, with notable reductions. This is because the reference genome of *S. uvarum* showed only 32.2% synteny with the primary reference genome *S. cerevisiae* (Suppl. Table 4D). Meanwhile, the genomes of *S. kudriavzevii*, *S. mikatae*, *and S. paradoxus* showed proportions of 42%, 44%, and 55%, respectively (Suppl. Table 4D). Furthermore, the *S. uvarum* genome has a higher percentage of masked regions due to repetitive sequences (38% of masked genome) than the other *Saccharomyces* reference genomes (25–27% of masked genome). The hybrid *S. mikatae* x *S. kudriavzevii* showed a very similar proportion of SHPs (Smi: 52%; Sku: 48%; Suppl. Table 6D) to mapped reads (Smi: 51%; Sku: 48%; Suppl. Table 3D). However, the hybrid *S. mikatae* x *S. uvarum* and *S. cerevisiae* x *S. uvarum* showed a marked predominance of SHPs from the parent *S. mikatae* genome (78%), in the Smi x Suv hybrid, and from the *S. cerevisiae* genome (81%) in the Sce x Suv hybrid, compared to its other parent *S. uvarum* with percentages of SHPs of 22% and 19%, respectively (Suppl. Table 6D). It should also be noted that the proportions of each parental genome in the synthetic hybrids are variable (see Peris et al., 2020; [64]). The synthetic hybrid resulting from the cross of three parents, *S. uvarum* x *S. mikata*e x *S. kudriavzevii* also showed a notable reduction in the number of SHPs of the parent *S. uvarum* (SHPs: 11%, Suppl. Table 6D) versus 24% mapped reads (Suppl. Table 3D) compared to the predominance of its other two parents *S. kudriavzevii* (SHPs: 43%; mappings: 37%) and *S. mikatae* (SHPs: 46%; mappings: 38%). The synthetic hybrid resulting from the crossing of four parents, *S. cerevisiae* x *S. kudriavzevii* x *S. mikatae* x *S. uvarum*, also showed this trend with reduced values compared to the *S. uvarum* parent (SHPs: 3%; mappings: 11%), while the *S. cerevisiae* parent also showed reduced (11%) but similar percentages of SHPs to the mappings (12%). The genomes of the parents *S. mikatae* (SHPs: 45%; mappings: 40%) and *S. kudriavzevii* (SHPs: 41%; mappings: 38%) were predominant in this synthetic hybrid (Tables S5D; S3D).

Inference of subgenomic phylogenies from SHPs

The multiple alignments of SHPs used for phylogenetic analysis contain 5,958,612, 980,493, 9,664,388, and 1,540,088 total sites for the *Brachypodium*, *Brassica*, *Triticum-Aegilops*, and *Saccharomyces* groups, of which 4,378,851, 787,595, 7,165,502, and 1,441,544 are parsimony-informative, respectively (Suppl. Table 7).

The inferred phylogeny in the *Brachypodium distachyon* complex shows 100/100 SH-aLRT/UltraFast bootstrap support across all branches (Fig. 3A). A total of 642,818 SNPs from *Oryza sativa*, syntenic to the *B. distachyon* master genome, were included in the MSA to root the *Brachypodium* tree. Both *B. hybridum*-Bhyb26 subgenomes (S and D) are resolved as more ancestral than those of the recent *B. hybridum*-ABR113 subgenomes. Furthermore, these two subgenomic lineages, Bhyb26-S and Bhyb26-D, diverged earlier than their respective progenitor species lineages (*B. stacei* TE4.3 and ABR114; *B. distachyon* Bd21 and ABR2), suggesting that the ancestral *B. stacei* and *B. distachyon* parents of *B. hybridum* Bhyb26 went extinct or have not been sampled yet [51, 52]. This demonstrates the potential of our method applied to intra/inter-species studies to infer the different putative diploid progenitors (subgenomes), distinguishing within the same species the most ancestral and recent ones (Fig. 3A). The gwANI matrix (Suppl. Table 8 A) showed higher identities between *B. hybridum* ABR113 ecotype subgenomes D and S and the genomes of ecotypes of the diploid progenitor species *B. distachyon* and *B. stacei* (D-subgenome vs. *B. distachyon* ecotypes: 94–95%; S-subgenome vs. *B. stacei* ecotypes: 96–97%) than those obtained for the *B. hybridum* Bhyb26 ecotype subgenomes (D-subgenome vs. *B. distachyon* ecotypes: 87–88%; S-subgenome vs. *B. stacei* ecotypes: 85%) (Suppl. Table 8 A). This was reflected in the higher divergences of the Bhyb26 subgenomes from its diploid progenitor species compared to that of ABR113 as shown in the phylogenetic tree (Fig. 3A) and in agreement with previous finding [51, 52, 81].

The phylogeny of the *Brassica* complex showed high support across all branches. Only three branches had high support but below the recommended 80 and 95 SH-aLRT and UFboot thresholds (Fig. 3B). Two groups were distinguished based on the two diploid species, *Br. oleracea* and *Br. rapa*, progenitors of the allotetraploid *Br. napus*, and the respective subgenomes of the allotetraploid accessions clustered according to their diploid progenitor. Among *Br. rapa* subspecies, subsp. *tricularis* was more ancestral than subsp. *chinensis*. Regarding the Co-subgenome (*Br. oleracea* subgenome), the *Br. napus* R16G44 sample showed a more ancestral divergence than the other samples, which were resolved into two (R16GE06/ R16GE38 and R16GE39/ R16GE45) more recently evolved sister groups. In contrast, the Ar-subgenome clade (*Br. rapa*), showed the successive divergences of the R16G38, R16GE06 and R16GE39 lineages. As in the *Brachypodium distachyon* complex, different phylogenetic resolutions for one and the other subgenomic lineages were also detected in the *Brassica* allotetraploid species studied (Fig. 3B). The gwANI values of the Co (Bro) subgenomes of *Br. napus* showed a higher similarity among them (96.6–98%) than among the Ar (Brr) subgenomes (90.5–94.8%) (Suppl. Table 8B). These results, at the intraspecific level in *Br. napus* and focusing only in the syntenic regions shared between progenitors, agree with those obtained by Khan et al. (2016) using HANDS2 [39] software to compare the subgenomes of *Brassica carinata* (BBCC), *Br. juncea* (AABB), and *Br. napus*

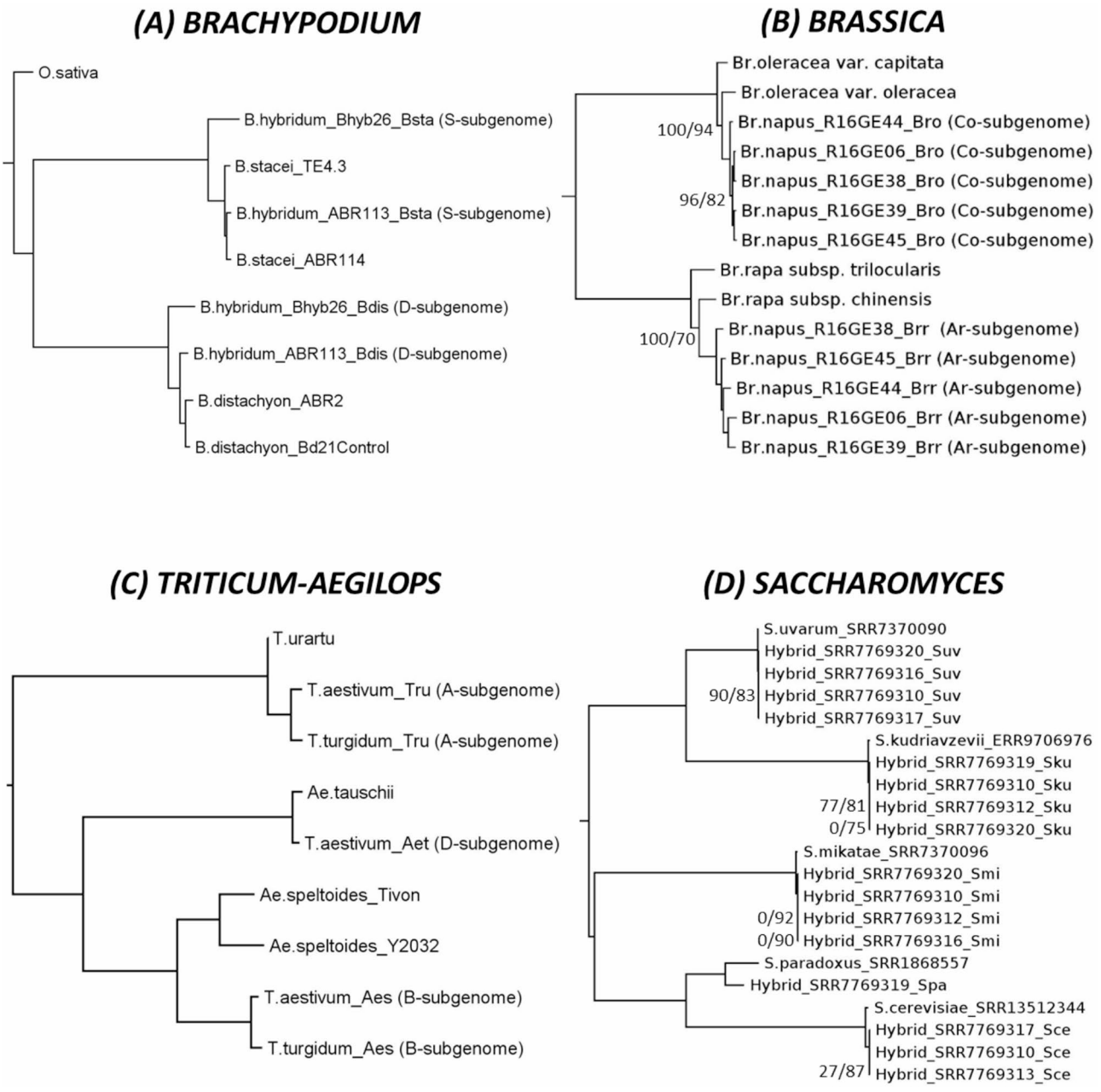


Fig. 3 Phylograms inferred using the SHPs alignment from the four diploid-polyploid complex datasets [**A***Brachypodium distachyon* complex, **B***Brassica* complex, **C***Triticum-Aegilops* complex, and **D***Saccharomyces* haploids and synthetic hybrids] used for the pipeline validation. Numbers indicate branches with SH-aLRT/UltraFast Bootstrap supports (BS) < 80/95; the remaining branches have 100/100 values

(AACC), indicating that the three C genomes of *Brassica* are more similar to each other than the three A genomes.

The inferred phylogeny in the *Triticum-Aegilops* complex showed 100/100 SH-aLRT/UltraFast Bootstrap support across all branches (Fig. 3C). The two main clades corresponded to *T. urartu* and the A-subgenomes of *T. aestivum* and T. *turgidum*, and the *Aegilops* clade, including *Ae. tauschii* and the D-subgenome of *T. aestivum*, and *Ae. speltoides*, both samples, with the B-subgenomes of *T. turgidum* and *T. aestivum* (Fig. 3C) matching previous studies [58, 82, 83]. The A subgenomes (Tru) of the allopolyploids *T. turgidum* and *T. aestivum* showed a high gwANI of 96.2%, as well as the B subgenomes (Aes) of the same species with 95.4% (Suppl. Table 8 C). The identity between the subgenomes of allopolyploids and their closest diploid ancestors varied considerably. The D subgenome of *T. aestivum* was 96.8% identical to *Ae. tauschii*. However, this decreased to 91.2 and 91.4% between the *T. turgidum* and *T. aestivum* A subgenomes and the *T. urartu* progenitor, and to 69–70% between the

B subgenomes and the *Ae. speltoides* progenitor (Suppl. Table 8 C).

The phylogeny of the yeast *Saccharomyces* showed two clades. One formed by the *species S. uvarum* and *S. kudriavzevii* and the subgenomes of hybrids shared with one of these parents, and the other clade formed by the species *S. mikatae* and the sister species *S. paradoxus* and *S. cerevisiae* (Fig. 3D). The subgenomes of the hybrids were grouped with the corresponding parental genomes. The support for the nodes of the main clades and groups was 100/100. This support dropped significantly within the subgenomic clades from the same parent. These clades often collapsed into polytomies due to the similarity of the recovered syntenic SNP sets, but were divergent from their parental genome (Fig. 3D). The gwANI for subgenomes within a single parent among the different synthetic hybrids studied was 100% (Suppl. Table 8D). Given the origin of these hybrids [64], their high level of identity is not surprising. However, the use of this set of synthetic samples resulting from multiparent crosses was used in the present study as a proof of concept to analyze the behavior of our protocol with such many reference genomes and the corresponding synthetic hybrids.

Discussion

Applications of AlloSHP in crop and wild population genomics

The use of AlloSHP enables scaling up of the number of individuals, populations, ecotypes, and/or accessions that can be analyzed simultaneously, and therefore related to each other. Together with the rise of pangenomes, all of this opens a new stage in the study of both wild populations and domesticated crops of diploid-allopolyploid complexes or groups. For instance, the analysis of syntenically aligned nuclear SNVs (single-nucleotide variants) from the *B. distachyon complex* pangenome revealed two independent origins for the allotetraploid *B. hybridum* using polymorphic orthologous nucleotide positions [52]. In bread wheat, the genome- and subgenome-wide base composition patterns were analyzed using polymorphic sites from multiple accessions of bread wheat, its diploid and tetraploid ancestral progenitors, and their reference genomes [84]. Wheat-T. *timopheevii* introgression lines were analyzed and homozygous introgressions detected through the development of a set of chromosome-specific Kompetitive allele-specific PCR (KASP) markers, where some of them were developed based on SNPs discovered through whole genome sequencing of *T. timopheevii*, being a majority of these KASP markers also found to be *T. timopheevii* subgenome specific [85]. In this context, subgenome-specific SNPs have been instrumental in designing markers targeting alleles in a particular genome, and for this reason, these SNPs have been curated and published in resources such as Ensembl Genomes [86].

Strengths and limitations of the protocol

This protocol does not require the targeted polyploids to be assembled nor annotated, as the sequenced reads are mapped directly against the reference genomes of the diploid progenitor species. This allows the selection of SHPs from both coding and intergenic regions (excluding repetitive regions). This approach allows us to reconcile a large number of polyploid subgenomes in a single alignment (Figs. 1 and 2). In the analyses performed on allotetraploids (*Brachypodium hybridum*, *Triticum turgidum*, *Brassica napus*), allohexaploid (*Triticum aestivum*), and synthetic yeast hybrids resulting from crosses of up to four different parents (synthetic *Saccharomyces* hybrids) (Suppl. Figure 1; Suppl. Table 1), we did not observe any limitation with respect to the ploidy of the allopolyploid under study, as long as the genomes of its extant diploid progenitor species are available and show syntenic regions between them.

Due to the mapping and syntenic alignment strategy, our protocol is only effective for inferring the subgenomes of allopolyploid organisms, not autopolyploids. Therefore, it is necessary to know which diploid species are the progenitor species or the closest extant relatives of the polyploid species under study, and the reference genomes of these diploids must be available. However, it is also possible to analyze polyploids whose progenitor species are uncertain, but it is mandatory to include as many diploid reference genomes as the putative diploid progenitors are involved in the allopolyploids under study. In this case, the resulting percentage of mappings could provide clues as to the most likely involvement of the progenitor species. In addition, previous tentative analyses, such as k-mer analysis (e.g., PolyCRACKER [87]) or simply a sequence similarity search (BLASTN [88]) of allopolyploid reads against the database, can provide clues about which species can be used as putative progenitors. However, caution should be exercised with allopolyploid species that have unknown, extinct, or closely related putative progenitor species.

Another limitation is related to the variants considered downstream, since to avoid conflicts between the positions extracted from the VCF file and the syntenic positions of the reference diploid progenitor genomes, indels are discarded and only SNPs are used for downstream analyses. Furthermore, the predetermined elimination of heterozygous sites, whose presence can complicate the identification of SHPs, can lead to a significant reduction in the number of final informative positions, especially in studies of wild populations.

Some sources of bias may arise from the quantity and sequencing quality of the reads, as well as the assembly

quality of the diploid reference genomes used. Other sources of bias that must be considered are inherent to the mapping and variant calling processes and are linked to the quality of the reads, reference genomes, and the complexity of those genomes [27]. To this end, the user must set thresholds and filters for FastQC, mapping quality, syntenic blocks, additional VCF filters resulting from the mappings and variant calling tools used.

AlloSHP has limitations in how it handles genomic rearrangements, such as introgressions and translocations, in allopolyploid genomes. Our pipeline is limited to mappings produced against included reference genomes. Therefore, chromosomal rearrangements arising from polyploidization events not captured in progenitor genomes will not be detected. However, we propose the following two-step approach for introgression detection using AlloSHP: (i) blast unmapped reads from allopolyploid samples against genome databases to infer candidate species that might have participated in the introgression event, and (ii) add the chromosome(s) involved to the concatenated diploid reference genomes. This will ultimately produce an "extra subgenome" from reads mapped against the chromosomes involved in the introgression into the polyploid genome. Note that this approach has not been tested in the present study. Furthermore, the percentage of SHPs recovered may conflict with the thresholds established for detecting artifactual subgenomes. Regarding the treatment of other chromosomal rearrangements, such as translocations, AlloSHP does not treat them in any special way as long as reads from those segments match the diploid parents used as references.

Although we have not found any limitation of our pipeline with respect to the size of the samples to be analyzed, it must be considered that some of the algorithms can generate bulky output files (tens of GB in the case of the *Triticum-Aegilops* complex) and might have high RAM requirements, with peaks that can exceed 100 GB in the case *of Triticum-Aegilops*. The size of the output files, the required RAM and the processing time are directly related to the size of the VCF input file and the size and number of reference genomes used (Suppl. Table 9 A, B).

Technical considerations of syntenic alignment, mapping and variant calling steps

Some of the technical and methodological considerations that users should consider are related to obtaining the VCF file, a step prior to using our pipeline. Although this study does not propose optimal filtering parameters and thresholds for the mapping and variant calling process in the polyploid genome analysis, there are previous studies and reviews such as those conducted by Clevenger et al. , Cooke et al. and Phillips [27, 89, 90], that suggest some values for these parameters as applied to polyploidy species studies. In any case, the parameters should be fine-tuned based on the polyploid organism and sequencing data [91]. Likewise, the parameters applied in the CGaln software used to obtain the syntenic regions must also be optimized in each case, although in this work those indicated gave the best results for the allopolyploid organisms studied, which suppose a good representation due to their wide range of genome sizes and ploidies.

The benchmark with single-copy orthologues revealed that synteny-based blocks mostly support the same SNPs, but often might capture pairwise alignments that are different to those resulting from protein sequence alignment and phylogenetic inference. The fact that the mismatched sites are less than 1kbp away suggests that the larger scale of genome alignments and the potential presence of tandem copies might be confounding SNP calling. As the final synteny-based results are parsed in BED format by the VCF2SYNTENY script, users can replace whole genome alignments and use orthology-based SNPs in cases where alignment quality might be an issue.

Threshold of artifactual subgenomes

Since our pipeline infers as many subgenomes as reference genomes are used, another methodological aspect to consider is setting the minimum threshold of SHPs required to establish a subgenome as plausible and not an artifact produced by non-specific mapping of reads against some of the reference genomes. These cases should be reduced by increasing parameters such as mapping quality (mapQ) or read depth (DP), while maintaining a balance so as not to miss an excessive number of SHPs.

We have established a cross-validation criterion to define artifactual subgenome based on the percentage of SHPs (see Suppl. Table 6) from non-specific mappings (see Suppl. Table 3) of diploid samples against the "erroneous" reference genome (the one that is not from the same species). All subgenomes presenting a percentage of SHP equal to or lower than the one recovered from the nonspecific mapping of the diploid sample, coinciding with the allopolyploid subgenome to be evaluated, will be defined as artifactual subgenomes and therefore eliminated. For example, the % SHP recovered from reads of the *B. distachyon* sample mapped against the *B. stacei* reference genome (from concatenated reference genomes) will set the threshold for identifying Bdis artifactual subgenomes in allopolyploid samples. Likewise, the % SHP recovered from reads of the *B. stacei* sample mapped against the *B. distachyon* reference genome (from concatenated reference genomes) will set the threshold for identifying Bsta artifactual subgenomes in allopolyploid samples. Below are the results of our case studies and the established criteria.

In our study, this threshold varies between the organisms studied, but the cross-validation criterion for setting the threshold for artifactual subgenomes is generalizable across the four databases evaluated. In *Brachypodium*, plausible *B. distachyon* (allopolyploid_Bdis) and *B. stacei* (allopolyploid_Bsta) subgenomes were those with a percentage of SNPs (in diploid species) or SHPs (in allopolyploids) greater than 0.9% and 0.8% (artifactual subgenome threshold) of SHP (Suppl. Table 6 A), according to the highest percentage of non-specific/artifactual SHPs recovered in the diploid samples (*B. distachyon*_Bsta [0.9 and 0.6%] and *B. stacei*_Bdis [0.8 and 0.2%]).

In *Brassica*, this percentage increased to 14.8% for *Br. oleracea* subgenomes (allopolyploid_Bro), as *Brassica oleracea* var. *capitata* had a non-specific mapping to the reference genome of 16.3% of the reads (Supplementary Table 3B), representing 14.8% of artifactual SHPs (Suppl. Table 6B). In *Br. rapa* subgenomes (allopolyploid_Brr), this threshold was reduced to 11.1% according to the non-specific SHP recovered on *Br. rapa* var. chinensis sample from mappings to *Br. oleracea* reference genome (Suppl. Table 6B).

In *Triticum* dataset, the thresholds were also raised to 19% and 12%, as these percentages of artifactual SHPs were recovered from the *Ae. speltoides* isolate Y2032 sample of reads non-specifically mapped to the *Ae. tauschii* and *T. urartu* reference genomes, respectively (Suppl. Table 6 C) and were fixed for removing allopolyploid_Aet and allopolyploid_Tru subgenomes with % SHPs equal to or lower these values. Similarly, the % of SHPs recovered from *Ae. tauschii* mappings against the *Ae. speltoides* reference genome showed a value of 1.9%, a value that was set as a threshold to eliminate artifactual subgenomes from allopolyploid_Aes with % SHP less than this value (Suppl. Table 6 C).

In the case of *Saccharomyces*, the percentages of artifactual SHPs from non-specific mappings ranged between 0% and the highest value of 0.36% from *S. paradoxus* sample mapping to *S. cerevisiae* reference genome. The percentages of SHPs from non-specific mappings between a haploid/diploid sample against reference genomes from another species were set as thresholds for the removal of artifactual subgenomes. (Suppl. Table 6D).

These thresholds might be biased by the quality of the assembly, the quality and number of reads, the evolutionary proximity and the genomic identity between the progenitor reference genomes used. To minimize this bias, it is recommended to include more than one sample of each diploid species, with said species coinciding, or as close as possible, with the ancestral progenitors of the allopolyploids. If the diploid progenitor species of the allopolyploids under study are known previously, this decision should not be problematic.

Specificity in the subgenomic assignment of SHPs

Two tests were performed on the *Brachypodium* data set to analyze the specificity of the mappings against the concatenated reference genomes. The first test consisted of using three reference genomes, including the two progenitor diploid species (*B. distachyon* and *B. stacei*; see Suppl. Table 10) and a third non-progenitor diploid species (*B. sylvaticum*; *Brachypodium sylvaticum* v1.1 DOE-JGI; http://phytozome.jgi.doe.gov/) of the allotetraploids *B. hybridum*. The results indicated that the two *B. hybridum* ecotypes (Bhyb26 and ABR113) showed reduced non-specific mappings against the non-progenitor reference genome *B. sylvaticum*, accounting for only 4.4% and 2.3% of the total reads, respectively (Suppl. Table 10).

The second test was designed to verify how the specificity of the mappings varies with the evolutionary proximity of the samples used as reference genomes. To do this, we mapped the reads of three *Brachypodium distachyon* ecotypes corresponding to each of the three clades/phylogenetic groups of this species (ABR2 from the Spanish group [S + from the S + T + clade], Bd21 from the Turkish group [T + from the S + T + clade], and BdTR8i from the EDF (Extremely Delayed Flowering) clade) against concatenated *B. distachyon* ecotypes genomes from the *B. distachyon* pangenome [14]. The EDF and S + T + clades diverged less than one million years ago, and the S + group is paraphyletic and, together with the monophyletic T + group, forms the Spanish-Turkish clade, which are evolutionarily close [14, 52, 92]. When we mapped the reads against two reference genomes (Suppl. Table 11 A), Tek2 from the EDF clade and ABR3 from the S + group (S + T + clade), 81% of the reads from the ABR2 sample (S + group) mapped against the ABR3 genome (S + group). The Bd21 reads (T + group) mapped mostly (64%) against the closest genome, in this case the ABR3 genome (S + group). The BdTR8i sample from the EDF clade mapped mostly (62%) against the Tek2 reference genome of this clade (Suppl. Table 11 A). When the same accessions, ABR2, Bd21 and BdTR8i, were mapped against the concatenated genomes of three *B. distachyon* ecotypes, one from each clade/group (Tek2 [EDF], ABR3 [S+] and BdTR12c [T+]) instead of only two, the percentages of mapped reads were more specific (Suppl. Table 11B). ABR2 reads (S+) mapped 13%, 58% and 29% against the EDF, S + and T + references, respectively. BdTR8i (EDF) mapped 55%, 27.5% and 17.5%, respectively (Suppl. Table 11B). Unlike the previous test, where the majority of the reads mapped to the reference genome of its own clade/group (Suppl. Table 11 A), in sample Bd21 (T+) 41% of reads mapped against the genome of the S + group, followed by 35% and 24% of the T + group and the EDF clade, respectively (Suppl. Table 11B). In this case, we used an extreme example, where the reference genomes were ecotypes of the same species,

and therefore they were extremely close, showing similar genomes [14, 52]. Therefore, for adequate application of our pipeline we need to consider the evolutionary proximity of the reference genomes, especially at the intraspecific level. Another aspect to consider is the quality of the assembly of the reference genomes. In this test, preliminary versions of the genomes assembled from the *B. distachyon* pangenome were used [14].

Accuracy and perspectives of the subgenomic phylogenetic reconstruction of Brachypodium, Triticum-Aegilops, and Brassica groups, and Saccharomyces haploid-synthetic hybrids

The phylogeny of *the B. distachyon* complex inferred from SHPs (Fig. 3A) showed the highest support for all the nodes and was consistent with that of Gordon et al. (2020) [52]. Both *B. stacei* and *B. distachyon* clustered with the respective S and D subgenomes of the allotetraploid *B. hybridum*, and the most divergent positions of both subgenomes of the older *B. hybridum* Bhyb26 ecotype were recovered with respect to their parents and the subgenomes of the more recent *B. hybridum* ABR113 ecotype. The *Brassica* phylogenetic tree showed divergences among and within the Ar and Co subgenomes of the five *Br. napus* accessions analysed (Fig. 3B). Multiple phylogenetic and population studies have been carried out on the diploid species *Br. oleracea* and *Br. rapa*, as well as on the allotetraploid *Br. napus* [93–97]. However, these studies require further expansion to confirm the multi-origin hypothesis of *Br. napus*. The present study used a small database to validate the pipeline, and using a large number of accessions is beyond this objective. Without further information on these samples and a much more extensive sample of *Br. rapa*, *Br. oleracea* and *Br. napus* accessions, no further conjectures on the inferred phylogeny can be made.

The genomes and phylogenies of the allopolyploids *T. aestivum* (6x) and *T. urartu* (4x) have been extensively studied, with *Ae. speltoides*, *Ae. tauschii* and *T. urartu* being proposed, albeit with some controversy, as their closest extant diploid progenitors [56, 58, 82, 83, 98–100]. The phylogeny of the *Triticum-Aegilops* complex recovered in the present study showed the expected groupings among the allopolyploid *Triticum* subgenomes (A, B and D) and their closest extant diploid ancestral relative [83, 100]. Comparing the number of SHPs detected in the *T. aestivum* and *T. turgidum* subgenomes, the number of SHPs in the A subgenomes of both species was similar (Suppl. Table 6 C). The *T. aestivum* D subgenome also recovered a high number of SHPs, even higher than the A subgenome. However, the number of SHPs in the B subgenome, although numerous, was reduced in both allopolyploid species by an order of magnitude compared to those recovered in the other subgenomes. Regarding base assignment to the B subgenome, Mithani et al. [38] detected that their HANDS protocol reduced the accuracy of base assignment for the B subgenome and attributed this to the fact that *Ae. speltoides*, used as the diploid progenitor species of the B genome, is evolutionarily closer to the *T. aestivum* B subgenome than the progenitor species of the A and D subgenomes. The *Saccharomyces* phylogeny shows polytomies in all subgenomes (Fig. 3D), with 100% gwANI for subgenomes within a single parent among the different synthetic hybrids studied (Suppl. Table 8D). Given the origin of these hybrids [64], their high level of identity is not surprising. However, the use of this set of synthetic samples resulting from multi-parent crosses was used in the present study as a proof of concept to analyze the precision of our protocol with such many reference genomes and the corresponding synthetic hybrids.

The proportions of SHPs in the synthetic hybrids (Suppl. Table 6D) from the different parents were generally consistent with the genomic contributions shown in Figs. 2b and 4 in Peris et al. [64], with the notable exception of SHPs from the *S. uvarum* progenitor. These were reduced compared to the actual proportions in each parental genome. This bias is due to the lower synteny between the *S. uvarum* genome, and the master or primary *S. cerevisiae* genome established in our protocol (Suppl. Table 4D).

The phylogenies obtained with AlloSHP in the four sets of diploid-allopolyploid species analyzed are consistent with previous studies that used different methodologies and nucleotide sequences (e.g. DNA and RNA loci). Therefore, AlloSHP can facilitate scaling up this type of analysis by increasing the number of allopolyploid and diploid progenitor populations and accessions, with the aim of studying their phylogenetic relationships at the subgenome level in greater detail.

Conclusions

A simple command-line pipeline has been developed to detect and extract SHPs from the homeologous subgenomes of allopolyploid species by mapping the reads against the concatenated reference genomes of their extant progenitor diploid species and reconciling SHPs into a subgenomic multiple sequence alignment using the syntenic positions of the reference genomes. This protocol allows to generate from a single VCF file the SHP alignment necessary to perform subgenome-scale phylogenetic studies of allopolyploid organisms, requiring only the genomes of their closest existing diploid progenitors and the genomic or transcriptomic sequences of the allopolyploids under study. This novel approach provides a valuable tool for the evolutionary study of allopolyploid species, both at the inter- and intra-specific levels, allowing the simultaneous analysis of a large number of

accessions and avoiding the complex process of assembling polyploid genomes.

Availability and requirements

Project name: AlloSHP. Project home page: https://github.com/eead-csic-compbio/AlloSHP. Operating system(s): Linux and MacOS. Programming language: Perl (90.9%), R (1.1%), Shell (4.0%) and Makefile (4.0%). Other requirements: Standard Linux utilities (gzip, grep, sort, perl, make, python3, g++, libdb-dev), Perl libraries (Getopt::Std, File::Temp, File::Basename, FindBin, DB_File, FileHandle) and third-party dependencies (Cgaln, GSAlign, Red, Red2Ensembl.py and gnuplot). License: Apache License 2.0.

Abbreviations

Aes	*Aegilops speltoides*
Aet	*Aegilops tauschii*
BAM	Binary Alignment Map format
BED	Browser Extensible Data format
Bdis	*Brachypodium distachyon*
Bsta	*Brachypodium stacei*
Bro	*Brassica oleracea*
Brr	*Brassica rapa*
DP	Total read depth
EDF	Extremely Delayed Flowering clade
gwANI	genome-wide Average Nucleotide Identity
homeoSNPs	homeologous Single Nucleotide Polymorphisms
HSPs	Homeolog-Specific Polymorphisms
KASP	Kompetitive Allele-Specific PCR
MSA	Multiple Sequence Alignment
NGS	Next-Generation Sequencing
Osat	*Oryza sativa*
QC	Quality Check
S+	Spanish group
SAM	Sequence Alignment Map format
Sce	*Saccharomyces cerevisiae*
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SHP	Single Homeologous Polymorphism
Sku	*Saccharomyces kudriavzevii*
Smi	*Saccharomyces mikatae*
Spa	*Saccharomyces paradoxus*
Suv	*Saccharomyces uvarum*
T+	Turkish group
Tru	*Triticum urartu*
VCF	Variant Call Format
WGD	Whole Genome Duplication

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13007-025-01458-6.

Supplementary Material 1.

Acknowledgements

The authors thank Miguel Campos (Escuela Politécnica Superior de Huesca, Universidad de Zaragoza) and Francesc Montardit-Tarda (Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas) for their valuable comments to improve the code and implementation of the AlloSHP scripts.

Author contributions

R.S., B.C.-M., and P.C. planned and designed the study and supervised the work. B.C.-M., R.S. wrote the source code and documentation. J.V. provided data. R.S. drafted the manuscript. All authors contributed to the discussion of the results, the writing, and approval of the final manuscript.

Funding

This work was supported by the Spanish Ministries of Economy and Competitivity (Mineco) and Science and Innovation (MICINN) [AGL2013-48756-R, CGL2016-79790-P, PID2019-108195GB-I00, PID2022-140074NB-I00], University of Zaragoza [UZ2016_TEC02] and CSIC [FAS2022_052]. RS was funded by a Mineco FPI PhD fellowship [BES-2013-066228], Mineco [EEBB-I-15-09760] and Ibercaja-CAI Mobility Grants 2016, Instituto de Estudios Altoaragoneses grant 2016 and RecoBar European project [PCI2022-135024-2]. BCM was funded in part by Fundacion ARAID. BCM, PC and RS were also funded by a European Social Fund/Aragon Government grants [A01-17R, A01-20R, A01-23R, A08-20R]. The work (proposal: https://doi.org/10.46936/10.25585/60001143) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

Data availability

The datasets generated and/or analysed during the current study are available in public databases (Phytozome13, Genome Portal and NCBI), Supplementary materials and AlloSHP repository, [https://github.com/eead-csic-compbio/AlloSHP] (https://github.com/eead-csic-compbio/AlloSHP).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 22 July 2025 / Accepted: 30 September 2025
Published online: 24 October 2025

References

1. Ramsey J, Schemske DW. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu Rev Ecol Syst. 1998; 29:467–501. http://www.annualreviews.org
2. Levin DA. The role of chromosomal change in plant evolution. Levin DA, editor. The role of chromosomal change in plant evolution. Oxford Series in Ecology and Evolution. New York: Oxford University Press; 2002.
3. Masterson J. Stomatal size in Fossil plants: evidence for polyploidy in majority of angiosperms. Science (1979). 1994;264. http://science.sciencemag.org/.
4. Stebbins GL. The evolutionary significance of natural and artificial polyploids in the family gramineae. Hereditas. 1949;35(1 S):461–85.
5. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. The frequency of polyploid speciation in vascular plants. Proceedings of the National Academy of Sciences (PNAS). 2009;106(33).
6. Halabi K, Shafir A, Mayrose I. PloiDB: the plant ploidy database. Volume 240. New Phytologist: John Wiley and Sons Inc; 2023. pp. 918–27.
7. Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, et al. Polyploidy and angiosperm diversification. Am J Bot. 2009;96(1):336–48.
8. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011;473(7345):97–100.
9. Renny-Byfield S, Wendel JF. Doubling down on genomes: polyploidy and crop plants. Am J Bot. 2014;101(10):1711–25.
10. Gregory TR, Mable BK. Polyploidy in animals. In: Gregory TR, editor. The evolution of the genome. Elsevier Inc; 2005. pp. 427–517.

11. Van de Peer Y, Meyer A. Large-scale gene and ancient genome duplications. In: Gregory TR, editor. The evolution of the genome. Elsevier Inc; 2005. pp. 329–68.
12. Albertin W, Marullo P. Polyploidy in fungi: evolution after whole-genome duplication. In: Proceedings of the Royal Society B: Biological Sciences. Vol. 279. Royal Society of London; 2012. pp. 2497–509.
13. Todd RT, Forche A, Selmecki A. Ploidy variation in fungi: polyploidy, aneuploidy, and genome evolution. Microbiol Spectr. 2017.
14. Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S et al. Extensive gene content variation in the brachypodium distachyon pan-genome correlates with population structure. Nat Commun. 2017;8(2184).
15. Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, et al. The barley pan-genome reveals the hidden legacy of mutation breeding. Nature. 2020;588(7837):284–9.
16. Jiao WB, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun. 2020;11(989).
17. Kang M, Wu H, Liu H, Liu W, Zhu M, Han Y, et al. The pan-genome and local adaptation of *Arabidopsis thaliana*. Nat Commun. 2023;14(6259).
18. Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee HT, Chan CKK, et al. The pangenome of hexaploid bread wheat. Plant J. 2017;90(5):1007–13.
19. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim HR, Martinez PA, et al. The pangenome of an agronomically important crop plant *Brassica Oleracea*. Nat Commun 2016;7(13390).
20. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. Nat Plants. 2020;6(8):914–20.
21. Schreiber M, Jayakodi M, Stein N, Mascher M. Plant pangenomes for crop improvement, biodiversity and evolution. Nat Rev Genet. 2024;25(8):563–77.
22. Kong W, Wang Y, Zhang S, Yu J, Zhang X. Recent advances in assembly of complex plant genomes. Genomics, proteomics and bioinformatics. Vol. 21. Beijing Genomics Institute; 2023. pp. 427–39.
23. Li K, Xu P, Wang J, Yi X, Jiao Y. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. Nat Commun 2023;14(6556).
24. Mason AS. Challenges of genotyping polyploid species. In: Batley J, editor. Plant genotyping methods in molecular biology. New York: Humana Press; 2015. http://www.springer.com/series/7651.
25. Soltis DE, Visger CJ, Blaine Marchant D, Soltis PS. Polyploidy: pitfalls and paths to a paradigm. Am J Bot. 2016;103(7):1146–66.
26. Wang Y, Yu J, Jiang M, Lei W, Zhang X, Tang H. Sequencing and assembly of polyploid genomes. In: Van de Peer Y, editor. Polyploidy methods in molecular biology. New York: Humana; 2023. http://www.springer.com/series/7651.
27. Phillips AR. Variant calling in polyploids for population and quantitative genetics. Appl Plant Sci. 2024;12(4).
28. Bombarely A, Coate JE, Doyle JJ. Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. PeerJ. 2014;2:e391.
29. Oxelman B, Brysting AK, Jones GR, Marcussen T, Oberprieler C, Pfeil BE. Phylogenetics of allopolyploids. Annu Rev Ecol Evol Syst. 2017; 48:543–57. https://doi.org/10.1146/annurev-ecolsys-110316-.
30. Than C, Ruths D, Nakhleh L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinform. 2008;9(322).
31. Than C, Nakhleh L. Species tree inference by minimizing deep coalescences. PLoS Comput Biol. 2009;5(9).
32. Jones G, Sagitov S, Oxelman B. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst Biol. 2013;62(3):467–78.
33. Marcussen T, Heier L, Brysting AK, Oxelman B, Jakobsen KS. From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae). Syst Biol. 2015;64(1):84–101.
34. Kamneva OK, Syring J, Liston A, Rosenberg NA. Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. BMC Evol Biol. 2017;17(180).
35. Sancho R, Inda LA, Díaz-Pérez A, Des Marais DL, Gordon S, Vogel JP, et al. Tracking the ancestry of known and 'ghost' homeologous subgenomes in model grass *Brachypodium* polyploids. Plant J. 2022;109(6):1535–58.
36. Page JT, Gingle AR, Udall JA, PolyCat:. A resource for genome categorization of sequencing reads from allopolyploid organisms. G3: Genes, genomes. Genetics. 2013;3(3):517–25.
37. Peralta M, Combes MC, Cenci A, Lashermes P, Dereeper A. SNiPloid: A utility to exploit high-throughput SNP data derived from RNA-Seq in allopolyploid species. Int J Plant Genomics. 2013;2013(890123).
38. Mithani A, Belfield E, Brown C, Jiang C, Leach L, Harberd N. HANDS: a tool for genome-wide discovery of subgenome-specific base-identity in polyploids. BMC Genomics. 2013;14(653).
39. Khan A, Belfield EJ, Harberd NP, Mithani A. HANDS2: Accurate assignment of homoeallelic base-identity in allopolyploids despite missing data. Sci Rep. 2016;6(29234).
40. Kulkarni R, Zhang Y, Cannon SB, Dorman KS. CAPG: comprehensive allopolyploid genotyper. Bioinformatics. 2023;39(1).
41. Page JT, Udall JA. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet. 2015;16(Suppl 2).
42. Session AM. Allopolyploid subgenome identification and implications for evolutionary analysis. Trends Genet. 2024;40(7):621–31.
43. Walden N, Schranz ME. Synteny identifies reliable orthologs for phylogenomics and comparative genomics of the brassicaceae. Genome Biol Evol. 2023;15(3).
44. Liu L, Chen M, Folk RA, Wang M, Zhao T, Shang F, et al. Phylogenomic and syntenic data demonstrate complex evolutionary processes in early radiation of the Rosids. Mol Ecol Resour. 2023;23:1673–88.
45. Zhao T, Zwaenepoel A, Xue JY, Kao SM, Li Z, Schranz ME, et al. Whole-genome microsynteny-based phylogeny of angiosperms. Nat Commun. 2021;12(1).
46. Thomas GWC, Ather SH, Hahn MW. Gene-Tree reconciliation with MUL-Trees to resolve polyploidy events. Syst Biol. 2017;66(6):1007–18.
47. Page JT, Liechty ZS, Huynh MD, Udall JA, BamBam. Genome sequence analysis tools for biologists. BMC Res Notes. 2014; 829(7).
48. Nakato R, Gotoh O. Cgaln: fast and space-efficient whole-genome alignment. BMC Bioinform. 2010; 11(224). http://www.biomedcentral.com/1471-2105/11/224.
49. Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, et al. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010;463(7282):763–8.
50. Joint Genome Institute. JGI Genome Portal. 2025. https://genome.jgi.doe.gov/portal/. Accessed 22 Apr 2025.
51. Scarlett VT, Lovell JT, Shao M, Phillips J, Shu S, Lusinska J, et al. Multiple origins, one evolutionary trajectory: gradual evolution characterizes distinct lineages of allotetraploid *Brachypodium*. Genetics. 2023;223(2).
52. Gordon SP, Contreras-Moreira B, Levy JJ, Djamei A, Czedik-Eysenberg A, Tartaglio VS, et al. Gradual polyploid genome evolution revealed by pan-genomic analysis of *Brachypodium hybridum* and its diploid progenitors. Nat Commun. 2020;11(1).
53. Shang L, He W, Wang T, Yang Y, Xu Q, Zhao X, et al. A complete assembly of the rice Nipponbare reference genome. Mol Plant. 2023;16. http://www.ricesuperpir.com/web/nip.
54. Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, et al. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica Oleracea*. Genome Biol. 2014;15(6).
55. Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, et al. Improved *Brassica Rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. Hortic Res. 2018;5(50).
56. Ling HQ, Ma B, Shi X, Liu H, Dong L, Sun H, et al. Genome sequence of the progenitor of wheat A subgenome *Triticum Urartu*. Nature. 2018;557(7705):424–8.
57. Wang L, Zhu T, Rodriguez JC, Deal KR, Dubcovsky J, McGuire PE, et al. *Aegilops tauschii* genome assembly Aet v5.0 features greater sequence contiguity and improved annotation. G3: Genes Genomes Genet. 2021;11(12).
58. Li LF, Zhang Z, Bin, Wang ZH, Li N, Sha Y, Wang XF, et al. Genome sequences of five sitopsis species of *Aegilops* and the origin of polyploid wheat B subgenome. Mol Plant. 2022;15(3):488–503.
59. NCBI. National Center for Biotechnology Information. 2025. https://www.ncbi.nlm.nih.gov/. Accessed 22 Apr 2025.
60. Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. Gigascience. 2017;6(11).
61. Engel SR, Wong ED, Nash RS, Aleksander S, Alexander M, Douglass E, et al. New data and collaborations at the *Saccharomyces* Genome Database: updated reference genome, alleles, and the Alliance of Genome Resources. Genetics. 2022;220(220).
62. Procházka E, Franko F, Poláková S, Sulo P. A complete sequence of *Saccharomyces paradoxus* mitochondrial genome that restores the respiration in S. cerevisiae. FEMS Yeast Res. 2012;12(7):819–30.
63. Yue JX, Li J, Aigrain L, Hallin J, Persson K, Oliver K, et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat Genet. 2017;49(6):913–24.

64. Peris D, Alexander WG, Fisher KJ, Moriarty RV, Basuino MG, Ubbelohde EJ et al. Synthetic hybrids of six yeast species. Nat Commun. 2020; 11(2085).
65. Andrews S. FastQC. A quality control tool for high throughput sequence data. 2019.
66. Bolger AM, Lohse M, Usadel B, Trimmomatic. A flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
68. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37(23):4572–4.
69. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO et al. Twelve years of samtools and BCFtools. Gigascience. 2021; 10(2).
70. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-Throughput sequencing data. Gigascience. 2021;10(2):1–6.
71. Lin HN, Hsu WL. GSAlign: an efficient sequence alignment tool for intra-species genomes. BMC Genomics. 2020; 21(182).
72. Girgis HZ, Red. An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics. 2015; 16(227).
73. Contreras-Moreira B, Filippi CV, Naamati G, García Girón C, Allen JE, Flicek P. K-mer counting and curated libraries drive efficient annotation of repeats in plant genomes. Plant Genome. 2021; 14e20143).
74. Williams T, Kelley C, Bersch C, Bröker HB, Campbell J, Cunningham R et al. gnuplot 6.0 An Interactive Plotting Program [Internet]. 2023. Available from: http://sourceforge.net/projects/gnuplot
75. Quinlan AR, Hall IM, BEDTools:. A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
76. Contreras-Moreira B, Naamati G, Rosello M, Allen JE, Hunt SE, Muffato M, et al. Scripting analyses of genomes in ensembl plants. Methods in molecular biology. Humana Press Inc.; 2022. pp. 27–55.
77. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2(4):e000056.
78. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37(5):1530–4.
79. Rambaut A, Figtree. 2018. Available from: https://tree.bio.ed.ac.uk/software/figtree/
80. Page AJ, Sjunnebo S, Seemann T, pANIto. Calculate genome wide average nucleotide identity (gwANI) for a multiFASTA alignment. 2018. Available from: https://github.com/sanger-pathogens/panito
81. Mu W, Li K, Yang Y, Breiman A, Yang J, Wu Y et al. Subgenomic stability of progenitor genomes during repeated allotetraploid origins of the same grass *Brachypodium hybridum*. Mol Biol Evol. 2023; 40(12).
82. Glémin S, Scornavacca C, Dainat J, Burgarella C, Viader V, Ardisson M et al. Pervasive hybridizations in the history of wheat relatives. Sci Adv [Internet]. 2019; 5(eaav9188). Available from: https://www.science.org
83. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, The International Wheat Genome Sequencing Consortium., Ancient hybridizations among the ancestral genomes of bread wheat. Science (1979). 2014; 345(6194).
84. Zhao Y, Dong L, Jiang C, Wang X, Xie J, Rashid MAR et al. Distinct nucleotide patterns among three subgenomes of bread wheat and their potential origins during domestication after allopolyploidization. BMC Biol. 2020; 18(1).
85. King J, Grewal S, Othmeni M, Coombes B, Yang CY, Walter N et al. Introgression of the *Triticum timopheevii* genome into wheat detected by Chromosome-Specific kompetitive allele specific PCR markers. Front Plant Sci. 2022; 13.
86. Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, Allen J, et al. Ensembl genomes 2020-enabling non-vertebrate genomic research. Nucleic Acids Res. 2020;48(D1):D689–95.
87. Gordon SP, Levy JJ, Vogel JP. PolyCRACKER, a robust method for the unsupervised partitioning of polyploid subgenomes by signatures of repetitive DNA evolution. BMC Genomics. 2019; 20(1).
88. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10.
89. Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA. Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. Mol Plant. 2015;8(6):831–46.
90. Cooke DP, Wedge DC, Lunter G. Benchmarking small-variant genotyping in polyploids. Genome Res. 2022;32(2):403–8.
91. Ning W, Meudt HM, Tate JA. A roadmap of phylogenomic methods for studying polyploid plant genera. Appl Plant Sci. 2024; 12e11580).
92. Sancho R, Cantalapiedra CP, López-Alvarez D, Gordon SP, Vogel JP, Catalán P, et al. Comparative plastome genomics and phylogenomics of *Brachypodium*: flowering time signatures, introgression and recombination in recently diverged ecotypes. New Phytol. 2018;218(4):1631–44.
93. An H, Qi X, Gaynor ML, Hao Y, Gebken SC, Mabry ME et al. Transcriptome and organellar sequencing highlights the complex origin and diversification of allotetraploid *Brassica Napus*. Nat Commun. 2019; 10(1).
94. Bird KA, An H, Gazave E, Gore MA, Pires JC, Robertson LD et al. Population structure and phylogenetic relationships in a diverse panel of *Brassica Rapa* L. Front Plant Sci. 2017; 8(321).
95. Gazave E, Tassone EE, Ilut DC, Wingerson M, Datema E, Witsenboer HMA et al. Population genomic analysis reveals differential evolutionary histories and patterns of diversity across subgenomes and subpopulations of *Brassica Napus* L. Front Plant Sci 2016; 7(525).
96. Lv H, Wang Y, Han F, Ji J, Fang Z, Zhuang M et al. A high-quality reference genome for cabbage obtained with SMRT reveals novel genomic features and evolutionary characteristics. Sci Rep. 2020; 10(12394).
97. Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica Napus*. Nat Plants. 2020;6(1):34–45.
98. Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science (1979). 2018; 361(eaar7191).
99. El Baidouri M, Murat F, Veyssiere M, Molinier M, Flores R, Burlot L, et al. Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). New Phytol. 2017;213(3):1477–86.
100. Petersen G, Seberg O, Yde M, Berthelsen K. Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). Mol Phylogenet Evol. 2006;39(1):70–82.

Publisher's Note