the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **109,** 1535–1558 doi: 10.1111/tpj.15650

Tracking the ancestry of known and 'ghost' homeologous subgenomes in model grass *Brachypodium* polyploids

Rubén Sancho[1,2], Luis A. Inda[1,3], Antonio Díaz-Pérez[1,4], David L. Des Marais[5,†], Sean Gordon[6], John P. Vogel[6,7], Joanna Lusinska[8], Robert Hasterok[8], Bruno Contreras-Moreira[2,9,‡,*] and Pilar Catalán[1,2,10,*]

[1]*Department of Agricultural and Environmental Sciences, High Polytechnic School of Huesca, University of Zaragoza, Huesca, Spain,*

[2]*Grupo de Bioquímica, Biofísica y Biología Computacional (BIFI, UNIZAR), Unidad Asociada al CSIC, Zaragoza, Spain,*

[3]*Instituto Agroalimentario de Aragón (IA2), Universidad de Zaragoza, Zaragoza, Spain,*

[4]*Instituto de Genética, Facultad de Agronomía, Universidad Central de Venezuela, Caracas, Venezuela,*

[5]*The Arnold Arboretum of Harvard University, Boston, Massachusetts, USA,*

[6]*DOE Joint Genome Institute, Berkeley, California, USA,*

[7]*Department of Plant and Microbial Biology, University of California, Berkeley, California, USA,*

[8]*Plant Cytogenetics and Molecular Biology Group, Institute of Biology, Biotechnology and Environmental Protection, Faculty of Natural Sciences, University of Silesia in Katowice, Katowice, Poland,*

[9]*Department of Genetics and Plant Breeding, Estación Experimental de Aula Dei-Consejo Superior de Investigaciones Científicas, Zaragoza, Spain, and*

[10]*Tomsk State University, Tomsk, Russia*

Received 8 June 2021; revised 10 December 2021; accepted 20 December 2021; published online 24 December 2021.

*For correspondence (e-mails pcatalan@unizar.es; bcontreras@eead.csic.es).

[†]Present address: Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[‡]Present address: Ensembl Plants, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK

SUMMARY

Unraveling the evolution of plant polyploids is a challenge when their diploid progenitor species are extinct or unknown or when genome sequences of known progenitors are unavailable. Existing subgenome identification methods cannot adequately infer the homeologous genomes that are present in the allopolyploids if they do not take into account the potential existence of unknown progenitors. We addressed this challenge in the widely distributed dysploid grass genus *Brachypodium*, which is a model genus for temperate cereals and biofuel grasses. We used a transcriptome-based phylogeny and newly designed subgenome detection algorithms coupled with a comparative chromosome barcoding analysis. Our phylogenomic subgenome detection pipeline was validated in *Triticum* allopolyploids, which have known progenitor genomes, and then used to infer the identities of three subgenomes derived from extant diploid species and four subgenomes derived from unknown diploid progenitors (ghost subgenomes) in six *Brachypodium* polyploids (*B. mexicanum, B. boissieri, B. retusum, B. phoenicoides, B. rupestre* and *B. hybridum*), of which five contain undescribed homeologous subgenomes. The existence of the seven *Brachypodium* progenitor genomes in the polyploids was confirmed by their karyotypic barcode profiles. Comparative phylogenomics of nuclear versus plastid trees allowed us to formulate hypothetical homoploid hybridizations and allo- and autopolyploidization scenarios that could have generated the six *Brachypodium* polyploids.

Keywords: *Brachypodium*, chromosomal barcodes, 'ghost' progenitor genomes, phylogenomic subgenome detection pipeline, polyploids.

INTRODUCTION

While the genomic origins of some polyploid plants have been deduced using comparative genomics (e.g. wheats; Marcussen et al. 2014a; Appels et al. 2018), deciphering the genomic history of many allopolyploids has proven to be challenging when the progenitor species are extinct or unknown (Soltis and Soltis 2016), or when the parental genomes are highly similar (Brassac and Blattner 2015; Kamneva et al. 2017). Incomplete genome assemblies further complicate the delineation of homeologous genomes in

allopolyploid plants, which is the typical scenario in angiosperms outside of a few well-studied plants (Soltis et al. 2016; Scholthof et al. 2018). Multispecies coalescent species trees and networks, together with syntenic read-mapping phylogenetic approaches, have successfully reconstructed the history of the homeologous genomes of some allopolyploid plants (Bombarely et al. 2014; Bertrand et al. 2015; Marcussen et al. 2015; Novikova et al. 2016; Oxelman et al. 2017). However, most of these cases are allopolyploids with known diploid ancestors. Few studies have identified subgenomes that were derived from unknown diploid ancestors (hereafter referred to as ghost subgenomes; Kamneva et al. 2017) or have explicitly searched for 'ghost' subgenomes (Marcussen et al. 2015). Coalescence-based methods account for incomplete lineage sorting (ILS) events across the gene trees and create consensus merging scenarios for the subgenomes within allopolyploids (Marcussen et al. 2015; Kamneva et al. 2017). Nonetheless, these protocols are challenging due to the computational overhead for the likelihood or Bayesian-based methods (e.g. AlloppNET and AlloppMUL models; Jones 2017; Oxelman et al. 2017). Additionally, selecting the optimal hybridization scenarios is impeded in cases in which the progenitor diploid genomes are unknown and as the number of possible subgenome combinations increases with ploidy (Bertrand et al. 2015; Marcussen et al. 2015).

Allopolyploids are common in the grass family and account for 70% of the current species (Stebbins 1949; Kellogg 2015). The genus *Brachypodium* has been developed as a model system for cereals and biofuel grasses (Scholthof et al. 2018; Catalán and Vogel 2020). The development of numerous experimental and genomic resources has made this pooid genus an indispensable tool for investigating many aspects of the biology and evolution of grasses, and monocots more broadly, and as a tool for translating fundamental biological insights to crop species (Catalán and Vogel 2020). Recent phylogenetic studies suggest that allopolyploidy has been a prevalent speciation mechanism in *Brachypodium* (Catalán et al. 2016; Díaz-Pérez et al. 2018) and, indeed, that most allopolyploid *Brachypodium* species likely resulted from crosses of dysploid progenitor species that had different basic chromosome numbers (Betekhtin et al. 2014; Díaz-Pérez et al. 2018). The best-known case is the annual allotetraploid *B. hybridum* ($2n = 30$, $x = 10 + 5$), which was derived from the cross and subsequent genome doubling of the diploid *B. stacei*-type ($2n = 20$, $x = 10$) and *B. distachyon*-type ($2n = 10$, $x = 5$) progenitors (Catalán et al. 2012, 2014; López-Álvarez et al. 2012; Shiposha et al. 2020; Gordon et al. 2020). The re-creation of a stable synthetic allotetraploid that phenotypically resembles the natural *B. hybridum* corroborated the allopolyploid origin of this neopolyploid species (Dinh Thi et al. 2016). In contrast, the evolutionary history of the perennial *Brachypodium* allopolyploids is more intriguing due to their full or partial ghost subgenomes, which have only been studied with a limited set of nuclear and plastid loci (Catalán et al. 2016; Díaz-Pérez et al. 2018).

Here, we use a novel approach (phylogenomic subgenome detection; PhyloSD) to uncover the homeologous subgenomes in *Brachypodium* allopolyploids, and reconstruct their evolution focusing specifically on species whose diploid progenitors are extinct or unknown. We used the well-known phylogeny and diploid progenitor genomes of the allopolyploid *Triticum* species (Marcussen et al. 2014a) to benchmark our algorithms. Then, we applied the algorithms to six putative polyploid *Brachypodium* species in an attempt to retrieve their reticulate history. We used our subgenome detection algorithms as an *a priori* assignment of homeologs to the genomes of their hypothetical diploid progenitors. We further validated the computational pipeline and reconstructed a robust phylogeny for the genomes and homeologous subgenomes of 12 *Brachypodium* species and ecotypes. These computational approaches were validated using fluorescence *in situ* hybridization (FISH)-based comparative chromosome barcoding (CCB), which enables the specific painting of whole chromosomes or specific regions. CCB proved to be effective in tracking the structural and evolutionary trajectories of individual chromosomes and whole karyotypes in some dicots (e.g. *Arabidopsis thaliana* and its relatives; Lysak et al. 2006) and monocots (e.g. rice; Hou et al. 2018), and recently contributed to dissecting the karyotype organization of some *Brachypodium* species (Lusinska et al. 2019). Our combined phylogenomic and cytomolecular strategy enables us to propose hypotheses about the identities of both previously known as well as unknown progenitor genome donors in the studied *Brachypodium* polyploid species and to infer their times of origin. Genome and transcriptome-wide analyses together with comparative phylogenomics or network analyses of nuclear and chloroplast genomes have been used to postulate potential hybridization and polyploidization scenarios that gave rise to the plant polyploids (Doyle and Egan 2010; Marcussen et al. 2015; Oxelman et al. 2017). The use of these comparative nuclear versus plastome approaches in our study allowed us to advance hypothetical evolutionary scenarios that explain the likely origins of allo- and autopolyploid *Brachypodium* species.

RESULTS

The PhyloSD pipeline

The PhyloSD pipeline employs three sequential algorithms: *Nearest Diploid Species Node*; *Bootstrapping Refinement*; and *Subgenome Assignment* (Figure 1a–c). The input data are a set of pre-computed multiple sequence alignments (MSAs) of coding sequences (CDS) and transcripts that contain the ingroup diploid orthologs and polyploid homeologs and outgroup orthologs. The pipeline consists of: (i) a computational filtering step; (ii) labeling the

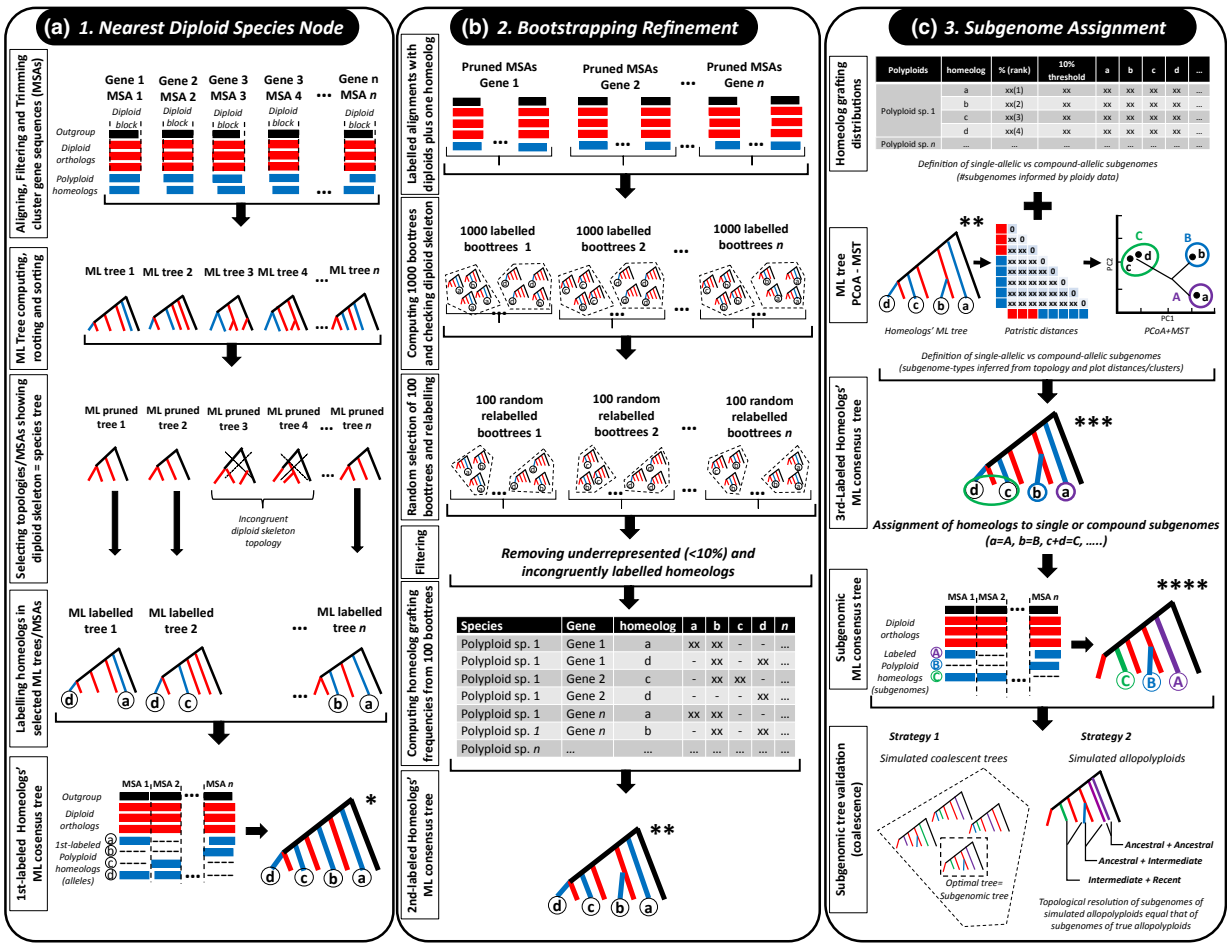

Figure 1. A summarized workflow of the phylogenomic subgenome detection (PhyloSD) pipeline highlighting the three *Nearest Diploid Species Node* (a), *Bootstrapping Refinement* (b) and *Subgenome Assignment* (c) algorithms. Black, red and blue colors indicate the diploid outgroup and diploid and polyploid ingroup sequences, respectively. The 1st-Labeled Homeologs' ML consensus tree (*) that was obtained from the *Nearest Diploid Node* algorithm (a) was fine-tuned by the *Bootstrapping Refinement* algorithm (b) resulting in the 2nd-Labeled Homeologs' ML consensus tree (**). The 2nd-Labeled Homeologs' ML consensus tree (**) was readjusted by the *Subgenome Assignment* algorithm (c) resulting in the 3rd-Labeled Homeologs' ML consensus tree (***). This tree was genomically relabeled, after the assignment of homeologs to single or compound subgenomes, resulting in the Subgenomic ML consensus tree (****). ML, maximum likelihood; MSA, multiple sequence alignment; PCoA–MST, principal coordinate analysis and superimposed minimum spanning tree.

homeologs; and (iii) allocating the homeologs to subgenomes; a detailed description of the pipeline is provided in Appendix S1 (see also Figure S1).

The *Nearest Diploid Species Node* algorithm labels the homeologous sequences according to their grafting positions with respect to the nearest diploid species in the optimal diploid skeleton tree and its stem branch (Figure 1a). MSAs with missing diploid sequences and non-overlapping alignment blocks are discarded. Maximum likelihood (ML) phylogenetic trees are subsequently estimated for each of the curated MSAs, thereby obtaining exploratory gene trees (Figure 1a). These trees are further filtered, keeping only the most frequent partitions with a diploid skeleton topology congruent with that of the diploid species tree. The diploid species tree was obtained from parallel coalescent analyses (ASTRAL, STEAC, STAR). Then, the homeologs are labeled in each partition tree according to their grafting positions concerning the nearest diploid species in the optimal diploid skeleton tree using *ad hoc* labeling rules ('a' to 'i'; Figures 1a, 2a and 3a), and assuming that each homeolog type would represent a subgenome in the polyploid. A labeled ML consensus tree is then computed from all of the labeled partitions (Figures 1a and S1).

The *Bootstrapping Refinement* algorithm tests the labeling of the homeologs. Bootstrap analyses are performed to generate patterns of the branch distribution for each homeolog type, assuming that a single homeolog could have been grafted in topological-vicinity branches by accident. The labeled MSAs from the previous step are pruned and used to generate new datasets, each of which contains all of the diploid orthologs plus one polyploid homeolog at a time (Figure 1b). Next, 1000 labeled ML bootstrapping trees (boottrees) are computed for each pruned alignment.

Table 1 Homeolog allelic and subgenomic datasets: (a) *Triticum–Aegilops*; (b) *Brachypodium*

(a)

	Taes		Ttur	
	#	%	#	%
Homeolog type				
a	3	2.2*	4	4.7*
b	1	0.7*	1	1.2*
c	35	25.5	31	36.0
d	5	3.6*	4	4.7*
e	2	1.5*	1	1.2*
f	45	32.8	–	–
g	5	3.6*	5	5.8*
h	4	2.9*	7	8.1*
i	37	27.0	33	38.4
Total	137	100	86	100
Inferred subgenome				
A (a)	–	–	–	–
B (b)	–	–	–	–
C (c)	35	29.9	31	48.4
D (d)	–	–	–	–
E (e)	–	–	–	–
F (f)	45	38.5	–	–
G (g)	–	–	–	–
H (h)	–	–	–	–
I (i)	37	31.6	33	51.6
Total	117	100	64	100

(b)

	Bmex		Bboi		Bret		Bhyb		Brup		Bpho6		Bpho422	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Homeolog type														
a	89	47.8	70	39.5	26	12.9	2	0.9*	–	–	3	1.4*	1	0.5*
b	43	23.1	31	17.5	8	4.0*	126	55.8	1	0.5*	–	–	1	0.5*
c	39	21.0	39	22.0	38	18.9	2	0.9*	–	–	3	1.4*	2	1*
d	6	3.2*	8	4.5*	2	1.0*	96	42.5	1	0.5*	3	1.4*	2	1*
e	9	4.8*	22	12.4	59	29.4	–	–	42	21	54	25	46	23
f	0	0.0*	3	1.7*	15	7.5*	–	–	24	12	43	20	38	19
g	0	0.0*	3	1.7*	31	15.4	–	–	58	29	50	23	57	28
h	0	0.0*	–	–	12	6.0*	–	–	40	20	26	12	25	12
i	0	0.0*	1	0.6*	10	5.0*	–	–	34	17	31	15	30	15
Total	186	100	177	100	201	100	226	100	200	100	213	100	202	100
Inferred subgenome														
A1 (a + b + c)	171	100	–	–	–	–	–	–	–	–	–	–	–	–
A2 (a + b + c + e; a + c)	–	–	162	100	64	41.6	–	–	–	–	–	–	–	–
B (b)	–	–	–	–	–	–	126	56.8	–	–	–	–	–	–
D (d)	–	–	–	–	–	–	96	43.2	–	–	–	–	–	–
E1 (e + g)	–	–	–	–	90	58.4	–	–	–	–	–	–	–	–
E2 (e)	–	–	–	–	–	–	–	–	42	21.2	54	26.5	46	23.47
G (f + g + h + i)	–	–	–	–	–	–	–	–	156	78.8	150	73.5	150	76.53
Total	171	100	162	100	154	100	222	100	198	100	204	100	196	100

Number (#) and percentage (%) of polyploid homeolog alleles that were detected in the studied species by our *Nearest Diploid Species Node* and *Bootstrapping Refinement* algorithms using the aligned genes (a) and core transcripts (b). The homeologs were classified into nine homeolog types ('a' to 'i') according to their grafting positions in the diploid skeleton tree (Figures 2a and 3a). Those occurring in less than 10% of the selected genes in each accession (see asterisks) were removed from the downstream analyses. The inferred homeologous subgenomes of the studied polyploids that were selected and labeled according to the S*ubgenome Assignment* algorithm and the ploidy level of each polyploid species that was inferred by cytogenetic data.

Taes, *Triticum aestivum*; Ttur, *Triticum turgidum*; Bmex, *Brachypodium mexicanum*; Bboi, *Brachypodium boissieri*; Bret, *Brachypodium retusum*; Bhyb, *Brachypodium hybridum*; Brup, *Brachypodium rupestre*; Bpho, *Brachypodium phoenicoides*.

The robustness of the grafted homeologs is assessed using a low bootstrap support cut-off (BS < 10%) in each of the consensus boottrees with a congruent diploid skeleton topology. Poorly represented homeolog types are removed. Then, 100 boottrees are randomly selected from each group and the homeologs are relabeled. The homeolog grafting frequencies from each of the 100 boottrees are computed, and the resulting homeologs' ML consensus tree is constructed (Figures 1b and S1).

The *Subgenome Assignment* algorithm allocates the homeologs to the corresponding defined polyploid subgenomes based on: (i) the frequency ranks of their grafting distributions (Figure 1c); and (ii) the clusters of homeologs in a principal coordinate analysis (PCoA) with a superimposed minimum spanning tree (MST) plot (Figure 1c). The PCoA–MST is obtained from the pairwise patristic distances computed from the ML consensus tree retrieved in the second step (Figures 1b and S1). The bootstrap grafting distributions of the homeologs are evaluated to determine their circumscription to a single or a few contiguous branches of the species tree (Figure 1c). The homeologs are assigned to single-type (single-allelic) subgenomes if they were grafted to single branches with the highest frequency, and the remaining graftings were below the cut-off threshold (≤ 10% of the main grafting frequency). They are assigned to compound-type (multi-allelic) subgenomes if the secondary and subsequent grafting frequencies are above the threshold. The 10% threshold was selected as the minimum threshold to discard the clustering of spurious graftings in closely related branches. It should be noted that a low threshold is indicative of clear separation and selection of alternative branch graftings. By contrast, a higher threshold would have placed less related branch graftings (different homeologs) into the same compound subgenome. The most frequently observed homeolog types ('a', 'b', ...) are selected according to the expected number of subgenomes that are suggested by ploidy (two for tetraploids and three for hexaploids) and are re-coded as subgenomes ('A', 'B', ...), and the low-frequency homeolog types incompatible with the ploidy level of the polyploid are discarded (Figures 1c, 2b,c and 3b,c). The labeled subgenomic MSAs are used to compute the subgenomic consensus ML tree (Figures 1c, 2c and 3c).

Benchmarking the PhyloSD pipeline in the *Triticum–Aegilops* allopolyploid complex

The initial *Triticum–Aegilops* dataset consists of 275 ortholog clusters that were obtained from Marcussen et al. (2014a,b; see Experimental Procedures) of which only 48 MSAs with 236 homeologs (having a diploid skeleton topology that was congruent with that of the coalescent-based species tree; Figure S2a) remained after the filtering steps of the first algorithm. The homeologs were labeled according to the *ad hoc* rules that are presented in Figure 2(a). When the incongruently labeled and underrepresented homeologs were removed from the second algorithm, a total of 48 MSAs and 181 homeologs were obtained. These were used to compute the *Triticum–Aegilops* ML consensus tree (Figure 2b). The grafting distributions of the homeologs (Table S1a) and the PCoA–MST clusterings (Figure S2b) of the third algorithm presented a simple scenario in which all of the selected homeologs were assigned to single subgenomes that contained only one homeolog type (Tables 1a and S1b). Thus, homeolog 'c' corresponded to the subgenome C that is present in *T. turgidum* and *T. aestivum*, 'f' to the subgenome F that is present in *T. aestivum*, and 'i' to the subgenome I that is present in *T. turgidum* and *T. aestivum* (Figure 2b; Table S1b,c). The highly supported subgenomic tree (Figure 2c) recovered the expected phylogeny for the studied diploid *Triticum* and *Aegilops* taxa, and demonstrated sister relationships of the homeologous C, F and I subgenomes to their respective *Ae. speltoides*-like, *Ae. tauschii*-like and *T. urartu*-like diploid progenitor genomes, which corroborated the accuracy of our subgenome assignment method. According to Marcussen et al. (2014b), our 'C' subgenome is equivalent to the subgenome B, 'F' to the subgenome D and 'I' to the subgenome A, respectively, in the nomenclatural system of the Triticeae genomes.

Retrieving the known and ghost subgenomes of the *Brachypodium* polyploids

The initial *Brachypodium* dataset contained 3675 transcript clusters (Tables S2, S3 and S4a,b), which were obtained from a wide transcriptomic analysis (see Experimental Procedures). These were reduced to 329 MSAs with 1965 homeologs after the successive filtering steps of the first algorithm (Figures 1a and S3a). The homeologs were labeled according to the *ad hoc* rules presented in Figure 3 (a). The bootstrapping refinement algorithm left 322 MSAs and 1307 homeologs (Table S4a; Data S1), which were used to build the ML consensus tree (Figure 3b). In contrast to the wheats, *Brachypodium* had an excess of homeolog types that required ranking, selecting the most frequent types and merging some of them in order to retrieve the plausible subgenomes of some polyploids (Tables 1b and S4a,b). The allotetraploid *B. hybridum* was the exception; it fitted a simple scenario where its 'b' and 'd' homeologs corresponded to its subgenomes B (*B. stacei*-type) *and D (B. distachyon*-type). *For the remaining Brachypodium* polyploids, the relatively close 'a' and 'c' homeolog types (plus 'b' in some species) were considered to be variants of the ancestral A subgenomes, the separate 'e' homeolog types were assigned to the intermediately evolved E subgenomes, and the recently diverged and close 'f', 'g', 'h' and 'i' homeolog types were assumed to represent variants of a single core perennial clade G subgenome (Figure 3b,c; Table S4b,c). In addition, the clear

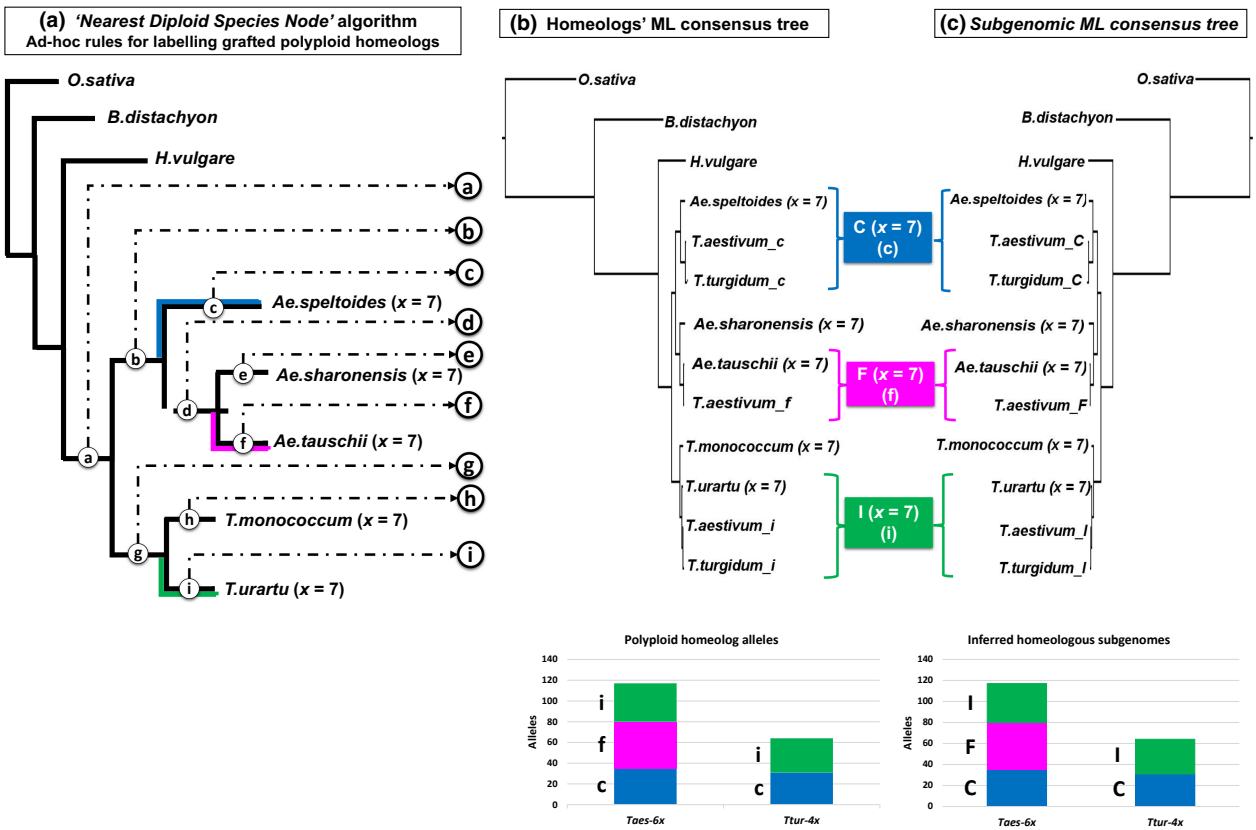

Figure 2. (a) Schematic *Triticum–Aegilops* tree illustrating the diploid skeleton tree (thick black branches) of the orthologous diploid genome sequences (*Ae. speltoides, Ae. sharonensis, Ae. tauschii, T. monococcum, T. urartu*) of $x = 7$ showing the *ad hoc* labeling rules (lowercase letters, 'a'–'i') for the grafting positions of the *Triticum* polyploid homeolog sequences according to the *Nearest Diploid Species Node* algorithm; (b) *Triticum–Aegilops* homeologs' ML consensus tree based on 48 core genes and 181 homeologs (Table 1a) with the polyploid homeolog sequences labeled according to the *Nearest Diploid Species Node* algorithm ('c', 'i', 'f'); (c) *Triticum–Aegilops* subgenomic ML consensus tree based on 48 core genes with the homeolog subgenomes labeled according to the *Subgenome Assignment* algorithm ('C', 'F', 'I'; Table 1a; Figure S2b). *Oryza sativa, Brachypodium distachyon* and *Hordeum vulgare* were used as the outgroups. Asterisks indicate branches with SH-aLRT/UltraFast Bootstrap support (BS) <80/95; the remaining branches have 100/100 values. The bar diagrams represent the frequencies of the homeologs in each polyploid and their assignments to their respective homeologous subgenomes (Table 1a). ML, maximum likelihood.

divergence of the *B. mexicanum* 'a' homeolog type from that of *B. boissieri* and *B. retusum* in the ML tree and the PCoA–MST plot (Figures 3b and S3b) supported their respective assignment to the independent ancestral A1 and A2 subgenomes. In contrast, the sequential divergences of the *B. retusum* 'e' homeolog type from that of *B. rupestre* and *B. phoenicoides* in the ML tree (Figure 3b) maintained their respective assignments to the independent intermediate E1 and E2 subgenomes (Figures 3c and S3b; Table 1b). In this sense and considering the estimated ploidy levels of the studied polyploids (Table S2), the subgenomic assignments were constrained as follows: the tetraploid *B. mexicanum* 'a', 'b' and 'c' homeolog types were assigned to the ancestral A1 subgenome; the hexaploid *B. boissieri* 'a', 'b', 'c' and residual 'e' homeolog types to the ancestral A2 subgenome; the tetraploid *B. retusum* 'a' and 'c' homeolog types to the ancestral A2 subgenome and intermediate 'e' (plus recent 'g') homeolog types to the intermediate E1 subgenome; the tetraploids *B. rupestre* and *B. phoenicoides* intermediate 'e' homeolog types to the intermediate E2 subgenome and the recent core perennial clade 'f', 'g', 'h' and 'i' homeolog types to the recent G subgenome (Figures 3c and S3b; Tables 1b, S2 and S4b, c).

The correct identification and correspondence of transcripts to orthologous (true homeolog) alleles was further assessed in the allotetraploid *B. hybridum* (accession BdTR6g) and in one of its extant diploid progenitor species (*B. stacei* accession TE4.3) using comparative matching analyses of assembled transcripts to reference transcriptomes and to annotated reference genomes (see Experimental Procedures). Transcript data from the other diploid progenitor species, *B. distachyon*, were already retrieved from its reference transcriptome (accession Bd21; Table S2). In the first reference-transcriptome matching analysis, Blastn analysis of the 322 *B. stacei* TE4.3 and 222 *B. hybridum* BdTR6g transcripts used in this study to their respective *B. stacei* ABR114 and *B. hybridum* ABR113

reference transcriptomes showed high identities for the first matches to their corresponding reference primary transcripts (>98%; Data S2a,b), and low identities for the second-best matches to any other transcript (<82%, *B. stacei*) or diverse quality matches to primary transcripts (*B. hybridum*), though those with high scores (>95%) matched to the reference transcripts from the other (homeologous) subgenome (Data S2b). In the second reference-genome matching analysis, Illumina reads from the resequenced genomes of the *B. stacei* TE4.3 and *B. hybridum* BdTR6g accessions were used to assemble their encoding genes by mapping their respective reads to the reference genomes of *B. stacei* (ABR114) and *B. hybridum* (ABR113; see Experimental Procedures). These genic datasets were then used to extract the assembled genes that encoded the 322 *B. stacei* TE4.3 and 222 *B. hybridum* BdTR6g transcripts using a Blastn matching strategy (Data S3a,b). One *B. stacei* TE4.3 gene could not be assessed as it was not annotated in the reference genome of this species (Data S3a). The first matches of the assembled genes to their respective reference genes showed very high identities (> 99%) and long sequence overlaps (*B. stacei*; *B. hybridum* D and S subgenomes), whereas the second-best matches to other reference genes showed either low identities or relatively high identities but for short overlapping sequences (Data S3a,b). Therefore, the results from the two analyses supported the correspondence of our selected *B. stacei* TE4.3 and *B. hybridum* BdTR6g transcripts to their respective orthologous encoding genes, discarding the potential influence of paralogy in the performance of our PhyloSD algorithms and the correct assignation of the *B. hybridum* BdTR6g 'b' and 'd' alleles to their corresponding progenitor subgenomes.

A further validation of the goodness-of-fit of our PhyloSD pipeline was performed with the CDSs of 160 out of the 322 encoding genes filtered by the PhyloSD algorithms (Table S5a), retrieved in some cases from the reference genomes of other accessions of the same species. We used different accessions (than those used in the transcriptome study) of the allotetraploid *B. hybridum* (ABR113), its diploid progenitor species *B. stacei* (ABR114), and the diploid perennial *B. sylvaticum* (Ain-1), plus the orthologous CDSs from the reference genomes of the same accessions of the diploids *B. distachyon* (Bd21) and *B. arbuscula* (Barb1) and outgroups *O. sativa* and *H. vulgare* (see Experimental Procedures; Table S2). This PhyloSD analysis, restricted to only the allopolyploid *Brachypodium* species with known progenitor subgenomes, also fit a simple scenario where the *B. hybridum* ABR113 'b' and 'd' homeologs corresponded to its respective *B. stacei* ABR114-type (B) and *B. distachyon* Bd21-type (D) subgenomes (Figure S4a–c; Table S5a–c), supporting the likely accuracy and applicability of our subgenomic designation approach for any infraspecific genome of the same allopolyploid species and of their potential diploid progenitor subgenomes.

Independent validation of the *Subgenome Assignment* algorithm using coalescence methods

The accuracy of our results was validated independently using a coalescence analysis of the confirmed *Brachypodium* allopolyploid species. We used two strategies: (i) simulated coalescent trees; and (ii) simulated allopolyploids (Figures 1c and S1). With strategy (i) (Table S6a–c), we evaluated whether the *Subgenome Assignment* algorithm grafted known and 'ghost' homeologous subgenomes to the correct branches of the species tree under the hypothetical existence of ILS. Different hypothetical species trees (with variable effective population sizes, see Experimental Procedures) that contained all of the diploid genomes and one polyploid homeologous subgenome at a time were evaluated with the COAL program (Degnan and Salter 2005), which assays all possible gene tree distributions that can be constructed with a specific number of tips of a species tree. In all cases, the highest probability model corresponded to the species tree branch in which the homeologous subgenome was grafted by our *Subgenome Assignment* algorithm (Table S6a). There was a complete qualitative agreement between the most frequently observed versus the theoretical topologies for *B. hybridum* [B (b) + D (d)] and for *B. rupestre* and *B. phoenicoides* (Bpho422, Bpho6) [E2 + G (g)], which suggests that the theoretical distributions fit closely to the observed data. Similarly, the *B. retusum* [A2 (a) + E1 (e)] and *B. rupestre* [E2 (e) + G (g)] theoretical distribution scenarios had higher frequencies of gene trees in which the A2 subgenome grafted to branch 'a' than to 'c' and the G subgenome to branch 'g' than to 'h', respectively, as was scored for the observed data (Figures 3c and 4a; Table S6b).

Strategy (ii) enabled us to measure the ability of the *Subgenome Assignment* algorithm to select the correct allopolyploid subgenomes under different levels of ILS (Table S7a–c). We generated hypothetical subgenomic allopolyploids that matched the real *Brachypodium* allopolyploids [ancestral-ancestral (A + B), similar to *B. mexicanum* (see Discussion); ancestral-intermediate (A + E), similar to *B. retusum*; and intermediate-recent (E + G), similar to *B. rupestre* and *B. phoenicoides*] and the relative frequency of the theoretical gene tree distributions were calculated for each using COAL. In all cases, the *Subgenome Assignment* algorithm recovered the expected placements of the subgenomes despite the different topological graftings (ancient, intermediate or recent branches) in the diploid skeleton tree at different coalescent-unit levels (0.5 CU, 1 CU) of ILS (Table S7b). This suggests that our algorithm is able to place the subgenomes onto their correct branch independently of any deep or shallow

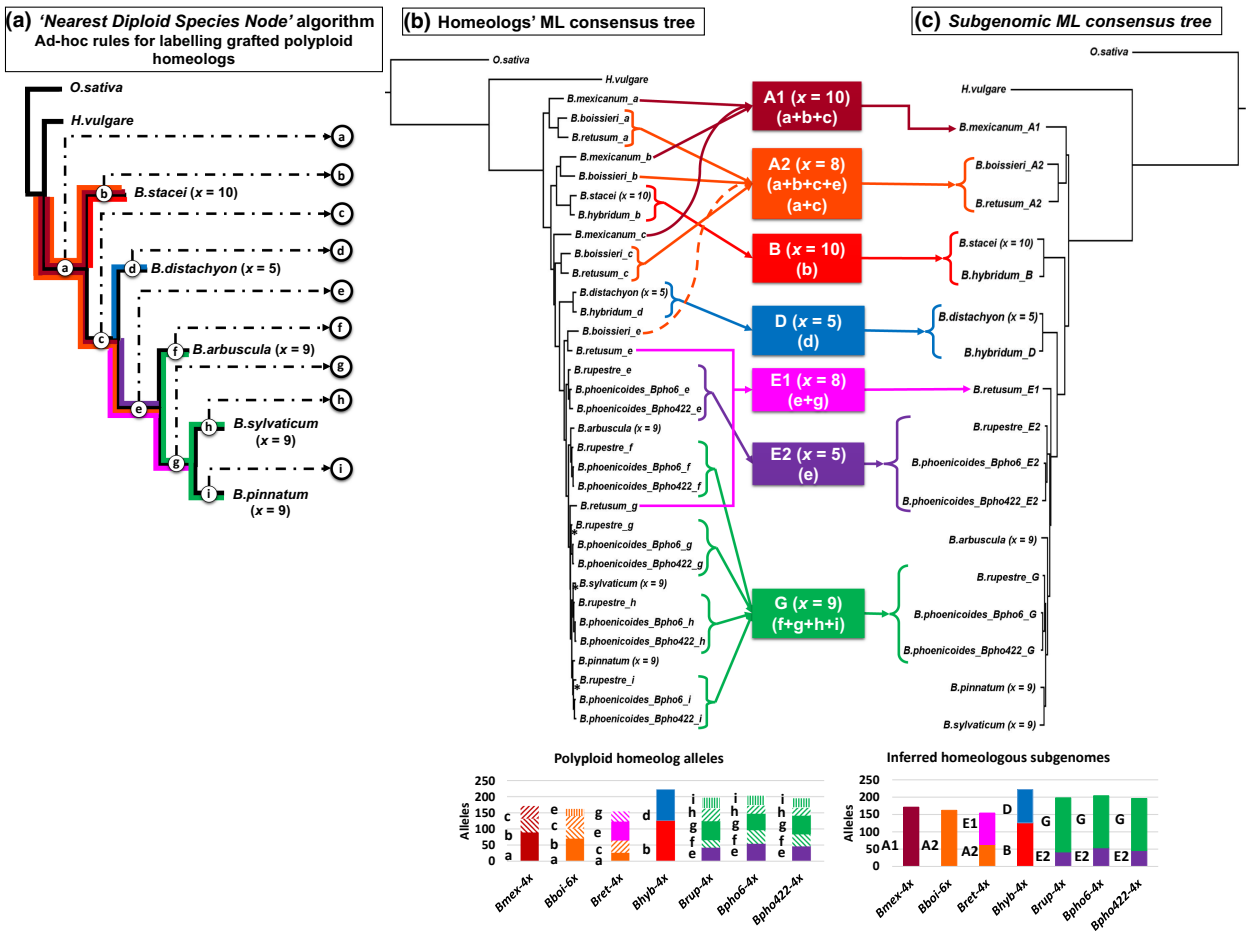

Figure 3. (a) Schematic *Brachypodium* tree illustrating the diploid skeleton tree (thick black branches) of the orthologous diploid genome sequences and their respective chromosome base numbers (*B. stacei* $x = 10$; *B. distachyon* $x = 5$; core perennial *B. arbuscula*, *B. sylvaticum* and *B. pinnatum* clade $x = 9$), and the nesting positions of the *Brachypodium* polyploid homeolog sequences showing the *ad hoc* labeling rules (lowercase letters, 'a'–'i') for the grafting positions of the *Brachypodium* polyploid homeolog sequences according to the *Nearest Diploid Species Node* algorithm; (b) *Brachypodium* homeologs' ML consensus tree based on 322 core transcripts and 1307 homeologs (Table 1b), with the polyploid homeolog sequences labeled according to the *Nearest Diploid Species Node* algorithm ('a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i'); (c) *Brachypodium* subgenomic ML consensus tree based on 322 core genes with the homeolog subgenomes labeled according to the *Subgenome Assignment* algorithm ('A1', 'A2', 'B', 'D', 'E1', 'E2', 'G'; Table 1b; Figure S3b). *Oryza sativa* and *Hordeum vulgare* were used as the outgroups. Asterisks indicate branches with SH-aLRT/UltraFast Bootstrap supports (BS) <80/95; the remaining branches have 100/100 values. The bar diagrams represent the frequencies of the homeologs in each polyploid and their assignments to their respective homeologous subgenomes (Table 1b). ML, maximum likelihood.

coalescences (branch lengths) or the effective population sizes of the *Brachypodium* lineages.

The *Brachypodium* dated nuclear tree and plastid tree

A strongly supported nuclear subgenomic ML consensus tree (Figure 3c), computed from the 322 validated core clusters (transcripts) with single and compound subgenome homeolog types, yielded the same *Brachypodium* gene topology as the dated Bayesian maximum clade credibility (MCC) BEAST tree (Figure 4a). The *Brachypodium* stem and crown nodes were estimated to have had Late-Eocene (36.3 Ma) and Mid-Miocene (12.1 Ma) ages, respectively (Figure 4a), which is consistent with the previous estimates that were based on a plastome analysis (Sancho et al. 2018). The basic chromosome numbers of the 'ghost' and merged subgenomes were inferred from their respective phylogenetic positions and the ploidy levels of the species that contained them [e.g. tetraploid *B. mexicanum* $2n = 40$: A1 ($x = 10$); hexaploid *B. boissieri* $2n = 48$: A2 ($x = 8$); allotetraploid *B. retusum* $2n = 32$: A2 ($x = 8$) and E1 ($x = 8$); allotetraploids *B. rupestre* and *B. phoenicoides* $2n = 28$: E2 ($x = 5$) and G ($x = 9$); Figures 3c and 4a; Table S2]. The ML tree shows the early divergence of the sister ancestral *B. mexicanum*_A1 ($x = 10$) and *B. boissieri*_A2/*B. retusum*_A2 ($x = 8$) subgenomic 'ghost' lineages, which was followed by the successive splits of *B. stacei* ($x = 10$) and its sister derived *B. hybridum*_B subgenomic lineage ($x = 10$) and of *B. distachyon* ($x = 5$) and its sister derived *B. hybridum*_D subgenomic lineage ($x = 5$; Figure 3c). The split of

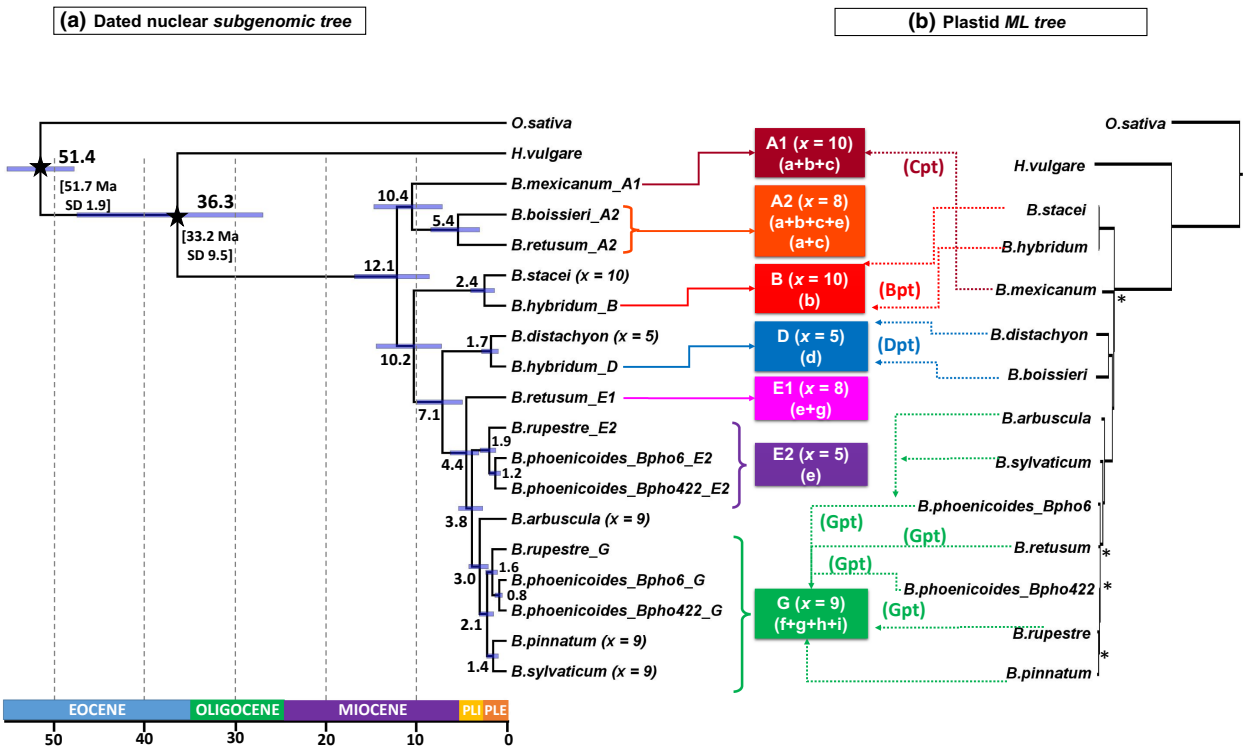

Figure 4. (a) *Brachypodium* Bayesian MCC dated chronogram of 322 independent core genes (with the polyploid homeologous subgenomes labeled according to subgenome-types 'A1', 'A2', 'B', 'D', 'E1', 'E2', 'G'; Table 1b), showing the estimated nodal divergence times (medians, in Ma) and the 95% highest posterior density (HPD) intervals (bars). Stars indicate secondary nodal calibration priors (means ± SD, in Mya) for the crown nodes of the BOP [*Oryza* + *Brachypodium* + *Hordeum*] and *Brachypodium* + core pooids [*Brachypodium* + *Hordeum*] clades. Accessions codes of *B. phoenicoides* correspond to those indicated in Table S2. (b) ML plastid consensus tree based on 31 plastome transcripts. *Oryza sativa* was used to root the trees. Asterisks indicate branches with SH-aLRT/UltraFast Bootstrap supports (BS) <80/95; the remaining branches have 100/100 values. MCC, maximum clade credibility; ML, maximum likelihood.

the ancestral (A1/A2) clade was inferred to have occurred in the Mid-Late-Miocene (10.4 Ma), a time close to that of the split of the oldest extant diploid *B. stacei*-type (genome B) clade (10.2 Ma; Figure 4a). The tree also revealed the successive divergences of the intermediate *B. retusum*_E1 ($x = 8$) and *B. rupestre*_E2/*B. phoenicoides*_E2 ($x = 5$) subgenomic 'ghost' lineages and the recently evolved [*B. arbuscula*, (*B. sylvaticum*/*B. pinnatum*)] core perennial clade species ($x = 9$), where the derived *B. rupestre*_G/*B. phoenicoides*_G subgenomic lineages ($x = 9$) were nested within (Figure 3c). The inferred dates indicate that the *B. retusum*_E1 subgenomic 'ghost' lineage is more ancestral (4.4 Ma, Early-Pliocene) than the *B. rupestre* and *B. phoenicoides*_E2 subgenome 'ghost' lineages (3.8 Ma; Figure 4a). Additionally, the estimated ages for the splits of the core perennial clade (3.0 Ma, Late-Pliocene), the diploid *B. pinnatum*/*B. sylvaticum* clade (2.1 Ma, Pleistocene) and the *B. rupestre*/*B. phoenicoides* G subgenomic lineages (2.1 Ma), and the origins of the *B. stacei*-type (2.4 Ma) and *B. distachyon*-type (1.7 Ma) homeologous lineages of *B. hybridum* (Figure 4a) are also in agreement with previous datings (Díaz-Pérez et al. 2018; Sancho et al. 2018; Gordon et al. 2020).

A *Brachypodium* plastid ML tree (Figure 4b) was constructed using 31 concatenated plastid transcripts (see Experimental Procedures). The plastid topology showed the successive moderate to well-supported divergences of *B. stacei* (and its sister *B. hybridum* with a *B. stacei*-type plastotype), *B. mexicanum*, *B. distachyon*/*B. boissieri* and *B. arbuscula* lineages, and the recent splits of the *B. sylvaticum*, *B. phoenicoides*_Bpho6, *B. retusum*, *B. phoenicoides*_Bpho422 and *B. pinnatum*/*B. rupestre* lineages (Figure 4b). The nuclear and plastid data recovered the same topology in their respective highly to moderately supported diploid skeleton trees (Figure S5a,b). Therefore, the diploid + polyploid plastid versus nuclear trees were compared with each other to infer the putative maternal genome donors of the polyploid accessions studied (Figure 4a,b). The plastid tree showed a full support for the maternal inheritance of a *B. distachyon* plastome-type (Dpt) by *B. boissieri*, a high support for that of a core perennial clade plastome-type (Gpt) by *B. retusum*, *B. phoenicoides* and *B. rupestre*, and a moderate support

for that of a C plastome-type (Cpt) by *B. mexicanum* (Figure 4b).

Karyotypic identification of the new *Brachypodium* genomes using CCB

The karyotypes of the previously unstudied *B. arbuscula* (2*n* = 18), *B. boissieri* (2*n* = 48) and *B. retusum* (2*n* = 32) species were analyzed using CCB mapping as described in Lusinska et al. (2019). The mapping was done with reference to the *B. distachyon* karyotype, and its genome was compared with the ancestral rice genome (IBI 2010). The arrangement of all of the BAC clones (Figure S6a–c) is shown on the cytogenetic maps of *B. arbuscula* (Figure S6a), *B. boissieri* (Figure S6b) and *B. retusum* (Figure S6c) chromosomes.

The cytomolecular mapping of the diploid *B. arbuscula* showed that each of the BAC clones hybridized to a single chromosome pair (Figures S6a, S7 and S8). The karyotypic pattern of this species was revealed to be the same as that of the genomes of the core perennial clade diploids *B. sylvaticum* and *B. pinnatum* with *x* = 9 (Figure 5; Lusinska et al. 2019). *Brachypodium arbuscula* chromosomes Ba1, Ba2, Ba4, Ba6, Ba7, Ba8 and Ba9, which correspond, respectively, to the ancestral *Oryza sativa* Os1, Os3, Os2, Os7, Os6, Os5 and Os4 chromosomes did not undergo any fusions, whereas one nested chromosome fusion (NCF) of Os8 and Os10 was observed on chromosome Ba3, and two NCFs of Os12 + Os9 + Os11 on chromosome Ba5 (Figures S6a, S7 and S8).

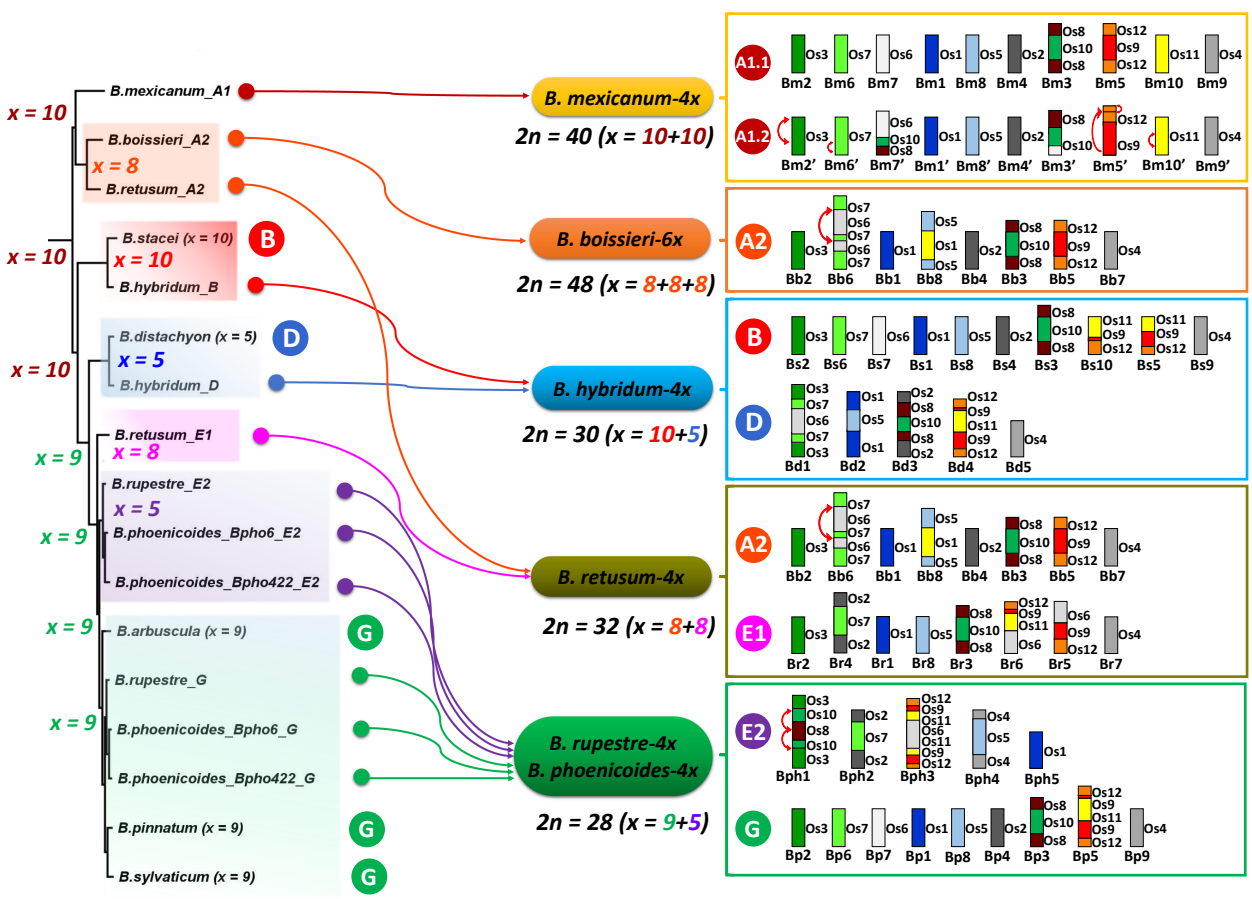

Figure 5. A comprehensive evolutionary framework for the origin of *Brachypodium* allopolyploids based on the combined phylogenomic and CCB analyses. Colors indicate the different types of (sub)genomes that were retrieved in the phylogenomic analysis, and letters designate the karyotype profiles that were found in the diploids and polyploids. The arrows link the inferred (sub)genomes and karyotypes of each studied *Brachypodium* polyploid. The karyotype models are based on the CCB analysis of the *B. arbuscula* (2*x*), *B. retusum* (4*x*) and *B. boissieri* (6*x*) species that were analyzed in this study (Figures S6–S17), and other *Brachypodium* representatives that had been previously studied (Lusinska et al. 2018, 2019; Gordon et al. 2020). Within the karyotypes, each chromosome or homeologous chromosome region corresponded to the relevant ancestral rice chromosome equivalents (Os1-Os12; Os – *Oryza sativa*; IBI 2010). The basic chromosome numbers (*x*) that were obtained for each genome and karyotype and inferred for the ancestors of the subgenomic tree are shown in the topology; their colors correspond to their respective (sub)genomic and karyotypic assignments. (Sub)genome designations: 'A1' – ancestral *B. mexicanum* (dark red); 'A2' – ancestral *B. boissiei* (orange); 'B' – *B. stacei* (red); 'D' – *B. distachyon* (blue); 'E1'– intermediate *B. retusum* (purple); 'E2' – intermediate *Brachypodium* core perennials (violet); 'G' – recent *Brachypodium* core perennials (green; Table 1b). Chromosome designations within the (sub)genomes: Bb – *B. boissiei*; Bd – *B. distachyon*; Bm, Bm′ – *B. mexicanum*; Bp – *Brachypodium* core perennials *x* = 9; Bph – *Brachypodium* core perennials *x* = 5; Br – *B. retusum*; Bs – *B. stacei*. CCB, comparative chromosome barcoding.

The CCB of the *B. boissieri* ($2n = 48$) chromosomes revealed that each BAC had six hybridization sites that were localized on three chromosome pairs (Figures S6b, S9–S13). The identical triplicated distribution pattern of the BAC-FISH signals in the morphologically uniform chromosomes supports the autohexaploid nature of *B. boissieri* with $x = 8$ (Figures 5 and S6b). The CCB analysis also detected the presence of chromosome fusions and rearrangements that were specific for the *B. boissieri* genome. *Brachypodium boissieri* chromosomes Bb1, Bb2, Bb4 and Bb7 correspond, respectively, to the Os1, Os3, Os2 and Os4 chromosomes, whereas Bb3 resulted from the NCF of Os8 and Os10, Bb5 from the NCF of Os12 and Os9, Bb6 from the NCF of Os7 and Os6 that were complemented with pericentromeric inversion, and Bb8 from the NCF of Os5 and Os11, which is a unique trait of the *B. boissieri* karyotype (Figures 5, S6b, S9–S13). In addition, the *B. boissieri* genome does not have Os12 + Os9 fused with Os11, a trait that is only shared with *B. mexicanum* in the genus *Brachypodium* (Figure 5; Lusinska et al. 2019).

The BAC-FISH analysis of *B. retusum* ($2n = 32$) demonstrated that each clone hybridized to four sites that were located on two chromosome pairs (Figures S6c, S14–S17). Unlike *B. boissieri*, this species had two distinct groups of chromosomes, each consisting of eight pairs of chromosomes, thereby revealing an allotetraploid nature for this *B. retusum* cytotype (Figures 5 and S6c). One of the chromosomal sets corresponded to a subgenome with the same karyotypic pattern as that of the *B. boissieri* genome and the other to a ghost subgenome with a *B. retusum*-type karyotype, both with $x = 8$ (Figures 5, S6c, S14–S17). The *B. retusum* chromosomes Br1, Br2, Br7 and Br8 corresponded, respectively, to Os1, Os3, Os4 and Os5 chromosomes, whereas Br3 resulted from the NCF of Os8 and Os10, and Br4 from the NCF of Os2 and Os7 (Figures 5 and S6c), which is a trait that was shared with the 'ghost' subgenome $x = 5$ present in the core perennial clade allotetraploids with $2n = 28$ (Figure 5; Lusinska et al. 2019). The *B. distachyon* Bd1- and Bd4-BAC-derived probes hybridized to two different *B. retusum* Br5 and Br6 chromosomes, both of which are specific to this subgenome in their syntenic segment composition. The distinctive arrangement of the BAC-FISH signals indicates that these two chromosomes originated via the reciprocal translocation of two ancestral *Brachypodium* chromosomes that correspond to Os12 + Os9 + Os11 and Os6 (Figures 5, S6c, S14–S17; Lusinska et al. 2019).

DISCUSSION

Deciphering the diploid origins of allopolyploids faces the challenge of accurately capturing their progenitor subgenomes (Doyle and Egan 2010; Levin 2013; Bombarely et al. 2014; Soltis et al. 2016). Approaches using coalescent-based analyses of multi-labeled trees and networks are hindered by homeolog loss and ILS (Marcussen et al. 2015; Thomas et al. 2017). The deconvolution of hybrid subgenomes is especially challenging in the absence of any known extant parents and in the absence of whole-genome sequence data for the studied species (Soltis et al. 2016; Liston et al. 2020). Recently, a phylogenetic subgenome-tree searching (PhyDS) pipeline was developed to retrieve the four progenitor genomes of allooctoploid *Fragaria x ananassa* from a wide transcriptome analysis of candidate species (Edger et al. 2019). This method explored the exclusive clades that contained the syntenic ortholog and homeolog sequences, and identified the progenitor subgenomes by grouping genes based on trees that meet bootstrap support cut-off values (Edger et al. 2019, 2020). There is current debate, however, on the accuracy of this approach, which although it correctly identified two extant progenitor diploid species (*F. vesca*, *F. iinumae*), it may have failed to identify the other two (Liston et al. 2020; Session and Rokhsar 2020; Feng et al. 2021).

Our PhyloSD pipeline refines previous methods and additionally enables the progenitor genomes of an unknown origin in the polyploids to be inferred. Our method was validated in the thoroughly studied *Triticum–Aegilops* polyploid complex in which the homeologous subgenomes of the *T. turgidum* 4*x* and *T. aestivum* 6*x* allopolyploids (Marcussen et al. 2014a) were accurately inferred, and in the allotetraploid *B. hybridum*, which was used as the internal control species with known progenitor genomes (Catalán et al. 2012; Gordon et al. 2020). In *Brachypodium,* our strategy enabled us to uncover three known (B, D and amalgamated G) and four unknown (A1, A2, E1, E2) diploid progenitor genomes of six polyploid *Brachypodium* species that had different dysploid ancestral origins (Figures 3 and S3b). Moreover, the inferences of the *Subgenome Assignment* algorithm were robust to the presence of the 'ghost' subgenomes and to the moderate existence of ILS in *Brachypodium* (strategy (i), Table S6a; strategy (ii), Table S7a).

One of the caveats of our approach is that only a small percentage of the pre-filtered gene clusters had a topology that was congruent with that of the diploid species tree (18% in *Triticum–Aegilops*; 17% in *Brachypodium*), and only those genes could be used to infer the homeologous subgenomes of the allopolyploids. Although other phylogenetic approaches such as PhyDS also use low percentages of the total number of expressed genes (< 17.9%) to identify the progenitor subgenomes of allopolyploids, in this case this was done by selecting the high-confidence syntenic homeologs that are present in each of the subgenomes of the polyploid (Edger et al. 2019, 2020). Recent approaches have proposed the inclusion of paralogs in order to increase the amount of data that can be used to infer a species tree (Smith and Hahn 2020). However, the

all-gene total evidence principle could lead to misleading phylogenies if the increasing amount of data also increases the phylogenetic noise. By contrast, the utility of our PhyloSD pipeline concurs with a restricted total evidence genomic scenario that favors the use of selected components of the data partitions that better fit the evolutionary models as the most reliable method for phylogenetic reconstruction (Goremykin et al. 2015) and, consequently, for homeolog subgenomic detection.

We examined alternative evolutionary scenarios that could explain the origins of the studied *Brachypodium* polyploids (Figure 6) considering the information obtained from the nuclear and plastid phylogenies and the cytogenetic barcoding data (Figures 4 and 5). The hypothetical nuclear genomes A1 of *B. mexicanum*, A2 of *B. boissieri* and *B. retusum*, E1 of *B. retusum*, and E2 and G of *B. phoenicoides* and *B. rupestre* are similar to those that were retrieved using cloned nuclear ribosomal genes (Catalán et al. 2016; Díaz-Pérez et al. 2018); however, here, they are supported by a larger set of 322 core-expressed genes (Figure 3c; Table 1b; Data S1). Our CCB karyotypes undisputedly identified the three known and four 'ghost' diploid progenitor genomes that are present in the six studied *Brachypodium* polyploids (Figure 5). The feasibility of our approach was facilitated by the high synteny that was observed across the *Brachypodium* reference genomes (Scholthof et al. 2018; Gordon et al. 2020), and by the high integrity of the progenitor genomes that were found in the subgenomes of some of the *Brachypodium* allopolyploids (Gordon et al. 2020). The *Brachypodium* nuclear genomes likely derived from a karyotype evolution model of successive centromeric chromosome fusions with a relatively low incidence of other types of rearrangements (Figure 5; Lusinska et al. 2019).

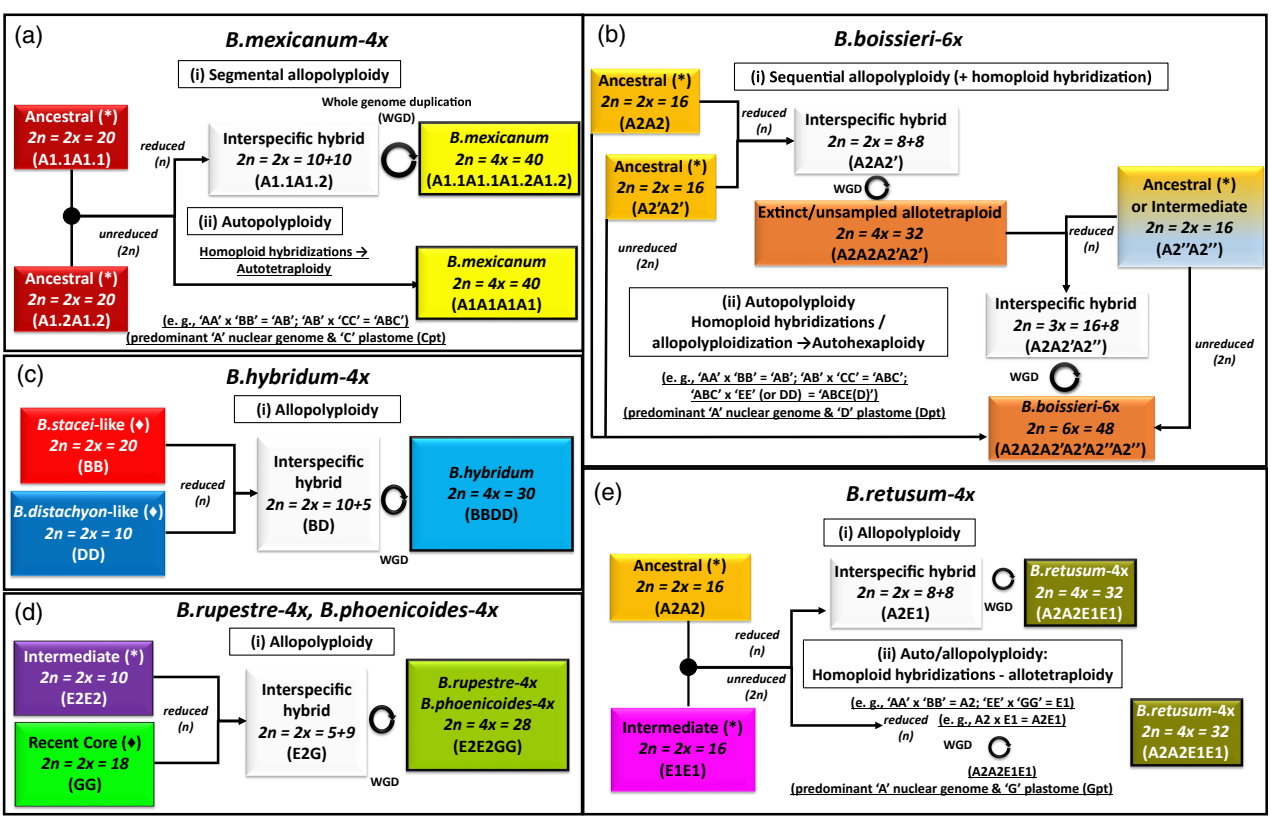

Figure 6. Diagrams representing the hypothetical evolutionary scenarios for the origins of the studied *Brachypodium* polyploids based on their nuclear homeolog and subgenomic trees (Figure 3b,c), the dated nuclear tree (Figure 4a), the plastid tree (Figure 4b), and the cytogenetic and karyotype evolution data (Figure 5; Table S2). (a) *B. mexicanum*-4x: (i) segmental allopolyploidy scenario [interspecific hybridization followed by whole-genome duplication (IH + WGD) of closely related ancestral genomes A1.1 and A1.2] *versus* (ii) autopolyploidy scenario [homoploid hybridizations (HH) of closely related ancestral genomes A, B, C followed by autotetraploidization]. (b) *B. boissieri*-6x: (i) sequential allopolyploidy scenario [two rounds of allopolyploidizations (tetraploidy and hexaploidy), with or without homoploid hybridizations, of closely related ancestral and ancestral-intermediate genomes A2, A2′, A2′′] *versus* (ii) autopolyploidy scenario [HH of closely related ancestral and ancestral-intermediate genomes A, B, C and E (or D) (with or without allopolyploidizations) followed by autohexaploidization]. (c) *B. hybridum*-4x: (i) allopolyploidy scenario [IH + WGD of ancestral and relatively ancestral B and D genomes]. (d) *B. rupestre*-4x and *B. phoenicoides*-4x: (i) allopolyploidy scenario [IH + WGD of intermediate E2 and recent G genomes]. (e) *B. retusum*: (i) allopolyploidy scenario [IH + WGD of ancestral A2 and intermediate E1 genomes] *versus* (ii) auto/allopolyploidy scenario [HH of closely related ancestral A and C and intermediate and recent E and G genomes (with autotetraploidizations), or HH of those pairs of genomes followed by allotetraploidization]. Asterisks indicate ghost genomes derived from progenitor diploids that are unknown in current diploid species; diamonds indicate progenitor genomes present in current diploid species.

Although our subgenome detection algorithms were initially designed to retrieve the homeologous subgenomes of known and putative allopolyploids, two of the studied *Brachypodium* polyploids revealed homeolog types that could be assigned to the compound (sub)genomes that pertain to autopolyploids (*B. mexicanum* A1A1, *B. boissieri* A2A2A2; Figures 3 and 4a). These findings were confirmed by our CCB data (Figures 5, S6b, S9–S13; Lusinska et al. 2019). By contrast, all of the remaining *Brachypodium* polyploid species were undisputedly identified as allopolyploids (Figures 3c, 4a, 5, S6c, S14–S17; Table 1b; Lusinska et al. 2019). The *Brachypodium* plastid tree, however, detected maternal plastome traits from more recently evolved ancestors than those inferred by the nuclear tree and CCB data in *B. mexicanum* (Cpt), *B. boissieri* (*B. distachyon*, Dpt) and *B. retusum* (core perennial Gpt), whereas the plastome inheritances observed in the remaining polyploids *B. hybridum* (*B. stacei*, Bpt), *B. rupestre* and *B. phoenicoides* (Gpt) were compatible with one of their respectively inferred nuclear progenitor genomes (Figure 4b). The large genome size (GS) of *B. mexicanum*, which is unusual among extant *Brachypodium* species (Table S2), and the uncertainty in the assignment of its close ancestral homeologs to a single nuclear genome (A1) or to two closely related genomes [e.g. A1.1 ('a' + 'c') and A1.2 ('b')] would also favor an alternative segmental allopolyploid scenario (Mason and Wendel 2020) for this species. In fact, this interpretation agrees with the similar karyotypic barcoding patterns that were observed in its two chromosome complements (Figure 5; Lusinska et al. 2019). The detection of three types of nuclear alleles ('a', 'b', 'c'; Figure 3b) and a Cpt plastotype (Figure 4b) in this species would support the alternative hypothetical scenarios for the origin of *B. mexicanum* (Figure 6a). Thus, assuming that there were three potential ancestral diploid progenitor species (AA, BB and CC) each with $x = 10$ and high nuclear chromosomal collinearity, we could hypothesize: (i) that a new species arose from homoploid hybridization between B and C (BC), which then hybridized with A [or between A and B and then hybridized with C] followed by allopolyploidization to generate *B. mexicanum* (segmental allopolyploidy scenario; Figure 6a); (ii) that homoploid hybridizations occurred between the three ancestral species (at diploid level) generating a new (ABC-type) species that finally underwent autopolyploidization to generate *B. mexicanum* (autopolyploidy scenario; Figure 6a); (iii) that the three ancestors underwent autopolyploidization (at tetraploid level) followed by homoploid hybridizations to generate *B. mexicanum*. In any of these potential scenarios, the *B. mexicanum* outcome retained a predominant A-type nuclear genome and a C-type plastome (Figures 4a,b and 6a). Ancestral homoploid hybridizations that predated subsequent polyploidization events have been common in some grass lineages (Marcussen et al. 2014a). The elucidation of the potential segmental allotetraploidy versus autotetraploidy of *B. mexicanum* and any of the proposed scenarios would require more precise genomic and cytomolecular data.

Brachypodium boissieri was revealed to be autohexaploid by cytogenetic analysis (Figures 5, S6b, S9–S13). The assignment of its close 'a', 'b' and 'c' plus the residual 'e' homeologs to a unique A2 $x = 8$ genome (Figure 3; Table 1b) was corroborated by its unambiguous karyotype (Figures 5, S6b, S9–S13). However, more complex hypothetical evolutionary scenarios should be considered (Figure 6b) to explain the origin of this species based on its nuclear alleles (Figure 3b) and its Dpt plastotype (Figure 4b). Hypothesized evolutionary scenarios could involve: (i) two rounds of allopolyploidization (tetraploidy and hexaploidy) between the potential ancestral diploid progenitor species (AA, BB, CC) with putative $x = 8$ and a more recent DD-(or EE)-type species could have been accompanied by homoploid hybridization (sequential allopolyploidy scenario; Figure 6b); (ii) different homoploid hybridizations (at either diploid or polyploid levels) of diploid ancestors with putative $x = 8$, coupled with allopolyploidizations, could have rendered a new ABCD (orE)-type species that finally underwent autopolyploidization (autopolyploidy scenario; Figure 6b). In any of these scenarios, *B. boissieri* retained a predominant A-type nuclear genome and a D-type plastome (Figures 4a,b and 6b). Although this *B. boissieri* genome shares unique karyotypic features with the reduced D ($x = 5$) genome of *B. distachyon*, such as the nested fusion of Os6 into the centromere of Os7, the evolutionary time frame and events that resulted in its A2 ($x = 8$) genome are unclear (Figure 5). This Os6-Os7 NCF could have occurred in parallel from the common ancestor with $x = 10$ to genome A2 with $x = 8$ and genome D with $x = 5$ (Figures 5 and 6), or an intermediate ancestor with the Os6-Os7 NCF separately gave rise to A2 and D through different chromosomal rearrangements. To date, this species constitutes the only putative autopolyploid within *Brachypodium*.

Even though the origins of the confirmed *B. hybridum*, *B. phoenicoides*, *B. rupestre* and *B. retusum* allotetraploids could be more easily conjectured, alternative scenarios could be also proposed for them (Figure 6c,d) based on the compared nuclear and plastid topologies and the CBB data (Figures 3b, 4b and 5). Homoploid hybridization of diploid progenitor ancestors followed by genome doubling, rather than autotetraploidization of the ancestors followed by tetraploid homoploid hybridization, represents the most straightforward hypothetical scenario for their respective BBDD (*B. hybridum*), A2A2E1E1(*B. retusum*) and E2E2GG (*B. phoenicoides*, *B. rupestre*) origins (Figure 6c, d). It has been confirmed in *B. hybridum* for which synthetic allotetraploids were obtained after selected crosses of *B. stacei*-2*x* and *B. distachyon*-2*x* parents followed by

colchicine-induced genome duplication, whereas all attempted crosses between the synthetic parental autotetraploids failed to produce descendants (Dinh Thi et al. 2016). Protoancestral homoploid BD hybrids ($2n = 15$) of the annual *B. hybridum* have not been found in the wild, presumably because they would be sterile and would disappear at the end of the growing season unless they experience spontaneous whole-genome duplication to become allotetraploid. The studied *B. hybridum* accession shows a BD-type nuclear genome and a maternal *B. stacei*-type (S) plastome (Figures 4a,b, 5 and 6c), though individuals with maternal *B. distachyon*-type (D) plastomes have also been found in nature (López-Álvarez et al. 2012; Gordon et al. 2020; Shiposha et al. 2020), corroborating the recurrence and bi-directionality of this allotetraploid speciation event in this species. A similar homoploid hybridization followed by genome doubling scenario is hypothesized to explain the origins of the perennial *B. rupestre* and *B. phoenicoides* (Bpho6, Bpho422) allotetraploids, with all the three accessions showing a E2G-type nuclear genome and a G-type plastome (Figures 3b, 4a,b, 5 and 6d). However, the deconvolution of the origin of *B. retusum* requires more complex hypothetical scenarios (Figure 6e) that may involve: (i) interspecific hybridization of ancestral A2A2 and intermediate E1E1 species with $x = 8$ followed by genome doubling (direct allopolyploidy scenario); (ii) different homoploid hybridizations of AA × CC and EE × GG diploid genomes that, respectively, originated the A2 (AC) and E1 (EG) subgenomes, followed by allotetraploidization, with the *B. retusum* outcome retaining a predominant A-type nuclear genome and a G-type plastome (homoploid hybridization + allopolyploidy scenario; Figures 3b, 4a,b and 6e). Deeper genomic and cytomolecular analyses should be carried out to confirm the origins of this perennial *Brachypodium* allotetraploid. The successive divergences of the *Brachypodium* polyploid subgenomes and their karyotype structures support an evolutionarily descendant dysploidy trend from the ancestral $x = 10$ (A1, B) to the recent $x = 9$ (G) genomes (Figures 4 and 5), which corroborates the findings of Lusinska et al. (2019) that inferred the existence of an Ancestral *Brachypodium* Karyotype (ABK) of $x = 10$. Our newly emerged karyotype evolutionary scenario of *Brachypodium* also involves two independent reductions from $x = 10$ to $x = 8$ (ancestral A2) and from $x = 9$ to $x = 8$ (intermediate E1) plus two independent reductions from $x = 9$ to $x = 5$ (intermediate D and E2; Figures 4 and 5).

The nearly contemporary Mid-Late Miocene inferred origins of the divergent A1 and B $x = 10$ genomes (Figure 4) resulted in highly syntenic karyotypes that only had rearrangements within some of the homeologous chromosomes (e.g. chromosomes Bm5 and Bm10 of subgenome A1.1 and intrachromosomal rearrangements in Bm5′ and Bm10′ of subgenome A1.2 of *B. mexicanum versus* chromosomes Bs5 and Bs10 of *B. stacei* that probably originated via a reciprocal translocation or chromosome split; Figure 5). By contrast, the parallel but separate reductions to $x = 8$ and $x = 5$ genomes imply major structural changes that primarily affected the number and compositions of the chromosomal fusions. Thus, the two hypothesized NCFs that resulted in the ancestral Late-Miocene A2 $x = 8$ genome (chromosomes Bb6 and Bb8) differed from the more complex pattern of the three NCFs plus one translocation that resulted in the Early-Pliocene E1 $x = 8$ genome (chromosomes Br4, Br5 and Br6; Figures 4 and 5). Similarly, the increasing reduction that was caused by the four NCFs from the hypothetical $x = 9$ Intermediate ABK (Figure 5; Lusinska et al. 2019) that ended in the Late-Miocene D $x = 5$ genome (chromosomes Bd1 to Bd4) was distinct from the four NCFs that resulted in the Late-Pliocene E2 $x = 5$ genome (chromosomes Bph1 to Bph4; Figures 4 and 5).

Despite the large hypothesized rearrangements that were experienced by the *Brachypodium* genomes (Figure 5), their chromosomes are highly collinear as is demonstrated by the high synteny that was observed between the reference genomes of the ancestral *B. stacei* $x = 10$ (B) genome and the intermediate and highly reduced *B. distachyon* $x = 5$ (D) genome (Gordon et al. 2020). Interestingly, the diploid progenitor genomes have experienced divergence and diversification in different *B. hybridum* lines [ancient (1.4 Ma) and recent (0.4 Ma) allotetraploids, spanning 1 Ma; Gordon et al. 2020], although those genomes have remained almost intact in the derived allopolyploid subgenomes as shown in the inherited karyotypes and collinear sequences of the allotetraploid *B. hybridum* reference subgenomes and those of its diploid progenitors' reference genomes (Figure 5; Gordon et al. 2020). Similar genomically diversified but syntenically conserved progenitor subgenomes have been observed in other allopolyploid plants such as tetraploid cottons (Chen et al. 2020) and octoploid strawberry (Hardigan et al. 2020). These findings, together with the inferred ages and karyotype patterns of the studied *Brachypodium* species (Figures 4 and 5), support a highly dynamic evolutionary scenario of chromosomal reshuffling that led to diploid species that have highly syntenic but rearranged genomes that were inherited by their descendant polyploids during the last 12 million years. The genomic evidence supports our assumption that the identified ghost genomes of the *Brachypodium* polyploids (A1, A2, E1, E2; Figures 3–6) are the preserved vestiges of the diploid progenitor genomes they had originated from.

While our current analyses suggest that the ancestral A1 and A2 and the intermediate E1 and E2 genomes likely correspond to extinct or unsampled diploid *Brachypodium* species, identifying the G genome is more problematic. The phylogenomic data could not accurately assign the

very close 'f', 'g', 'h' and 'i' homeologs to any of the core perennial clade diploid lineages (Figure 3; Table 1b), and the CCB data could not detect differences in the karyotypic patterns of *B. arbuscula* (Figures 5 and S6a), *B. sylvaticum* and *B. pinnatum* and the $x = 9$ subgenomes of the allotetraploids *B. phoenicoides* and *B. rupestre* (Figure 5; Lusinska et al. 2019). The characterization of the Plio-Pleistocene-originating core perennial *Brachypodium* diploids would require the use of a large number of highly variable genomic loci and chromosomal barcodes. Still, the accurate identification of the four 'ghost' genomes by our combined PhyloSD and CCB methods makes *Brachypodium* a unique case within the angiosperms. This model genus constitutes an excellent study system to investigate the impact of the 'ghost' subgenomes on the functional, adaptive and evolutionary behavior of their hosting polyploids. Future studies focusing on the differential expression of genes under abiotic or biotic stresses, responses of individuals and populations to environmental cues or diverse ecological niche inheritances caused by the ghost subgenomes of the polyploids will open new avenues of advanced research in *Brachypodium* and other polyploid plants.

Conclusions

Our study has contributed to unraveling the origins of six *Brachypodium* polyploids and their hypothetical homoploid hybridization and allo- and autopolyploidization scenarios. Our results demonstrate the value of the PhyloSD pipeline coupled with the CCB approach in detecting polyploid subgenomes. The wheat benchmark indicated that our method can identify the diploid homeologous subgenomes from extant progenitors. More importantly, our analysis also identified three known and four novel 'ghost' subgenomes in *Brachypodium*, thus shedding light on the complex and intricate evolutionary history of this grass model genus. Our method could be of significant value in studies of polyploid plants that have complex histories of hybridizations and polyploidizations.

EXPERIMENTAL PROCEDURES

Genomic data of *Triticum–Aegilops*

Genomic sequence data of *Triticum* and *Aegilops* species with known chromosome numbers and ploidy levels were retrieved from Marcussen et al. (2014a) and (2014b). These data included 275 orthogroups of diploids *T. urartu*, *T. monococcum*, *Ae. speltoides*, *Ae. tauschii* and *Ae. sharonesis*, and allohexaploid *T. aestivum*. Additionally, cDNA sequences of the allotetraploid *T. turgidum* (*T. turgidum* subsp. *durum*; Svevo.v1; release 47 of https://plants.ensembl.org/Triticum_turgidum/Info/Index; Maccaferri et al. 2019; Howe et al. 2020) and genome data of the close outgroups *O. sativa* (Osativa_323_v7.0; http://phytozome.jgi.doe.gov/; Ouyang et al. 2007), *B. distachyon* (Bdistachyon_314_v3.1; http://phytozome.jgi.doe.gov/; IBI 2010) and *H. vulgare* (ftp://ftpmips.helmholtz-muenchen.de/plants/barley/genome_release2017/; Mascher et al. 2017) were added to this dataset.

Brachypodium sampling, chromosome counting and GS determination

Eleven *Brachypodium* species and two ecotypes [all known diploids *B. arbuscula*, *B. distachyon*, *B. pinnatum*, *B. stacei* and *B. sylvaticum*, and polyploids *B. boissieri, B. hybridum, B. mexicanum, B. phoenicoides* (Bpho6 and Bpho422 accessions), *B. retusum* and *B. rupestre*] were studied (Table S2; Methods S1). The GS and chromosome counting estimations were performed using flow cytometry and on DAPI-stained meristematic root cells following the protocols of Doležel et al. (2007) and Jenkins and Hasterok (2007), respectively. The ploidy levels were inferred from the chromosome counts ($2n$) and the GS (pg/2C) estimations that were performed in the same accessions that were used in the transcriptome study, and through the GS and $2n$ values that were obtained in conspecific accessions that showed similar values (Table S2; Methods S1).

Transcriptomic data of *Brachypodium*

Plants for transcriptome analysis were grown in the greenhouses of the Arnold Arboretum of Harvard University. Total RNA was isolated from the leaf tissue of each individual plant under one of the following conditions: control, soil-drying stress, heat stress and salt stress; pooled RNAs were used for the sequencing. RNA was isolated using Rneasy extraction kits (Qiagen, Hilden, Germany), quantified using a NanoDrop spectrophotometer (ThermoFisher, Waltham, MA, USA), and checked for integrity using a BioAnalyzer (Agilent, Santa Clara, CA, USA). RNASeq libraries were prepared at the Whitehead Institute Genome Technology Core and sequenced at the Bauer Core of Harvard University on Illumina HiSeq 2500 sequencers. Transcript sequences were assembled using trinityrnaseq-r20140717 (Grabherr et al. 2011). A *de novo* assembly of the *Brachypodium* RNA-seq reads (Table S3a) produced 72–160 thousand transcript isoforms with median lengths ranging between 414 and 555 bp (Table S3b). The *Brachypodium* RNA-seq data were deposited in the ENA (European Nucleotide Archive; https://www.ebi.ac.uk/ena; Methods S1). The RNA-seq data of *B. distachyon* (Bd21) and *B. sylvaticum* (Brasy-Esp) were obtained from Bettgenhaeuser et al. (2017) and Fox et al. (2013), respectively, and data of the outgroups *Oryza sativa* (SRX738077) and *Hordeum vulgare* (ERR159679) were obtained from the INSDC archives.

Plastid data of *Brachypodium*

Plastid reads of the studied *Brachypodium* samples were filtered from the pool of RNA-seq data with DUK (http://duk.sourceforge.net; Li et al. 2011) using a reference set of 23 grass plastomes and a matching K-mer composition of $K = 24$. *De novo* assembling and clustering of *B. pinnatum*, *B. rupestre, B. phoenicoides* (Bpho6 and Bpho422 ecotypes), *B. mexicanum*, *B. boissieri* and *B. retusum* transcripts plus CDS sequences extracted from plastomes of *B. distachyon* Bd21 (NC_011032; Bortiri et al. 2008), *B. stacei* ABR114 (NC_036837), *B. hybridum* ABR113 (NC_036836), *B. sylvaticum* (Sin1, Phytozome) and *B. arbuscula* (Barb1, Phytozome) were performed with NOVOPlasty (Dierckxsens et al. 2017) and the pipeline described in Sancho et al. (2018) rendering an aligned data matrix. A total of 31 plastome core transcripts (atpA, atpF, ccsA, cemA, clpP, matk, ndhB, ndhJ, ndhK, petA, petB, petD, psaA, psaB, psaI, psbA, psbB, psbC, psbE, psbH, psbI, psbK, psbM, psbN, rbcL, rpl22, rpoA, rpoB, rps16, rps4 and rps7 3′end

partial sequences) were recovered from this dataset, aligned and concatenated for phylogenomic analyses.

Orthology assessments of nuclear *B. stacei-2x* and *B. hybridum-4x* transcripts used in the phylogenomic analyses of *Brachypodium* polyploids

For the reference-transcriptome assessment, Blastn (blast-2.6.0+; Altschul et al. 1990; Camacho et al. 2009) searches were conducted to compare and match the respective 322 and 222 assembled transcripts of *B. stacei* (accession TE4.3) and *B. hybridum* (accession BdTR6g) to the available reference primary transcripts of *B. stacei* (accession ABR114; Phytozome v13: Bstacei_316_v1.1.-transcript_primaryTranscriptOnly) and *B. hybridum* (accession ABR113; Phytozome v13: Bhybridum_463_v1.1.transcript_primaryTranscriptOnly; Goodstein et al. 2012). For the reference-genome assessment, raw Illumina reads from the resequenced *B. stacei* TE4.3 (NCBI SRR1802178) and *B. hybridum* BdTR6g (Phytozome) accessions were mapped to their respective *B. stacei* (ABR114; https://phytozome-next.jgi.doe.gov/info/Bstacei_v1_1) and *B. hybridum* (ABR113; https://phytozome-next.jgi.doe.gov/info/Bhybridum_v1_1) reference genomes with BWA v. 0.7.12-r1039 (Li and Durbin 2009). Samtools and bcftools (Li et al. 2009; Li 2011) were used to assemble the consensus gene sequences (Methods S1). The genes encoding their respective 322 and 222 transcripts were filtered from each genic dataset, and their identities compared with their respective reference genes (orthologs) using the Blastn approach. Best Blastn matches with percentages of identity > 95% to the primary reference transcript or to its encoding reference gene were considered a strong evidence of gene orthology, whereas other matches with percentages of identity < 95% or showing high identity percent but short sequence overlap were considered as potential paralogs. Matches with high percentages of identity but corresponding to genes from the other progenitor subgenome (*B. hybridum*) were treated as homeologs.

CDS data from *Brachypodium* reference genomes

Full CDSs of the 322 *Brachypodium* transcripts used in the subgenome assignation analysis were retrieved from the available reference genomes of diploid *B. distachyon* Bd21 (Phytozome v13: Bdistachyon_314_v3.1), *B. stacei* ABR114 (Phytozome v13: Bstacei_316_v1.1), *B. arbuscula* (Phytozome: BarbusculaBARB1v3.1), *B. sylvaticum* (Phytozome 13: Bsylvaticum_490_v1.1), *H. vulgare* (Hvulgare_IBSC_PGSB; Mascher et al. 2017) and *O. sativa* (Phytozome 13: Osativa_323_v7.0) species, and of the allotetraploid *B. hybridum* ABR113 (Phytozome v13: Bhybridum_463_v1.1), and clustered using GET_HOMOLOGUES-EST v09112017 (Contreras-Moreira et al. 2017). The default parameters of the program were relaxed to an identity percent value of 75% and a coverage percent value of 10% due to the high variability of sequence lengths among the samples, recovering 193 CDS clusters. These were employed to assess the accuracy of the PhyloSD pipeline while assigning *B. hybridum* ABR113 homeologs (Figure S4), which ultimately produced 160 filtered CDSs.

Phylogenomic and dating analyses

The *Brachypodium* nuclear and plastid datasets and *Triticum–Aegilops* datasets were aligned using GET_HOMOLOGUES-EST v09112017 (Contreras-Moreira et al. 2017) and MAFFT v7.222 (Katoh et al. 2002; Katoh and Standley 2013), respectively. Multiple alignments were trimmed with TrimAl 1.4.1 (Capella-Gutiérrez et al. 2009). The ML analyses were performed using IQ-TREE v.1.6.1 imposing the optimal substitution model selected by ModelFinder for each partition or dataset in terms of the AICc (Minh et al. 2013; Nguyen et al. 2014; Chernomor et al. 2016; Kalyaanamoorthy et al. 2017). The ultrafast bootstrap searches were replicated 1000 times (Methods S1). When checking the diploid species topology, trees were pruned with the Newick Utils (Junier and Zdobnov 2010). The Bayesian phylogenetic dating analysis of the *Brachypodium* nuclear dataset was conducted using BEAST 2.4.7 (Bouckaert et al. 2014) imposing the GTR substitution model, lognormal relaxed clock (clock rate = 1×10^{-4}) and Birth–Death tree models and secondary calibrations (Methods S1). The distance-based coalescence analyses of the *Brachypodium* and *Triticum–Aegilops* nuclear diploid MSAs that reconstructed their respective diploid species tree were performed using ASTRAL v5.7.3 (Zhang et al. 2018), and STAR and STEAC (R v.3.5.1; Liu and Yu 2010; Methods S1).

Performance of the *Subgenome Assignment* algorithm in the presence of ILS in *Brachypodium*

The COAL program (Degnan and Salter 2005) was used to compute the theoretical probabilities of the gene tree topologies from fixed species trees under a multispecies ILS scenario using two strategies (Tables S6 and S7). The *Subgenome Assignment* algorithm was applied to each set of probabilities that had been computed from COAL for a single species tree, and the selected subgenomes were matched to the expected subgenomes in order to validate the algorithm. In strategy (i), the homeologous subgenomes of the *Brachypodium* allopolyploids were coded as for the observed data, and the divergence time for each polyploid lineage was inferred from the closer ancestral node that included it as a sister lineage to its diploid species (Figures 3c and 4a; Tables S6a–c). This species tree was used to compute the theoretical distribution of all of the gene tree topologies that had the optimal diploid skeleton topology and that contained the polyploid subgenome. The branch lengths were transformed to CU, where CU = g/2Ne, and assuming g = 1.5 years per generation and optimal effective population sizes of Ne = 5E5, Ne = 1E6 and Ne = 2E6. In strategy (ii), the theoretical distributions from two homeologous subgenomes were proportionally merged in order to recreate the genomic compositions of the *Brachypodium* allotetraploids, following the same criteria as for the observed homeolog types (Figure 3b,c). An effective population size of Ne = 5E5 individuals and two branch lengths were tested for all of the lineages of a fixed tree using different CUs [deep coalescence (1 CU), equivalent to 1.5 My; and shallow coalescence (0.5 CU), equivalent to 0.75 My (Tables S7a–c; Methods S1)].

Comparative chromosome barcoding

Three perennial *Brachypodium* species, *B. arbuscula* Barb502 ($2n = 2x = 18$), *B. boissieri* Bbois10 ($2n = 6x = 48$) and *B. retusum* Bret504 ($2n = 4x = 32$), were analyzed in this study along with the reference *B. distachyon* Bd21 (Table S2). Multisubstrate chromosome preparations (reference *B. distachyon* plus one other *Brachypodium* species at a time) were prepared from root-tip meristems as described in Hasterok et al. (2006). The 43 BAC clones (Table S8) that were used in this study were previously employed in the construction of the karyotypes of other *Brachypodium* genomes (Figure 5; Lusinska et al. 2019). These probes came from the BD_ABa and BD_CBa genomic DNA libraries generated from the five assemblies of FingerPrinted Contigs that had been assigned to the respective reference chromosomes of *B. distachyon* (Febrer et al. 2010). In order to determine any potential intraspecific variation, each clone was mapped to the chromosome preparations of at least three individuals of each species or

accession. The probe labeling with nick translation using tetramethylrhodamine-5-dUTP, digoxigenin-11-dUTP or biotin-16-dUTP (Sigma-Aldrich) and FISH were performed according to Jenkins and Hasterok (2007) with minor modifications (Lusinska et al. 2018). The images were acquired using a wide-field epifluorescence microscope (AxioImager.Z.2, Zeiss, Oberkochen, Germany) and a high-sensitivity monochromatic camera (AxioCam Mrm, Zeiss), and then uniformly processed using ZEN 2.3 Pro (Zeiss) and Photoshop CS3 (Adobe, San José, CA, USA).

ACKNOWLEDGEMENTS

The authors thank two anonymous reviewers for their valuable comments on an early version of the manuscript, Salvador Talavera for fruitful discussion on allopolyploidizations of wild annual plants, and the following institutions for providing the facilities to develop this study. The *Brachypodium* plants that were used in the study were collected in the field and cultivated at the High Polytechnic School of Huesca (University of Zaragoza), and vouchered specimens were deposited in the JACA (Pyrenean Institute of Ecology-CSIC) and UZ (University of Zaragoza) herbaria. The propagated clones were imported to and grown at The Arnold Arboretum of Harvard University under USDA APHIS PPQ PCIP-16-0433. The RNA-seq laboratory analyses were conducted at the Universities of Zaragoza and Harvard, the cytomolecular analyses at the University of Silesia in Katowice, and the bioinformatic and phylogenomic analyses at the EEAD-CSIC and University of Zaragoza. Access to the transcriptome data of *B. sylvaticum* Sin1, resequenced genome data of *B. hybridum* BdTR6g and reference genome data of *B. arbuscula* Barb1 was done with permission under the DOE Contracts no. DE-AC02-05CH11231 and FP00006675. This work was supported by the Spanish Ministries of Economy and Competitivity (Mineco) and Science and Innovation (MICINN) CGL2016-79790-P and PID2019-108195GB-I00 and University of Zaragoza UZ2016_TEC02 grant projects and funding from Harvard University. The work conducted by the US DOE Joint Genome Institute was supported by the Office of Science of the US Department of Energy (DOE) under Contracts no. DE-AC02-05CH11231 and FP00006675. RS was funded by a Mineco FPI PhD fellowship, Mineco and Ibercaja-CAI Mobility Grants and Instituto de Estudios Altoaragoneses Grant. BCM was funded by Fundacion ARAID. PC and RS were also funded by a European Social Fund/Aragon Government Bioflora Research Grants A01-17R and A01-20R. The work that was conducted at the University of Silesia in Katowice was supported by the Research Excellence Initiative program.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interests/competing interests.

AUTHOR CONTRIBUTIONS

PC, RS, BCM, AD, DD and RH designed the study. RS, LI, DD and JL developed the experimental work. RS, BCM, AD, LI, JL, RH and PC analyzed the data and interpreted the results. RS, AD, BCM and PC prepared the manuscript. All of the authors revised the manuscript.

DATA AVAILABILITY STATEMENT

The complete PhyloSD protocol (source code, step-by-step instructions, commands and examples) is available in Github (https://github.com/eead-csic-compbio/allopolyploids). The supporting information (Tables, Figures, Appendix, and Methods) and data used in this study, detailed pipelines and algorithms, alignments, BEAST xml file and phylograms of the Brachypodium and Triticum–Aegilops groups are available in Dryad (https://doi.org/10.5061/dryad.ncjsxksqw).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Development of the PhyloSD pipeline. Step-by-step application of the algorithms to the benchmarked *Triticum–Aegilops* dataset and the *Brachypodium* case study. Each step is numbered and referenced to the bioinformatic workflow shown in Figure S1.

Methods S1. Sampling, genome size, chromosome counting and ploidy level estimations of *Brachypodium*.

Data S1. List of the 322 nuclear validated core genes used in the phylogenomic analyses of the studied *Brachypodium* and outgroup samples. Information on *Brachypodium* gene identity (ID; *Brachypodium* database), gene annotation results from SwissProt and Ensemble Genomes databases, and gene description from SwissProt, are indicated for each gene.

Data S2. Blastn matching analysis of the *B. stacei* TE4.3 and *B. hybridum* BdTR6g nuclear validated core transcript used in this study to the primary transcripts of their respective reference transcriptomes. (a) 322 *B. stacei* TE4.3 core transcripts were used as query to match the corresponding primary transcripts of the *B. stacei* ABR114 reference transcriptome. (b) 222 *B. hybridum* BdTR6g core transcripts were used as query to match the corresponding primary transcripts of the *B. hybridum* ABR113 reference transcriptome. Values highlighted in bold indicate the best match for each query. In *B. hybridum*, values with red and blue background colors correspond to *B. stacei* and *B. distachyon* subgenomic matches, respectively.

Data S3. Blastn between the *B. stacei* TE4.3 and *B. hybridum* BdTR6g assembled genes that encode the nuclear validated core transcript used in this study and the annotated genes of their respective reference genomes. (a) 322 *B. stacei* TE4.3 genes were used as query to match the corresponding annotated genes of the *B. stacei* ABR114 reference genome (only 321 genes could be searched as one gene was not annotated in the *B. stacei* ABR114 genome). (b) 222 *B. hybridum* BdTR6g genes were used as query to match the corresponding annotated genes of the *B. hybridum* ABR113 reference genome. Values highlighted in bold indicate the best match for each query. In *B. hybridum*, values with red and blue background colors correspond to *B. stacei* and *B. distachyon* subgenomic matches, respectively.

Figure S1. A detailed workflow of our phylogenomic subgenome detection (PhyloSD) pipeline highlighting the scripts used in our '*Nearest Diploid Species Node*', '*Bootstrapping Refinement*' and '*Subgenome Assignment*' algorithms. Bioinformatics tools and analyses are shown in turquoise color, the outputs of each step in green color, and the two main final outputs of the first/second and third algorithms in violet and red, respectively. Parallel coalescence-based validations for the selection of the optimal diploid skeleton tree and the tested efficiency of the '*Subgenome Assignment*' algorithm are indicated in blue letters.

Figure S2. (a) The optimal *Triticum–Aegilops* diploid skeleton trees obtained through distance-based coalescence analyses with the ASTRAL, STAR and STEAC programs using 259 genes. (b) Bidimensional principal coordinate analysis (PCoA) plot constructed from pairwise patristic distances between diploid ortholog and

polyploid homeolog tips of the *Triticum–Aegilops* homeologs' ML tree computed with NTSYS-pc v2.10j. Each homeolog was labeled according to the branch of the skeleton diploid tree where it was grafted to (Figure 2a,b) using the '*Nearest Diploid Species Node*' algorithm. A minimum spanning tree was superimposed on the PCoA. The down-frequency-rank of grafting distributions of each polyploid homeolog-type across 100 bootstrap gene trees (Table S1a) is represented as a colored cluster in solid (1st rank), dashed (2nd rank) and dotted (3rd and 4th ranks) lines. Only the most frequent homeolog-types are shown according to ploidy level. Black lower-case letters in squares indicate groups of homeolog-types. Capital letters indicate the inferred homeolog subgenomes using the *Subgenome Assignment* algorithm (Table 1a, Inferred Subgenomes; Table S1b). Clusters represent bootstrap distributions of homeologs with threshold cut-off values over 10% (Table S1a). Color lower case letters next to clusters indicate the polyploids' homeolog-types that generated the overlapping distributions. Color codes for taxa and line styles for ranks of homeolog-type frequencies are indicated in the respective charts.

Figure S3. (a) The optimal *Brachypodium* diploid skeleton trees obtained through distance-based coalescence analyses with the ASTRAL, STAR and STEAC programs using 1877 core transcripts. (b) Bidimensional principal coordinate analysis (PCoA) plot constructed from pairwise patristic distances between diploid ortholog and polyploid homeolog tips of the *Brachypodium* homeologs' ML tree computed with NTSYS-pc v2.10j. Each homeolog was labeled according to the branch of the skeleton diploid tree where it was grafted to (Figure 3a,b) using the '*Nearest Diploid Species Node*' algorithm. A minimum spanning tree was superimposed on the PCoA. The down-frequency-rank of grafting distributions of each polyploid homeolog-type across 100 bootstrap gene trees (Table S4a) is represented as a colored cluster in solid (1st rank), dashed (2nd rank) and dotted (3rd and 4th ranks) lines. Only the most frequent homeolog-types are shown according to ploidy level. Black lower-case letters in squares indicate groups of homeolog-types. Capital letters indicate the inferred homeolog subgenomes using the *Subgenome Assignment* algorithm (Table 1b, Inferred Subgenomes; Table S4b). Clusters represent bootstrap distributions of homeologs with threshold cut-off values over 10% (Table S4a). Color lower case letters next to clusters indicate the polyploids' homeolog-types that generated the overlapping distributions. Color codes for taxa and line styles for ranks of homeolog-type frequencies are indicated in the respective charts.

Figure S4. (a) The *Brachypodium* diploid species trees obtained through distance-based coalescence analyses with the ASTRAL, STAR and STEAC programs using 193 orthologous coding sequences (CDSs) of the *Brachypodium* (*B. stacei* ABR114, *B. distachyon* Bd21, *B. arbuscula* Barb1, *B. sylvaticum* Ain-1) and outgroup diploid references genomes. (b) *Brachypodium hybridum* ABR113 homeologs' ML consensus tree based on 160 orthologous CDS (Table S5) with the polyploid homeolog sequences labeled according to the *Nearest Diploid Species Node* algorithm ('b' and 'd'). (c) *Brachypodium hybridum* ABR113 subgenomic ML consensus tree based on 160 orthologous CDS with its homeolog subgenomes labeled according to the *Subgenome Assignment* algorithm ('B' and 'D'). *Oryza sativa* and *Hordeum vulgare* were used as the outgroups. All SH-aLRT/UltraFast Bootstrap supports are 100/100.

Figure S5. (a) *Brachypodium* nuclear diploid skeleton tree based on 322 core transcripts. (b) *Brachypodium* plastid diploid skeleton tree based on 31 plastid transcripts. The values on branches indicate the SH-aLRT/UltraFast Bootstrap supports.

Figure S6. Distribution of the BAC clones derived from chromosomes Bd1–Bd5 of *B. distachyon* ($2n = 10$, $x = 5$) that were comparatively mapped to, respectively, the chromosomes of the (a) diploid *B. arbuscula* ($2n = 18$, $x = 9$), (b) autohexaploid *B. boissieri* ($2n = 48$, $x = 8 + 8 + 8$) and (c) allotetraploid *B. retusum* ($2n = 32$, $x = 8 + 8$). Only one homolog from a pair is shown. The diagrams next to the *Brachypodium* [Bd, Ba (a); Bd, Bb (b); Bd, Bb, Br (c)] chromosomes align the BAC clones to the homeologous regions (syntenic segments) in the relevant ancestral rice chromosome equivalents Os1–Os12. Black diamonds and dotted lines indicate the hypothetical fusion points of the ancestral rice chromosome equivalents (adapted from IBI 2010). Red dashed lines indicate the chromosomal breakpoints in the Ba-genome chromosomes of *B. arbuscula* (a), Bb-genome chromosomes of *B. boissieri* (b) and Bb- and Br-subgenome chromosomes in *B. retusum* (c) that were found using CCB. Red arrows point to a pericentric inversion that was found on chromosome Bb6 (b, c).

Figure S7. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd1 (a) and Bd2 (c) of *B. distachyon* ($2n = 10$, $x = 5$) mapped to chromosomes Ba2, Ba6 and Ba7 (b) and chromosomes Ba1 and Ba8 (d) of *B. arbuscula* ($2n = 18$, $x = 9$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6. BACs Bd1S/1 and Bd1L/14 from Bd1 mapped to chromosome Ba2, probes Bd1S/3 and Bd1L/12 hybridized to Ba6 and probes Bd1S/7-Bd1L/8 to Ba7 (b). BACs Bd2S/1-2 and Bd2L/5-6 from Bd2 mapped to chromosome Ba1, probes Bd2S/3 and Bd2L/4 hybridized to Ba8 (d). Probes Bd1S/1-3-7, Bd1L/8-12-14 from Bd1 and Bd2S/1-2-3, Bd2L/4-5-6 from Bd2 show the chromosomal breakpoints in *B. arbuscula* (b, d) compared with the chromosomal fusion points in *B. distachyon* (a, c). Probes Bd1S/1+ CEN + Bd1L/14 map to chromosome Ba2, whereas probes Bd1S/3 + CEN + Bd1L/12 hybridized to chromosome Ba6 and Bd1S/7 + CEN + Bd1L/8 to chromosome Ba7 (b), thus indicating the presence of two NCF events in the Bd genome of *B. distachyon* that involve three ancestral chromosomes, which were similar to Ba2, Ba6 and Ba7 of the $x = 9$ genome. Probes Bd2S/1 + CEN + Bd2L/6 hybridized to Ba1, whereas probes Bd2S/3 + CEN + Bd2L/4 map to chromosome Bp8 (d), thus indicating the presence of one NCF event in the Bd genome of *B. distachyon* that involved two ancestral chromosomes that were similar to Ba1 and Ba8 of the $x = 9$ genome of *B. arbuscula*.

Figure S8. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd3 (a) and Bd4, Bd5 (c) of *B. distachyon* ($2n = 10$, $x = 5$) mapped to chromosomes Ba4 and Ba3 (b) and chromosomes Ba5 and Ba4 (d) of *B. arbuscula* ($2n = 18$, $x = 9$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(a). BACs Bd3S/1 and Bd3L/10 from Bd3 mapped to chromosome Ba4, probes Bd3S/3-5 and Bd3L/6-7 hybridized to Ba3 (b). BACs Bd4S/1-3-4 and Bd4L/6-8-10 from Bd4 mapped to chromosome Ba5 (d). Probes Bd5S/1 and Bd5L/2 hybridized to Ba9 (d). Probes Bd3S/1-3-5, Bd3L/6-7-10 from Bd3 show the chromosomal breakpoints in *B. arbuscula* (b) compared with the chromosomal fusion points in *B. distachyon* (a). Probes Bd3S/1+ CEN + Bd3L/10 map to chromosome Ba4, whereas probes Bd3S/5 + CEN + Bd3L/6 hybridized

to chromosome Ba3, thus indicating the presence of one NCF event in the Bd genome of *B. distachyon* that involved two ancestral chromosomes, which were similar to Ba4 and Ba3 of the $x = 9$ genome. Probes Bd2S/1 + CEN + Bd2L/6 hybridized to Ba1, whereas probes Bd2S/3 + CEN + Bd2L/4 map to chromosome Bp8 (d), thus indicating the presence of one NCF event in the Bd genome of *B. distachyon* that involved two ancestral chromosomes that were similar to Ba1 and Ba8 of *B. arbuscula*.

Figure S9. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd1 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to six Bb2 and six Bb6 chromosomes (b) of autohexaploid *B. boissieri* ($2n = 48$, $x = 8 + 8 + 8$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6b. BACs Bd1S/1-2 and Bd1L/13-14 mapped to chromosomes Bb2, probes Bd1S/3-7 and Bd1L/8-12 to Bb6. BAC clones Bd1S/1-3 and Bd1L/12-14 show chromosomal breakpoints in all three subgenomes of Bb compared with the chromosomal fusion points in the genome Bd. Probes Bd1S/1+ CEN + Bd1L/14 map to chromosome Bb2, whereas probes Bd1S/3 + CEN + Bd1L/12 hybridized to chromosome Bb6, thus indicating the presence of one NCF event in the Bd genome of *B. distachyon* that involved two ancestral chromosomes, which were similar to Bb2 and Bb6. Within the BAC triplet Bd1S/3-5, clones Bd1S/4-5 mapped to the opposite chromosome arm compared with Bd1S/3. Probe triplets Bd1S/5-6-7 were characterized by an inverted arrangement of clones on the long arm of Bb2 compared with the short arm of the chromosome Bd. Comparative mapping of Probe Bd1S/7 + CEN + Bd1L/8 indicated the presence of a pericentric inversion within chromosome Bb6 of *B. boissieri*.

Figure S10. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd2 and Bd4 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to six Bb1 and six Bb8 chromosomes (b) of autohexaploid *B. boissieri* ($2n = 48$, $x = 8 + 8 + 8$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(b). BACs Bd2S/1-2 and Bd2L/5-6 mapped to chromosomes Bb1, probes Bd2S/3 and Bd2L/4 to Bb8. BAC clones Bd2S/1-3 and Bd2L/4-6 show chromosomal breakpoints in all three subgenomes of Bb compared with the chromosomal fusion points in the genome Bd. Probes Bd2S/1+ CEN + Bd2L/6 map to chromosome Bb1, whereas probes Bd2S/3 + CEN + Bd2L/4 hybridized to chromosome Bb8, thus indicating the presence of one NCF event in the Bd genome of *B. distachyon* that involve two ancestral chromosomes, which were similar to Bb1 and Bb8. Within the BAC triplet Bd2S/3 + Bd4S/4 + Bd5L/2 and Bd2S/3 + CEN+ Bd4L/6 probes Bd2-derived and Bd4-derived mapped to the opposite arms of the chromosome Bb8 indicating the presence of chromosomal fusion.

Figure S11. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd3 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to six Bb4 and six Bb3 chromosomes (b) of autohexaploid *B. boissieri* ($2n = 48$, $x = 8 + 8 + 8$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(b). BACs Bd3S/1-2 and Bd3L/8-10 mapped to chromosomes Bb4, whereas probes Bd3S/3-5 and Bd3L/6-7 mapped to Bb3. BAC clones Bd3S/1-3 and Bd3L/6-8 show chromosomal breakpoints in all three subgenomes Bb compared with the chromosomal fusion points in the genome Bd. Probes Bd3S/1+ CEN + Bd3L/10 map to chromosome Bb4, whereas probes Bd3S/5 + CEN + Bd3L/6 hybridized to chromosome Bb3, thus indicating the presence of one NCF event in the Bd genome of *B. distachyon* that involve two ancestral chromosomes, which were similar to Bb4 and Bb3.

Figure S12. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd4 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to six Bb5 and six Bb8 chromosomes (b) of autohexaploid *B. boissieri* ($2n = 48$, $x = 8 + 8 + 8$) and from Bd5 of (a) *B. distachyon* mapped to six Bb7 chromosomes (b) of *B. boissieri*. Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(b). BACs Bd4S/1-3 and Bd4L/7-10 mapped to chromosomes Bb5, probes Bd4S/4-5 and Bd4L/6 to Bb8. BAC clones Bd4S/3-5 and Bd4L/6-7 show chromosomal breakpoints in all three subgenomes Bb compared with the chromosomal fusion points in the genome Bd. Probes Bd4S/1+ CEN + Bd4L/10 map to chromosome Bb5, whereas probes Bd4S/5 + CEN + Bd4L/6 hybridized to chromosome Bb8, thus indicating the presence of one NCF event in the Bd genome of *B. distachyon* that involve two ancestral chromosomes. Within the BAC triplets, Bd4S/5 + CEN + Bd4L/6 mapped to the same chromosome arm compared with Bd4 of *B. distachyon* and BACs Bd5S/1 and Bd5L/2 mapped to short and long arms of chromosomes Bb7 of *B. boissieri*, respectively.

Figure S13. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosomes Bd1–Bd4 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to various chromosomes (b) of autohexaploid *B. boissieri* ($2n = 48$, $x = 8 + 8 + 8$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(b). Within each BAC triplet, the probes correspond to different ancestral chromosomes. None of the applied BAC triplets indicate chromosomal fusion or different hybridization pattern among subgenomes.

Figure S14. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd1 and Bd3 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to chromosomes Bb2, Bbr2, Bb6, Br4, Br6 and Br5 (b) of allotetraploid *B. retusum* ($2n = 32$, $x = 8 + 8$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(c). BACs Bd1S/1-2 and Bd1L/13-14 mapped to chromosomes Bb2 and Br2. BAC clones

Bd1S/3-7 and Bd1L/8-12 mapped to chromosome Bb6, probes Bd1S/3-4 and Bd1L/10-12 hybridized to Br4. Moreover, BACs Bd1S/5-7 mapped to Br6 and Bd1L/8-9 hybridized to Br5. BAC clones Bd1S/1-3, Bd1S/3-5, Bd1S/7 + CEN + Bd1L/8, Bd1L/8-10 and Bd1L/12-14 show chromosomal breakpoints in subgenomes Bb and Br compared with the chromosomal fusion points in the genome Bd. Probes Bd1S/1+ CEN + Bd1L/14 map to chromosomes Bb2 and Br2, whereas probes Bd1S/3 + CEN + Bd1L/12 hybridized to chromosomes Bb6 and Br4. Within the BAC triplet Bd1S/3-5, clones Bd1S/4-5 mapped to the opposite chromosome arm of Bb6 compared with probe Bd1S/3. Probe triplets Bd1S/5-6-7 were characterized by an inverted arrangement of clones on the long arm of Bb2 compared with the short arm of the chromosome Bd. Comparative mapping of probes Bd1S/7 + CEN + Bd1L/8 indicating the presence of a pericentric inversion within chromosome Bb6. Within BAC triplet Bd1S/3 + Bd1S/7 + Bd3S/1 probes Bd1S/3 + Bd1S/7 mapped to the same arm of the chromosome Bb6 indicating the presence of chromosomal fusion that involve two ancestral chromosomes, similar to chromosome Bd1 of *B. distachyon*. Moreover, in contrast to *B. distachyon* chromosomes probes Bd1S/3 + Bd3S/1 mapped to the same arm of the chromosome Br4 indicating the presence of chromosomal fusion that involved two ancestral chromosomes.

Figure S15. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd2, Bd4 and Bd5 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to chromosomes Bb1, Br1, Bb8, Br8, Br6 and Br7 (b) of allotetraploid *B. retusum* ($2n = 32$, $x = 8 + 8$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(c). BACs Bd2S/1-2 and Bd2L/5-6 mapped to chromosomes Bb1 and Br1 probes Bd2S/3 and Bd2L/4 to Bb8 and Br8. BAC clones Bd2S/1-3 and Bd2L/4-6 show chromosomal breakpoints in subgenomes Bb and Br compared with the chromosomal fusion points in the genome Bd. Probes Bd2S/1+ CEN + Bd2L/6 map to chromosomes Bb1 and Br1, whereas probes Bd2S/3 + CEN + Bd2L/4 hybridized to chromosomes Bb8 and Br8. Within the BAC triplet Bd2S/3 + Bd4S/4 + Bd5L/2 probes Bd2-derived and Bd4-derived mapped to the opposite arms of the chromosome Bb8 indicating the presence of chromosomal fusion in Bb that differentiate it from genome Br.

Figure S16. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd3 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to chromosomes Bb4, Br4, Bb3 and Br3 (b) of allotetraploid *B. retusum* ($2n = 32$, $x = 8 + 8$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(c). BACs Bd3S/1-2 and Bd3L/8-10 mapped to chromosomes Bb4 and Br4, whereas probes Bd3S/3-5 and Bd3L/6-7 mapped to Bb3 and Br3. BAC clones Bd3S/1-3 and Bd3L/6-8 show chromosomal breakpoints in subgenomes Bb and Br compared with the chromosomal fusion points in the genome Bd. Probes Bd3S/1+ CEN + Bd3L/10 map to chromosomes Bb4 and Br4, whereas probes Bd3S/5 + CEN + Bd3L/6 hybridized to chromosomes Bb3 and Br3.

Figure S17. BAC-FISH-based comparative chromosome barcoding with the clones derived from chromosome Bd4 and Bd1 of (a) *B. distachyon* ($2n = 10$, $x = 5$) mapped to chromosomes Bb5, Br5, Bb6, Br6 and Bb8 (b) of allotetraploid *B. retusum* ($2n = 32$, $x = 8 + 8$). Only one homolog from a pair is shown. The colors of the BAC identifiers in the first column indicate the fluorochrome that was used [green, FITC; red, tetramethylrhodamine; yellow (false color), Alexa Fluor 647]. The chromosomes were counterstained with DAPI (blue). The colored bars on the left and the BAC identifiers that were assigned to specific clones correspond to those on the cytogenetic maps in Figure S6(c). BACs Bd4S/1-3 and Bd4L/7-10 mapped to chromosome Bb5 and probes Bd4S/4 and Bd4L/6 hybridized to chromosome Bb8. Probes Bd4S/1-5 and Bd4L/6 hybridized to Br6, BAC clones Bd4L/7-10 mapped to chromosomes Br5. BAC clones Bd4L/6-8 and Bd4S/1 + Bd4S/5 + Bd4L/8 show chromosomal breakpoints in both Bb and Br subgenomes compared with the chromosomal fusion points in the genome Bd. In contrast to Bd genome, BAC clones Bd4S/3-5 show chromosomal breakpoint only in subgenome Bb and probes Bd4S/1 + CEN + Bd4L/10 show chromosomal breakpoint only in subgenome Br. Within BACs triplet Bd4S/1 + Bd4S/5 + Bd4L/8, probes Bd4S/1 and Bd4S/5 are mapped to the same chromosome arm of Br6, probes Bd4S/1 and Bd4L/8 are mapped to the opposite chromosome arms of Bb5. Within BACs triplets Bd1S/7 + Bd4S/1 + Bd4L/8 and Bd1S/7 + Bd4S/1 + Bd4S/4 and Bd1S/5 + Bd1S/7 + Bd4S/1, probes derived from interstitial part of the short arm of Bd1 and short arm of Bd4 mapped to the opposite arms of the chromosome Br6 indicating the presence of chromosomal fusion in contrast to genome Bd and subgenome Bb. Moreover, within BACs triplet Bd4S/4 + Bd4L/8 + Bd1L/8, probes derived from interstitial part of the long arm of Bd1 and long arm of Bd4 mapped to the opposite arms of the chromosome Br5 indicating the presence of chromosomal fusion in contrast to genome Bd and subgenome Bb.

Table S1. (a) Relative frequency (%) of polyploid homeologs detected in the studied *Triticum* species by our *Nearest Diploid Species Node* algorithm across the aligned core genes and their grafted ranked position (rank) across 100 bootstrap replicates. Only those occurring in more than 10% of the selected genes in each accession (see Table 1a, Homeolog-types) were included in the analysis. Homeologs were classified into nine types ('a' to 'i') according to their grafting positions in the diploid tree. Columns a–i indicate the global percentage of a homeolog-type across 100 random bootstrap replicates from pruned alignments (all diploid orthologs plus the polyploid homeolog), in which the *Nearest Diploid Species Node* algorithm grafted the homeolog in a particular branch. The branch with the highest frequency is underlined, and 10% of this value represents the lowest threshold allowed to include additional branches in the distribution of that homeolog-type. Only the most frequent homeologs (indicated with asterisks) were assigned to subgenomes, matching the ploidy level of each studied polyploid species informed by cytogenetic data (Marcussen *et al.*, 2014). (b) Frequency of polyploid heterologous copies across genes and subgenomes (homeologous subgenomes inferred according to our *Subgenome Assignment* algorithm) in *Triticum* polyploid species. All the *Triticum* homeologs were assigned to simple subgenomes (containing only one homeolog-type). Heterolog-types' ranked bootstrap values are represented as subgenomic components in Figure S2. (c) Pairwise patristic distances between the *Triticum–Aegilops* diploid orthologous branches and the polyploid homeologous subgenomic branches of the consensus ML tree based on 48 nuclear core genes, and 181 orthologous and homeologous sequences. Patristic distances were calculated with Geneious R11.1.5. Abbreviations of accessions correspond to *Triticum urartu* (Tura) ($2x$, $2n = 14$), *T. aestivum* (Taes) ($6x$, $2n = 42$), *T. turgidum* (Ttur) ($4x$, $2n = 28$), *T. monococcum* (Tmon)

(2*x*, 2*n* = 14), *Aegilops sharonensis* (Asha) (2*x*, 2*n* = 14), *Ae. tauschii* (Atau) (2*x*, 2*n* = 14) and *Ae. speltoides* (Aspe) (2*x*, 2*n* = 14).

Table S2. List of *Brachypodium* species and ecotypes and outgroup taxa used in the study. Information on locality of origin, accession code, voucher code, genome size (GS), chromosome number (2*n*), chromosome base number (*x*), inferred ploidy level, life cycle, and source of transcriptomic, cytogenetic, genomic and comparative chromosome barcoding (CCB) data are provided for each accession. Genome size and chromosome number values obtained in this work are shown in bold. Vouchers are deposited in the JACA (Pyrenean Institute of Ecology-CSIC, Jaca, Spain) and University of Zaragoza (UZ) herbaria.

Table S3. (a) Total filtered paired-end (PE) and single-end (SE) reads used to build the RNA-seq dataset of the *Brachypodium* species, ecotypes and outgroup taxa under study. Newly generated data are indicated in bold. Crosses indicate transcriptome data obtained in other studies. (b) Statistics of the assembled transcripts obtained from the *Brachypodium* species and ecotypes under study using Trinity assembler. Genes correspond to Trinity components, while transcripts include all the assembled isoforms. Contig N50 indicates that at least half of all assembled nucleotides are in transcript contigs of at least the detected N50 length value. Sources of accessions are indicated in Table S2.

Table S4. (a) Relative frequency (%) of polyploid homeologs detected in the studied *Brachypodium* species by our *Nearest Diploid Species Node* algorithm across the aligned core transcripts and their grafted ranked position (rank) across 100 bootstrap replicates. Only those occurring in more than 10% of the selected genes in each accession (see Table 1b, Homeolog-types) were included in the analysis. Homeologs were classified into nine types ('a' to 'i') according to their grafting positions in the diploid skeleton tree. Columns a–i indicate the global percentage of a homeolog-type across 100 random bootstrap replicates from pruned alignments (all diploid orthologs plus the polyploid homeolog), in which the *Nearest Diploid Species Node* algorithm grafted the homeolog in a particular branch. The branch with the highest frequency is underlined, and 10% of this value represents the lowest threshold allowed to include additional branches in the distribution of that homeolog-type. Only the most frequent homeologs (indicated with asterisks) were assigned to subgenomes, matching the ploidy level of each studied polyploid species informed by cytogenetic data (Table S2). (b) Frequency of polyploid heterologous copies across genes and subgenomes (homeologous subgenomes inferred according to our '*Subgenome Assignment*' algorithm) in *Brachypodium* polyploid species. The *Brachypodium* homeologs-types were assigned to simple or to compound subgenomes (containing more than one homeolog-type) following the circumscription of bootstrap distribution to contiguous branches [e.g. Bpho422 homeolog-type 'g' showed the highest frequency for its grafted branch 'g' (67.7% BS), but also included the graftings in branches 'f' (8.9% BS), 'h' (8.1% BS) and 'i' (9.4% BS) that were above the 10% threshold, whereas the homeolog-type 'e' only showed a highest frequency grafting to branch 'e' (76.2% BS)]. *Brachypodium* compound subgenome A1 is represented by a + b + c heterolog-types in *B. mexicanum*, A2 by a + b + c + e and a + c in *B. boissieri* and *B. retusum*, respectively, E1 by e + g in *B. retusum*, and G by f + g + h + i in *B. rupestre* and *B. phoenicoides*. Homeolog-types' ranked bootstrap values are represented as subgenomic components in Figure S3. (c) Pairwise patristic distances between the *Brachypodium* diploid orthologous branches and the polyploid homeologous subgenomic branches of the consensus ML tree based on 322 nuclear core genes, and 1307 orthologous and homeologous sequences. Patristic distances were calculated with Geneious R11.1.5. Polyploid accession codes and estimated ploidy correspond to those indicated in Table S2.

Table S5. (a) Homeolog allelic and subgenomic datasets of *Brachypodium hybridum* ABR113. Number (#) and percentage (%) of polyploid homeolog alleles that were detected in the encoding genes of the 322 selected transcripts in this allotetraploid reference genome by our *Nearest Diploid Species Node* and *Bootstrapping Refinement* algorithms using orthologous CDS of the available *Brachypodium* diploid reference genomes (*B. stacei* ABR114, *B. distachyon* Bd21, *B. arbuscula* Barb1, *B. sylvaticum* Ain-1). The homeologs were classified into seven homeolog types ('a'; 'b'; 'c'; 'd'; 'e'; 'f' and 'h') according to their grafting positions in the diploid skeleton tree (Figure S4a–c). The inferred homeologous subgenomes of *B. hybridum* ABR113 were selected and labeled according to the S*ubgenome Assignment* algorithm. (b) Relative frequency (%) of polyploid homeologs detected in allotetraploid *B. hybridum* ABR113 by our *Nearest Diploid Species Node* algorithm across the aligned CDS and their grafted ranked position (rank) across 100 bootstrap replicates. Columns 'a–f' and 'h' indicate the global percentage of a homeolog-type across 100 random bootstrap replicates from pruned alignments (all diploid orthologs plus the polyploid homeolog), in which the *Nearest Diploid Species Node* algorithm grafted the homeolog in a particular branch. The branch with the highest frequency is underlined, and 10% of this value represents the lowest threshold allowed to include additional branches in the distribution of that homeolog-type. Only the most frequent homeologs (b, d) were assigned to subgenomes. (c) Frequency of polyploid heterologous copies across genes and subgenomes (homeologous subgenomes inferred according to our *Subgenome Assignment* algorithm) in *B. hybridum* ABR113; all its homeologs were assigned to simple subgenomes (containing only one homeolog-type).

Table S6. (a) Theoretical distribution of gene trees obtained from the program COAL under ILS applying Strategy (i). Only gene trees showing a topology congruent with the skeleton diploid species trees were selected. Each column represents the theoretical probabilities (> 1%) obtained by grafting homeologous subgenomes ('A', 'B', 'D', 'E1', 'E2', 'G') onto their respective diploid species tree branches ('a + c', 'b', 'd', 'e', 'e', 'g + h') (see Figures 3 and 4). Rows indicate the frequency of gene trees with topologies identical to the species tree with the subgenome grafted to that branch. The most frequent COAL gene tree topology is highlighted in gray. Species tree topologies are shown in Table S6(c). A(a) and A(c) = Bret_A2; B(b) = Bhyb_B; D(d) = Bhyb_D; E1 (e) = Bret_E1; E2(e) = Brup_E2, Bpho6_E2, Bpho422_E2; G(g) = Brup_G, Bpho6_G, Bpho422_G; G(h) = Brup_G represent the subgenomes of the polyploids. Ne is the effective population size (in individuals) for ancestral populations of the tree. Three effective population sizes (Ne = 5E5, 1E6 and 2E6) were tested. (b) Relative frequency (%) of observed (Obs) gene tree grafting distribution of homeologs for *Brachypodium* polyploids and theoretical (COAL) distributions according to three effective population sizes (N = 5E5, 1E6 and 2E6) applying the coalescence-based strategy (i). Theoretical values were obtained by amalgamating in equal proportions the column distributions of Table 1(b). Cells with frequencies < 10% were excluded. The final values were normalized so that each column totals 100%. Theoretical distributions of grafted subgenomes into branches were amalgamated: E2(e) + G(g) for *B. rupestre* and *B. phoenicoides* Bpho6 and Bpho422; A (a) + E1(e) for *B. retusum*; B(b) + D(d) for *B. hybridum*. In addition, we included the scenarios where *B. rupestre* was formed by E2 (e) + G(h) and *B. retusum* by A(c) + E1(e) distributions because of the high observed frequencies of homeolog graftings in branches 'h' and 'c' in these species, respectively. Branches that represent

the most likely placement of subgenomes according to the *Subgenome Assignment* algorithm are highlighted in color. (c) Species trees used to generate homeologous subgenome distributions (columns of Table S6a). Branch lengths are given in coalescence units (CU) = g/2Ne, where 'g' is the number of generations. Parameters used for COAL were g = 1.5 years and Ne = 5E5, 1E6 and 2E6 individuals. P = *B. pinnatum*, Y = *B. sylvaticum*, A = *B. arbuscula*, D = *B. distachyon*, S = *B. stacei*, O = outgroup, X = polyploid subgenome. Species trees were created from the skeleton diploid tree generated by the standard BEAST analysis (((((P,Y):1.605, A):4.1,D):3.1,S):26.2,O) in My or (((((P,Y):1.07,A):2.73,D):2.07, S):17.4,O) in CU for Ne = 5E5, (((((P,Y):0.535,A):1.365,D):1.035, S):8.7,O) for Ne = 1E6 and (((((P,Y):0.268,A):0.683,D):0.518,S):4.35, O) for Ne = 2E6 by inserting subgenomes in their respective branches. To avoid placing the outgroup within *Brachypodium* we enlarged the stem branch to 17.4 CU. Theoretical probabilities were estimated in each case for the 10 395 topologies that could be computed with 7 tips.

Table S7. (a) Theoretical distribution of gene trees obtained from the program COAL under ILS applying strategy (ii). Only gene trees showing a topology congruent with the diploid skeleton species trees were selected. Each column represents the theoretical probabilities obtained by grafting hypothetical homeologous subgenomes (ancestral: 'A' and 'B'; intermediate: 'E'; recent 'G') to their respective diploid species tree branches. Rows indicate the frequency of gene trees with topologies identical to the species tree with the subgenomes grafted to that branch. The most frequent COAL gene tree topology is highlighted in gray; this topology agrees with the observed species tree topology. Branch lengths of species trees were set to 1 CU (A1, B1, E1 and G1) or 0.5 CU (A05, B05, E05, G05). (b) Relative frequency (%) of theoretical (ILS) gene tree grafting distribution of homeologs for hypothetical polyploids A + B, A + E and E + G. ILS values were obtained by amalgamating in equal proportions the column distributions of (a). Cells with frequencies < 10% were excluded. The final values were normalized so that each column totals 100%. Polyploid subgenomes recovered by the *Subgenome Assignment* algorithm are highlighted in dark gray; they correspond to the expected gene tree topology. (c) Species trees used to generate homeologous subgenome distributions (columns) of (a). Branch lengths are given in coalescence units (CU) = g/2Ne, where 'g' is the number of generations and Ne is the effective population size. P = *B. pinnatum*, Y = *B. sylvaticum*, A = *B. arbuscula*, D = *B. distachyon*, S = *B. stacei*, O = outgroup, X = polyploid subgenome. Theoretical probabilities were estimated in each case for the 10 395 topologies that could be computed with 7 tips.

Table S8. BAC clones used for the comparative chromosomes barcoding analysis.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J. et al. (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, eaar7191.

Bertrand, Y.J.K., Scheen, A.-C., Marcussen, T., Pfeil, B.E., De Sousa, F. & Oxelman, B. (2015) Assignment of homoeologs to parental genomes in Allopolyploids for species tree inference, with an example from *Fumaria* (Papaveraceae). *Systematic Biology*, **64**, 448–471.

Betekhtin, A., Jenkins, G. & Hasterok, R. (2014) Reconstructing the evolution of brachypodium genomes using comparative chromosome painting. *PLoS One*, **9**, e115108.

Bettgenhaeuser, J., Corke, F.M.K., Opanowicz, M., Green, P., Hernández-pinzón, I., Doonan, J.H. et al. (2017) Natural variation in *Brachypodium* links vernalization and flowering time loci as major flowering determinants. *Plant Physiology*, **173**, 256–268.

Bombarely, A., Coate, J.E. & Doyle, J.J. (2014) Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. *PeerJ*, **2**, e391.

Bortiri, E., Coleman-Derr, D., Lazo, G.R., Anderson, O.D. & Gu, Y.Q. (2008) The complete chloroplast genome sequence of *Brachypodium distachyon*: sequence comparison and phylogenetic analysis of eight grass plastomes. *BMC Research Notes*, **1**, 61.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D. et al. (2014) BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, **10**, e1003537.

Brassac, J. & Blattner, F.R. (2015) Species-level phylogeny and polyploid relationships in *Hordeum* (Poaceae) inferred by next-generation sequencing and in silico cloning of multiple nuclear loci. *Systematic Biology*, **64**, 792–808.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 1–9.

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 972–973.

Catalán, P. & Vogel, J. (2020) Advances on genomics, biology, ecology and evolution of *Brachypodium*, a bridging model grass system for cereals and biofuel grasses. *The New Phytologist*, **227**, 1587–1590.

Catalán, P., Müller, J., Hasterok, R. et al. (2012) Evolution and taxonomic split of the model grass *Brachypodium distachyon*. *Annals of Botany*, **109**, 385–405.

Catalán, P., Chalhoub, B., Chochois, V. et al. (2014) Update on the genomics and basic biology of *Brachypodium*. *Trends in Plant Science*, **19**, 414–418.

Catalán, P., López-Alvarez, D., Díaz-Pérez, A., Sancho, R. & López-Herranz, M.L. (2016) Phylogeny and evolution of the genus *Brachypodium*. In: Vogel, J.P. (Ed.) *Genetics and genomics of Brachypodium. Plant genetics and genomics: crops models*. Cham; Heidelberg; New York; Dordrecht; London: Springer, pp. 9–38.

Chen, Z.J., Sreedasyam, A., Ando, A. et al. (2020) Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. *Nature Genetics*, **52**, 525–533.

Chernomor, O., von Haeseler, A. & Minh, B.Q. (2016) Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology*, **65**, 997–1008.

Contreras-Moreira, B., Cantalapiedra, C.P., García-Pereira, M.J., Gordon, S.P., Vogel, J.P., Igartua, E. et al. (2017) Analysis of plant pan-genomes and transcriptomes with GET _ HOMOLOGUES-EST, a clustering solution for sequences of the same species. *Frontiers in Genetics*, **8**, 1–16.

Degnan, J.H. & Salter, L.A. (2005) Gene tree distributions under the coalescent process. *Evolution (N. Y).*, **59**, 24.

Díaz-Pérez, A., López-Alvarez, D., Sancho, R. & Catalán, P. (2018) Reconstructing the origins and the biogeography of species' genomes in the highly reticulate allopolyploid-rich model grass genus *Brachypodium* using minimum evolution, coalescence and maximum likelihood approaches. *Molecular Phylogenetics and Evolution*, **127**, 256–271.

Dierckxsens, N., Mardulyn, P. & Smits, G. (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, **45**, e18.

Dinh Thi, V.H., Coriton, O., Le Clainche, I. et al. (2016) Recreating stable *Brachypodium hybridum* allotetraploids by uniting the divergent genomes of *B. distachyon* and *B. stacei*. *PLoS One*, **11**, e0167171.

Doležel, J., Greilhuber, J. & Suda, J. (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, **2**, 2233–2244.

Doyle, J.J. & Egan, A.N. (2010) Dating the origins of polyploidy events. *The New Phytologist*, **186**, 73–85.

Edger, P.P., Poorten, T.J., Vanburen, R. et al. (2019) Origin and evolution of the octoploid strawberry genome. *Nature Genetics*, **51**, 541–547.

Edger, P.P., McKain, M.R., Yocca, A.E., Knapp, S.J., Qiao, Q. & Zhang, T. (2020) Reply to: revisiting the origin of octoploid strawberry. *Nature Genetics*, **52**, 5–7.

Febrer, M., Goicoechea, J.L., Wright, J. et al. (2010) An integrated physical, genetic and cytogenetic map of *Brachypodium distachyon*, a model system for grass research. *PLoS One*, **5**, e13461.

Feng, C., Wang, J., Harris, A.J., Folta, K.M., Zhao, M. & Kang, M. (2021) Tracing the diploid ancestry of the cultivated octoploid strawberry. *Molecular Biology and Evolution*, **38**, 478–485.

Fox, S.E., Preece, J., Kimbrel, J.A., Marchini, G.L., Sage, A., Cruzan, M.B. et al. (2013) Sequencing and de novo transcriptome assembly of *Brachypodium sylvaticum*. *Applications in Plant Sciences*, **1**, 1–8.

Goodstein, D.M., Shu, S., Howson, R. et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, **40**, 1178–1186.

Gordon, S.P., Contreras-Moreira, B., Levy, J.J. et al. (2020) Gradual polyploid genome evolution revealed by pan-genomic analysis of *Brachypodium hybridum* and its diploid progenitors. *Nature Communications*, **11**, 1–16.

Goremykin, V.V., Nikiforova, S.V., Cavalieri, D., Pindo, D.M. & Lockhart, P. (2015) The root of flowering plants and total evidence. *Systematic Biology*, **64**, 879–891.

Grabherr, M.G., Haas, B.J., Yassour, M. et al. (2011) Trinity: reconstructing a full-length transcrptome without a genome from RNA-Seq data. *Nature Biotechnology*, **29**, 644–652.

Hardigan, M.A., Feldmann, M.J., Lorant, A., Bird, K.A., Famula, R., Acharya, C. et al. (2020) Genome synteny has been conserved among the octoploid progenitors of cultivated strawberry over millions of years of evolution. *Frontiers in Plant Science*, **10**, 1–17.

Hasterok, R., Dulawa, J., Jenkins, G., Leggett, M. & Langdon, T. (2006) Multi-substrate chromosome preparations for high throughput comparative FISH. *BMC Biotechnology*, **6**, 1–5.

Hou, L., Xu, M., Zhang, T. et al. (2018) Chromosome painting and its applications in cultivated and wild rice. *BMC Plant Biology*, **18**, 1–10.

Howe, K.L., Contreras-Moreira, B. et al. (2020) Ensembl genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Research*, **48**, D689–D695.

IBI. (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Jenkins, G. & Hasterok, R. (2007) BAC "landing" on chromosomes of *Brachypodium distachyon* for comparative genome alignment. *Nature Protocols*, **2**, 88–98.

Jones, G. (2017) Bayesian phylogenetic analysis for diploid and allotetraploid species networks. *bioRxiv* 129361. https://doi.org/10.1101/129361.

Junier, T. & Zdobnov, E.M. (2010) The newick utilities: high-throughput phylogenetic tree processing in the UNIX Shell. *Bioinformatics*, **26**, 669–670.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Haeseler, A. & von Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.

Kamneva, O.K., Syring, J., Liston, A. & Rosenberg, N.A. (2017) Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evolutionary Biology*, **17**, 1–19.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Katoh, K., Misawa, K., Kuma, K.I. & Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Kellogg, E.A. (2015) The families and genera of vascular plants. Vol. XIII. *Flowering Plants. Monocots. Poaceae*. K. Kubitzki, ed., New York: Springer.

Levin, D.A. (2013) The timetable for allopolyploidy in flowering plants. *Annals of Botany*, **112**, 1201–1208.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009) The sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, M., Copeland, A. & Han, J. (2011) DUK – a fast and efficient Kmer matching tool. Lawrence Berkeley Natl. Lab. LBNL Pap. LBNL-4516E-Poster p. Available from: http://duk.sourceforge.net/ [Accessed 11th January 2021].

Liston, A., Wei, N., Tennessen, J., Li, J., Dong, M. & Ashman, T.-L. (2020) Revisiting the origin of the octoploid strawberry. *Nature Genetics*, **52**, 2–4.

Liu, L. & Yu, L. (2010) Phybase: An R package for species tree analysis. *Bioinformatics*, **26**, 962–963.

López-Álvarez, D., López-Herranz, M.L., Betekhtin, A. & Catalán, P. (2012) A DNA barcoding method to discriminate between the model Plant *Brachypodium distachyon* and its close relatives *B. stacei* and *B. hybridum* (Poaceae). *PLoS One*, **7**, e51058.

Lusinska, J., Majka, J., Betekhtin, A., Susek, K., Wolny, E. & Hasterok, R. (2018) Chromosome identification and reconstruction of evolutionary rearrangements in *Brachypodium distachyon*, *B. stacei* and *B. hybridum*. *Annals of Botany*, **122**, 445–459.

Lusinska, J., Betekhtin, A., Lopez-alvarez, D., Catalan, P., Jenkins, G., Wolny, E. et al. (2019) Comparatively barcoded chromosomes of *Brachypodium* perennials tell the story of their karyotype structure and evolution. *International Journal of Molecular Sciences*, **20**, 5557.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K. & Schubert, I. (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 5224–5229.

Maccaferri, M., Harris, N.S., Twardziok, S.O. et al. (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. *Nature Genetics*, **51**, 885–895.

Marcussen, T., Sandve, S.R., Heier, L. et al. (2014a) Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**, 1250092.

Marcussen, T., Sandve, S.R., Heier, L., et al. (2014b) Data from: ancient hybridizations among the ancestral genomes of bread wheat, Dryad, Dataset. *Science*, **345**(6194):1250092.

Marcussen, T., Heier, L., Brysting, A.K., Oxelman, B. & Jakobsen, K.S. (2015) From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae). *Systematic Biology*, **64**, 84–101.

Mascher, M., Gundlach, H., Himmelbach, A. et al. (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature*, **544**, 427–433.

Mason, A.S. & Wendel, J.F. (2020) Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Frontiers in Genetics*, **11**, 1–10.

Minh, B.Q., Nguyen, M.A.T. & von Haeseler, A. (2013) Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution*, **30**, 1188–1195.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. (2014) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.

Novikova, P.Y., Hohmann, N., Nizhynska, V. et al. (2016) Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, **48**, 1077–1082.

Ouyang, S., Zhu, W., Hamilton, J. et al. (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Research*, **35**, 883–887.

Oxelman, B., Brysting, A.K., Jones, G.R., Marcussen, T., Oberprieler, C. & Pfeil, B.E. (2017) Phylogenetics of allopolyploids. *Annual Review of Ecology, Evolution, and Systematics*, **48**, 543–557.

Sancho, R., Cantalapiedra, C.P., López-Alvarez, D., Gordon, S.P., Vogel, J.P., Catalán, P. et al. (2018) Comparative plastome genomics and phylogenomics of *Brachypodium*: flowering time signatures, introgression and recombination in recently diverged ecotypes. *The New Phytologist*, **218**, 1631–1644.

Scholthof, K.-B.G., Irigoyen, S., Catalán, P. & Mandadi, K.K. (2018) *Brachypodium*: a monocot grass model system for plant biology. *Plant Cell*, **30**, 1673–1694.

Session, A. & Rokhsar, D. (2020) Discovering subgenomes of octoploid strawberry with repetitive sequences. *bioRxiv* 2020.11.04.330431, 1–21. https://doi.org/10.1101/2020.11.04.330431

Shiposha, V., Marques, I., López-Alvarez, D., Manzaneda, A.J., Hernandez, P., Olonova, M. et al. (2020) Multiple founder events explain the genetic diversity and structure of the model allopolyploid grass *Brachypodium hybridum* in the Iberian Peninsula hotspot. *Annals of Botany*, **125**, 625–638.

Smith, M.L. & Hahn, M.W. (2020) New approaches for inferring phylogenies in the presence of paralogs. *Trends in Genetics*, **S0168-9525**, 30212–30212.

Soltis, P.S. & Soltis, D.E. (2016) Ancient WGD events as drivers of key innovations in angiosperms. *Current Opinion in Plant Biology*, **30**, 159–165.

Soltis, D.E., Visger, C.J., Blaine Marchant, D. & Soltis, P.S. (2016) Polyploidy: pitfalls and paths to a paradigm. *American Journal of Botany*, **103**, 1146–1166.

Stebbins, G.L. (1949) The evolutionary significance of natural and artificial polyploids in the family Gramineae. *Hereditas*, **35**, 461–458.

Thomas, G.W.C., Ather, S.H. & Hahn, M.W. (2017) Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Systematic Biology*, **66**, 1007–1018.

Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 15–30.