

Genome sequencing reveals the genetic architecture of heterostyly and domestication history of common buckwheat

In the format provided by the
authors and unedited

Supplementary Information

Contents

1	Sequencing, assembly, and gene prediction	3
1.1	Constructing pseudomolecules of <i>F. esculentum</i> PL4	3
1.1.1	Sequencing and obtaining long scaffolds and contigs . .	3
1.1.2	Linkage analysis	3
1.1.3	Anchoring scaffolds to linkage groups	4
1.2	Constructing pseudomolecules of <i>F. homotropicum</i>	7
1.3	Assessing and validating the assemblies	8
1.3.1	BUSCO and LAI scores	8
1.3.2	Experimental verification by FISH	8
1.3.3	Comparison with published linkage map	9
1.4	Gene prediction	10
2	Transposable Elements	11
2.1	Proportion of Transposable Elements in <i>Fagopyrum</i> genomes .	11
2.2	Amplification dynamics of Transposable Elements in <i>Fagopyrum</i>	13
2.3	Distribution of Transposable Elements across <i>Fagopyrum</i> genomes	16
2.4	LTR retrotransposons associated with centromeric regions . . .	16
3	Collinearity and whole-genome duplications	18
3.1	Detecting whole-genome duplications in <i>Fagopyrum</i>	18

21	3.2	Number and timing of whole-genome duplications	19
22	3.3	Molecular dating of whole-genome duplications	21
23	4	Gene family evolution	22
24	4.1	Identification of gene families in <i>Fagopyrum</i>	22
25	4.2	Rutin biosynthesis	23
26	4.3	The first waxy buckwheat ever	25
27	4.4	2S albumins: key allergens triggering anaphylaxis	27
28	5	Genetic architecture of heterostyly and its transition to	
29		selfing	29
30	5.1	Background	29
31	5.2	Determining haplotypes associated with the different flower	
32		morphs	31
33	5.3	Knockout of <i>S-ELF3</i> results in long homostyle plants	33
34	5.4	Long homostyle <i>S-ELF3</i> knockout mutants are self-compatible	34
35	5.5	Analyses of <i>F. esculentum</i> PL4 and <i>F. homotropicum</i> genomes	36
36	5.6	Evolutionary origin of <i>S-ELF3</i>	37
37	5.7	Genetic architecture of heterostyly and its breakdown	39
38	6	Population genomic analyses and the domestication of com-	
39		mon buckwheat	41
40	6.1	Background	41
41	6.2	Resequencing and identifying single nucleotide polymorphisms .	42
42	6.3	Genetic relationship between the samples and pattern of poly-	
43		morphism	43
44	6.4	Targets of artificial selection	45

1 Sequencing, assembly, and gene prediction

1.1 Constructing pseudomolecules of *F. esculentum* PL4

1.1.1 Sequencing and obtaining long scaffolds and contigs

We sequenced and assembled the genome of *F. esculentum* PL4, a self-compatible accession [1], to be used as a reference genome of common buckwheat. We generated two types of assemblies. One was based on Illumina short reads which have low error rates to use as the main sequence assembly in constructing pseudomolecules (referred to as NRG scaffolds) (Supplementary Table 1). The other was based on PacBio HiFi long reads which are suitable for obtaining long contigs that were used as supporting information to correct misassembled NRG scaffolds and anchor additional NRG scaffolds to linkage groups (referred to as HiFi contigs) (Supplementary Table 2).

We first generated Illumina short reads (374-fold genome coverage) and constructed an assembly with DenovoMAGIC 3.0 (Supplementary Table 1). This resulted in 3,097 scaffolds with high N50 and N90 (28.8 Mb and 9.1 Mb, respectively) (Supplementary Table 3). The total length of the scaffolds was 1.27 Gb, which is consistent with the genome size of 1.27 Gb estimated based on *k*-mer distribution (Supplementary Fig. 2). We then generated PacBio HiFi long reads (15-fold genome coverage) and assembled the reads with hifiasm [2]. This resulted in 457 contigs with a total length of 1.24 Gb with an N50 and N90 of 16.9 Mb and 3.6 Mb, respectively (Supplementary Table 3).

1.1.2 Linkage analysis

Linkage maps have been extensively used in genome assembly since the beginning of the genomics era [3], and is a very important source of information especially for crops for which genetic maps are available or easy to construct.

Even in recent years, many tools that help utilize linkage information are still being developed (e.g. ALLMAPS [4], Chromonomer [5], Lep-Anchor [6]). For instance, linkage maps have been utilized in the recent genome sequencing papers of kiwifruit [7], *Sphagnum* [8], and *Linum* in which linkage information was used to correct misassemblies in the initial assembly that was assembled using Hi-C [9].

To anchor the NRG scaffolds of PL4 to linkage groups, we constructed a linkage map of *F. esculentum* using targeted amplicon sequencing markers from 111 F₂ progeny derived from the cross between cv. Kitawasesoba and PL4 (Supplementary Table 4). 8,664 AmpliSeq-derived variable sites were obtained using Ion AmpliSeq Targeted Sequencing technology. After various filtering steps, 295 SNPs on 59 NRG v1 scaffolds (1.21 Gb in total) were retained and used as AmpliSeq-derived SNP markers. These 295 markers and the flower morphology phenotype marker were mapped to linkage groups using AntMap [10], resulting in eight linkage groups (Supplementary Figs. 3 and 4, Supplementary Tables 5 and 6). We also mapped the 295 markers to the HiFi contigs, which resulted in 287 anchored to 83 HiFi contigs (1.08 Gbp in total). Thirty-six HiFi contigs contained multiple genetic positions on the linkage map to which AmpliSeq markers could be mapped and were subsequently anchored to eight linkage groups (Supplementary Table 6).

1.1.3 Anchoring scaffolds to linkage groups

We constructed pseudomolecules based on the NRG v1 scaffolds anchored to the linkage groups using the linkage map and HiFi contigs to refine the order and direction of the NRG scaffolds. 46 of the 59 NRG scaffolds harboring at least one AmpliSeq-derived SNP marker had more than one marker that mapped to different genetic map positions. We first used these markers as

anchors to determine the direction of the NRG scaffolds on the eight linkage groups. Two (sc0065.1 and sc0529.1) of these 46 scaffolds had markers anchored to different linkage groups. The markers on sc0065.1 were anchored to LG1 and LG4, and those on sc0529.1 were anchored to LG1 and LG7 (Supplementary Table 6). Judging from the high bootstrap values for the order of these markers (72–100%, Supplementary Table 6), sc0065.1 was divided into sc0065.2 and sc0065.3, and sc0529.1 was divided into sc0529.2 and sc0529.3 based on comparison of these two scaffolds and HiFi contigs using MashMap [11].

We then compared the marker positions on the NRG scaffolds and HiFi contigs and found that only one marker, sc0087_05749490 on LG6, did not match. sc0087.1 was therefore divided into two sequences, sc0087.2 and sc0087.3. In addition, there were discrepancies between the linkage order of the AmpliSeq markers and the physical order of the AmpliSeq markers on the NRG scaffolds in three regions on LG1, LG2, and LG7 (Supplementary Figs. 3 - 5 and Supplementary Table 6). The physical order was judged to be more likely because (i) there was no discrepancy between the orders of AmpliSeq markers on the NRG scaffolds and HiFi contigs, and (ii) the linkage order of AmpliSeq markers was not supported by high bootstrap values (0–76%, Supplementary Table 6).

We were able to anchor nine of the 13 NRG scaffolds with markers representing only one genetic position to linkage groups by using HiFi contigs that share AmpliSeq markers with multiple NRG scaffolds as a bridge. For instance, NRG scaffold sc0203.1 had seven AmpliSeq markers with the same map position (50.7 cM) on LG1. As HiFi contig pac001 shared markers with sc0203.1 as well as NRG scaffold sc0542.1 that was already anchored to LG1, we were able to determine the orientation of sc0203.1 (Supplementary Figs. 3 and 5, Supplementary Table 6).

While we could not anchor the remaining four NRG scaffolds harboring AmpliSeq markers based on the sharing of markers with HiFi contigs, we identified corresponding regions in HiFi contigs by MashMap for NRG scaffolds sc0088.1, sc0089.1, and sc0262.1 (Supplementary Table 7), and by BLASTN [12] for a short (471 bp) NRG scaffold sc1899.1. This allowed us to anchor these NRG scaffolds to the linkage group based on the positions of the corresponding HiFi contigs in the linkage map. For instance, NRG scaffold sc0088.1 had seven AmpliSeq markers with the same map position (50.7 cM) on LG1, and shared these markers with HiFi contigs pac063, pac096, and pac082. In addition, we found that ~ 1.5 Mb at the end of sc0088.1 corresponds to the end of pac001, which is linked to LG1, with $>99\%$ identity. As such, we were able to determine the orientation of sc0088.1 on LG1 (Supplementary Figs. 3 and 5, Supplementary Table 7).

We then attempted to anchor NRG scaffolds that do not harbour any AmpliSeq marker by utilizing the HiFi contigs. For each long NRG v1 scaffold (>500 kb) without any AmpliSeq marker, we searched for corresponding HiFi contigs using MashMap [11]. Seven NRG scaffolds whose entire length overlapped a HiFi contig with $>99\%$ identity, and an additional NRG scaffold (sc0412.1) whose ~ 1.7 Mb region at the 3'-end showed $>99\%$ identity to a HiFi contig (pac075) were identified (Supplementary Table 7). Based on the markers in the corresponding HiFi contigs, these eight NRG scaffolds could be anchored to the linkage group. As a result, we were able to anchor 67 NRG v1 scaffolds (including four that were split into two) in total to the eight linkage groups (Supplementary Table 8).

As the NRG v1 scaffolds may contain some misassemblies, we mapped 100 kb of the 5'- and 3'-ends of each 67 anchored NRG v1 scaffolds to the HiFi contigs using MashMap. We examined 36 of the end sequences of NRG

scaffolds that show $\geq 99.9\%$ identity over more than 95 kb to a HiFi contig and detected one discrepancy between the anchored positions on the linkage map of the NRG and HiFi contigs. Although sc0010.1 was anchored to LG4 by the AmpliSeq marker, its 3'-end sequence mapped to a HiFi contig pac051 on LG1 (Supplementary Table 9). Thus, sc0010.1 was split into sc0010.2 and sc0010.3 (Supplementary Table 8). However, this was the only discrepancy detected and the pseudomolecules and HiFi assembly showed high collinearity without large structural differences.

As a result of anchoring the NRG v1 scaffolds to the linkage map with the help of HiFi contigs, we obtained eight pseudomolecules, consistent with the haploid chromosome number of common buckwheat (Supplementary Table 8). The total length of the eight pseudomolecules was 1.22 Gb excluding unidentified bases (N), thus covering 96% of the total length of the NRG v1 assembly. The final assembly consisting of eight pseudomolecules and 3,033 unanchored scaffolds were designated as FesPL4.r1.1 and used throughout this study (Main Table 1).

1.2 Constructing pseudomolecules of *F. homotropicum*

We also sequenced and assembled the genome of *F. homotropicum*, a self-compatible close relative of *F. esculentum* by generating PacBio HiFi long reads (21-fold genome coverage) and assembling the reads with hifiasm (Supplementary Table 2). This resulted in 464 contigs with an N50 of 54 Mb, N90 of 14.7 Mb, and a total length of 1.24 Gb, similar to the estimate of 1.21 Gb based on flow cytometry [13] (Supplementary Table 3). Eight pseudomolecules with a total of 1.21 Gb (excluding Ns) were constructed by ordering the contigs based on the collinearity with the *F. esculentum* PL4 pseudomolecules (Supplementary Table 14). The final assembly consisting of eight pseudomolecules

and 428 unanchored contigs were designated as Fho_r1.0 and used throughout this study (Main Table 1).

1.3 Assessing and validating the assemblies

1.3.1 BUSCO and LAI scores

We assessed the completeness of the assemblies of *F. esculentum* PL4 and *F. homotorpicum* first by utilizing the 1,640 benchmarking universal single-copy orthologs (BUSCO) of embryophyta (BUSCO v5) [14]. The BUSCO complete scores were 96.9% and 93.3% for the FesPL4_r1.1 assembly with and without the unanchored scaffolds, respectively, and 96.9% for the Fho_r1.0 assembly both with and without the unanchored contigs. These scores are comparable to that of the *F. tataricum* pseudomolecules [15] (95.3%, Supplementary Table 10). The LTR Assembly Index (LAI) [16] for *F. esculentum* PL4 and *F. homotopicum* was 15.4 and 18.32, respectively, corresponding to ‘reference standard’ and higher than that of the *F. tataricum* genome (7.98, Supplementary Table 11).

1.3.2 Experimental verification by FISH

Due to the current lack of technology to read chromosomal DNA from start to end, genome assemblies are often riddled with gaps and errors due to repetitive sequences. Fluorescence in situ hybridization (FISH) analysis is a powerful method for verifying genome assemblies because it allows direct visualization of the correct sequence and orientation of contigs and scaffolds on chromosomes [17]. Even a high-quality *O. pumila* genome with only 21 sequence gaps constructed using long read sequencing, BioNano analysis, and Hi-C analysis required correction based on FISH analysis [18]. Oligo-FISH, which pools short and unique sequences to use as probes, was recently developed [19, 20]. This

method facilitates the generation of chromosomal markers to validate genome assemblies in non-model organisms that do not have BAC libraries etc.

Here, we designed 24 probes (12 red probes and 12 green probes) and synthesized oligos prelabeled with myTags technology based on the DNA sequence of *F. homotropicum* pseudomolecules (Supplementary Table 12). At least one probe was designed for each arm of the chromosome so that we can verify that the entire pseudomolecule crossing the centromere with many repeats are correctly assembled. Using these probes, we performed oligo-FISH painting on each chromosome of *F. esculentum* PL4, *F. homotropicum*, and *F. esculentum* cv. Harunoibuki. The pattern of oligo-FISH signals was in perfect agreement with its design and consistent between each homologous chromosome of *F. esculentum* PL4, *F. homotropicum*, and *F. esculentum* cv. Harunoibuki (Main Fig. 1d and Supplementary Fig. 6). These results suggest that the pseudomolecules of *F. esculentum* PL4 and *F. homotropicum* are in general correctly assembled.

1.3.3 Comparison with published linkage map

To further verify the assembly of the *F. esculentum* PL4 pseudomolecules, we examined the positions of previously published linkage markers of *F. esculentum* [21] on the PL4 pseudomolecules. These linkage markers, which were designed on contigs obtained from Illumina short reads, comprise nine linkage groups and include the phenotypic marker *S* locus. These contigs were mapped to the PL4 pseudomolecules by BLASTN and 568 contigs mapping to a single locus and 625 mapping to multiple loci were identified. Following various filtering steps, we obtained 291 contigs (24.4%) that map to chromosomes 1 to 8 of the PL4 pseudomolecules (Supplementary Table 13 and Supplementary Fig. 7). Comparison of the previously published linkage map and the PL4 pseudomolecules using these markers showed that the order of markers is largely

consistent between the pseudomolecules and the linkage map, providing further support for our assembly (Extended Data Fig. 1). Together with the linkage map constructed in this study consisting of 296 markers, the assembly of the PL4 pseudomolecules is therefore supported by one marker every 2-2.5 Mb on average distributed across each chromosome. Thus, while oligo-FISH provides visual verification of each assembled pseudomolecule at a large-scale, the two linkage maps provide a much finer-scale verification of our PL4 assembly.

Although our assemblies are supported by various sources of information as we described in this section, we note that there are still assembly gaps and uncalled bases in the pseudomolecules and scaffolds that could not be anchored to the pseudomolecules. As the sequencing and assembly technologies continue to improve, especially with long-read sequencing, our assemblies should serve as important materials to facilitate buckwheat research and the construction of more complete assemblies in the future.

1.4 Gene prediction

Protein-coding genes in the genomes of *F. esculentum* PL4 and *F. homotropicum* were predicted by combining *ab initio* prediction and protein homology using BRAKER2 [22]. For *F. esculentum* PL4, RNA-Seq data were generated to identify high-confidence gene models. The initial prediction of BRAKER2 resulted in 224,148 and 238,445 gene models in *F. esculentum* PL4 and *F. homotropicum*, respectively (Supplementary Table 18). After filtering gene models that are likely to be Transposable Elements (TEs), we obtained 40,446 and 35,350 gene models in *F. esculentum* PL4 and *F. homotropicum*, respectively. These gene models were further examined and based on the degree of transcript and homology support, 30,608 and 31,961 gene models in *F. esculentum* PL4 and *F. homotropicum*, respectively, were assigned as high-confidence

(HC) gene models to be used throughout this study (Supplementary Table 18). These numbers are comparable to the 31,869 gene loci annotated in *F. tataricum* [15]. We note that the number of predicted protein-coding genes reported in *F. tataricum* (33,666) includes multiple transcripts from the same loci.

97% (29,831) and 94% (30,115) of the *F. esculentum* PL4 and *F. homotropicum* HC gene models, respectively, were supported by homology to genes in UniprotKB, NR database of NCBI, or *A. thaliana* (Araport11) or transcript evidence based on RNA-Seq (Supplementary Table 18). The average number and length of CDS, introns, and coding exons were all similar between the two species (Supplementary Table 18). BUSCO analysis based on the 1,640 embryophyta genes showed that 88.3% and 87.4% of the HC genes were covered in *F. esculentum* PL4 and *F. homotropicum*, similar to that of *F. tataricum* (89.5%, Supplementary Table 10).

2 Transposable Elements

2.1 Proportion of Transposable Elements in *Fagopyrum* genomes

TEs are major determinants of the genome size variation in flowering plants [23]. By combining *ab initio* and similarity-based methods, we annotated 960 Mb and 938 Mb as TEs in the genomes of *F. esculentum* and *F. homotropicum*, respectively (Supplementary Table 15). LTR retrotransposons were the most abundant, comprising 940 Mb and 908 Mb of *F. esculentum* and *F. homotropicum*, respectively. In particular, *Gypsy*-type LTR retrotransposons comprised 826 Mb of *F. esculentum* and 798 Mb of *F. homotropicum*. These included 24,656 full-length copies that accounted for 266 Mb in *F. esculentum* and 24,426 that accounted for 260 Mb in *F. homotropicum* (Supplementary

Table 16). A much higher proportion of *Gypsy* elements compared to *Copia* elements is consistent with previous analyses based on fragmented draft genomes of *F. esculentum* [24, 25], and has also been reported for the *F. tataricum* genome [15] and many other species such as coffee [26], sunflower [27], durian [28], and broomcorn millet [29] (also see [18]).

To ensure direct comparison and to use in further analyses, we also annotated TEs in the genome of *F. tataricum* using exactly the same procedure. This resulted in 202 Mb being annotated as TEs, including 185 Mb of LTR retrotransposons of which 148 Mb were *Gypsy*-type LTR retrotransposons (Supplementary Table 15). 5,485 full-length LTR retrotransposons comprising 45 Mb were identified, including 4,392 *Gypsy* elements comprising 37 Mb (Supplementary Table 16). These numbers, especially those of LTR retrotransposons, were highly similar with the genome sequencing study of *F. tataricum* [15] which annotated 224 Mb as TEs, including 189 Mb of LTR retrotransposons of which 149 Mb were *Gypsy* elements.

Our results show that the differences in the total lengths of LTR retrotransposons, especially of *Gypsy* elements, largely explains the much larger genome size of *F. esculentum* and *F. homotropicum* compared to *F. tataricum*. The total lengths of the genome assemblies of *F. esculentum*, *F. homotropicum*, and *F. tataricum* used in this study were 1,267 Mb, 1,245 Mb, and 451 Mb [15], respectively, meaning that the differences in genome size between *F. esculentum* and *F. tataricum*, and between *F. homotropicum* and *F. tataricum* were 816 Mb and 794 Mb, respectively. Of these differences in genome size, all LTR retrotransposons accounted for 755 Mb between *F. esculentum* and *F. tataricum*, and 723 Mb between *F. homotropicum* and *F. tataricum*, whereas *Gypsy*-type LTR retrotransposons alone accounted for 678 Mb between *F. esculentum* and *F. tataricum*, and 650 Mb between *F. homotropicum* and *F. tataricum*.

2.2 Amplification dynamics of Transposable Elements in *Fagopyrum*

The nucleotide divergence between the flanking LTRs of a given copy corresponds to the time since the insertion of that copy [30]. As such, we examined the past amplification dynamics of LTR retrotransposons in the three *Fagopyrum* species by calculating the nucleotide divergence between the LTRs of each full-length LTR retrotransposon of each species. We found that the majority of the full-length *Gypsy* elements in *F. esculentum* and *F. homotropicum* have inserted very recently, suggesting an ongoing active retrotransposition in both species (Extended Data Fig. 3). By contrast, there were very few recent insertions in *F. tataricum* and the activity seems to have decreased in the recent years. A similar pattern was observed for the *Copia* elements (Extended Data Fig. 3). We note that the mode of the divergence of synonymous sites (K_s) between the orthologous protein-coding genes of *F. esculentum* and *F. tataricum* was ~ 0.11 (Extended Data Fig. 5), indicating that almost all full-length LTR retrotransposons analyzed most probably inserted after the divergence of *F. esculentum* and *F. tataricum*. By contrast, the mode of K_s between the orthologous protein-coding genes of *F. esculentum* and *F. homotropicum* was ~ 0.01 - 0.02 (Extended Data Fig. 5), suggesting that *F. esculentum* and *F. homotropicum* both contain many full-length LTR retrotransposons that inserted before and after the divergence of the two species.

Considering that *Gypsy* elements are by far the most abundant type of TEs in the three species and the major determinants of the genome size differences, we focused on the comparison of *Gypsy* elements and whether their much higher activity in *F. esculentum* and *F. homotropicum* was due to the amplification of specific lineages. We performed clustering based on sequence similarity using CD-HIT [31, 32] with the nucleotide sequences of the reverse

transcriptase domain for 13,327, 13,705, and 2,472 full-length *Gypsy* elements in *F. esculentum*, *F. homotropicum*, and *F. tataricum*, respectively for which reverse transcriptase domains could be identified. Many clusters reported by CD-HIT contained much larger number of copies of *F. esculentum* and *F. homotropicum* compared to *F. tataricum*, while there were also a few clusters where *F. tataricum* had a lot more copies than the other two species (Supplementary Table 17). Two clusters in particular (Clusters 18 and 11), one with no copies and the other with a low copy number (54) in *F. tataricum*, appear to have expanded in *F. esculentum* and *F. homotropicum* and make up 40% and 45% of the full-length *Gypsy* elements analyzed in *F. esculentum* and *F. homotropicum*, respectively. We did not observe any clusters where the copy numbers of *F. esculentum* and *F. homotropicum* were largely different from each other.

We then performed phylogenetic analysis in order to understand the relationship between each cluster and also with *Gypsy* elements from other plant species. A previous study identified seven *Gypsy* clades in angiosperms, namely *Tekay*, *Reina*, *Galadriel*, *CRM*, *Athila*, *Retand*, and *Ogre*, and constructed a database, REXdb with sequences from various species of each clade [33]. We obtained 10 randomly selected sequences from eudicot species of each clade in REXdb and constructed a phylogenetic tree together with the representative sequence reported by CD-HIT of each *Fagopyrum* cluster containing at least 100 sequences. The resulting tree showed that most of the *Gypsy* elements in *Fagopyrum* species are from the *Athlia*, *CRM*, and *Tekay* families (Supplementary Fig. 12). Multiple lineages from the *Athlia* clade have expanded in *F. esculentum* and *F. homotropicum*, whereas many of the clusters with high copy numbers in *F. tataricum* were from the *Tekay* clade.

We assigned each full-length *Gypsy* elements in clusters containing at least 100 sequences to *Athlia*, *CRM*, or *Tekay* according to the phylogenetic tree (Supplementary Fig. 12 and Supplementary Table 17), and then obtained the distribution of the nucleotide divergence between the flanking LTRs for the *Athlia*, *CRM*, and *Tekay* families in each of the three species and for the two largest subfamilies (Clusters 18 and 11). The distributions of all *Athlia* elements and the two largest clusters were both similar to the distribution of all *Gypsy* elements, indicating very high recent activity of the *Athlia* elements, especially of two subfamilies in *F. esculentum* and *F. homotropicum* (Extended Data Fig. 3 and Supplementary Fig. 11). We also constructed phylogenetic trees for the two largest subfamilies (Supplementary Fig. 13). There were no large, well-supported clades specific to *F. esculentum* or *F. homotropicum* in either subfamily, whereas all the *F. tataricum* sequences in Cluster 11 formed a species-specific, well-supported clade. By contrast, the distribution of the *Tekay* clade showed that *Tekay* underwent a slightly older amplification burst in *F. tataricum* and that its recent activity has decreased.

A recent study that sequenced the genome of *F. esculentum* and constructed a draft genome assembly also reported a much higher amount of *Athlia* elements in *F. esculentum* compared to *F. tataricum*, and that the *Tekay* elements are abundant in *F. tataricum* [25], consistent with our results. However, they did not report any difference in the recent activity of *Athlia* elements between *F. esculentum* and *F. tataricum*, although they found that some *Copia* families showed much higher recent activity in *F. esculentum* compared to *F. tataricum*. Our results clearly demonstrate that *Athlia* elements show a much higher recent activity in *F. esculentum* and *F. homotropicum*, and that two subfamilies in particular have recently expanded in both *F. esculentum* and *F. homotropicum* but not in *F. tataricum*.

2.3 Distribution of Transposable Elements across

Fagopyrum genomes

We then examined the distribution of TEs across each chromosome in the three *Fagopyrum* species. We found that the density of TE is low in the distal ends of some chromosomes but high in the rest of the genome in *F. esculentum*, *F. homotropicum*, and to a lesser extent, in *F. tataricum*. These distal ends of chromosomes with low TE density showed a very high density of protein-coding genes compared to the rest of the genome and low GC content in each species (Supplementary Fig. 8). The uneven TE density across the genome mostly reflects the distribution of *Gypsy* elements, in particular of the recently expanded *Athila* subfamilies in *F. esculentum* and *F. homotropicum*. By contrast, the density of *Copia* elements and other classes of TEs such as DNA transposons and LINEs was higher in the distal ends of chromosomes and lower in the rest of the genome (Extended Data Fig. 4, Supplementary Figs. 9 and 10). Similar observations of high density of protein-coding genes, low density of *Gypsy* elements, and low GC content in the distal regions, and different density across the genome for different TE families have been reported for many other plant species such as barley [34].

2.4 LTR retrotransposons associated with centromeric regions

We found that some, but not all subfamilies of the *CRM* clade were highly enriched at specific regions of each chromosome of *F. esculentum* and *F. homotropicum*, but not *F. tataricum* (Extended Data Fig. 4, Supplementary Figs. 9 and 10). For instance, full-length copies of Clusters 3, 8, 9, 15, and 16 tightly clustered within a small genomic region of each chromosome in *F. esculentum* and *F. homotropicum*.

The *CRM* clade has its name derived from Centromeric Retrotransposon of Maize, and due to the presence of a targeting domain in the integrase C-terminus, some lineages of this clade are known to preferentially accumulate in centromeric regions in a wide range of plant species [35]. It has been suggested that the presence of these *CRM* elements in centromeric regions contributes to the correct functioning of centromeres [35]. We therefore performed FISH analysis in *F. esculentum* and *F. homotropicum* using probes with sequences from the *CRM* clade and confirmed that these regions correspond to centromeric regions with the primary constriction (Main Fig. 1e and Supplementary Fig. 15). Although the exact delineation of centromeric or peri-centromeric region is difficult, we will here refer to the *CRM*-rich regions of *F. esculentum* and *F. homotropicum* as centromeric regions and refer to the copies in two species in the Clusters 3, 8, 9, 15, and 16 as *FeCEN*.

The five *CRM* subfamilies associated with centromeric regions, i.e., *FeCEN* showed reduced recent activity and most copies had a slightly older insertion age compared to the most active *Athila* subfamilies (Supplementary Fig. 11). The peak of the insertion ages of these subfamilies was much younger than the K_s between the orthologs of *F. tataricum* and the other two species (~ 0.11), and similar to that of the K_s between the orthologs of *F. esculentum* and *F. homotropicum* (~ 0.01 - 0.02). By contrast, two other subfamilies of the *CRM* clade, Clusters 44 and 88 did not cluster in centromeric regions and were distributed across the genome in both *F. esculentum* and *F. homotropicum* and showed high recent activity in both species. In *F. tataricum*, although Cluster 88 had amplified recently, the other *CRM* subfamilies were only present at low-copy and none of them clustered together at specific chromosomal regions.

The non full-length copies of Clusters 3, 8, 9, 15, and 16 were also enriched in the centromeric regions of *F. esculentum* and *F. homotropicum*, although

they were also observed at low copies in other parts the genome (Supplementary Figs. 14). No preferential accumulation of the Clusters 3, 8, 9, 15, and 16 in any specific region in *F. tataricum* was observed even when all the non full-length copies were considered, apart from a slight enrichment in a region of Chr 6 (Supplementary Fig. 14). We could not identify any other LTR retrotransposon lineage that tightly cluster at specific genomic regions of *F. tataricum*. These results suggest that the accumulation of certain *CRM* sub-families around centromeric regions is a feature that evolved in *F. esculentum* and *F. homotropicum* after the divergence of their common ancestor from the lineage of *F. tataricum*.

3 Collinearity and whole-genome duplications

3.1 Detecting whole-genome duplications in *Fagopyrum*

Whole-genome duplications (WGDs) have played a crucial role in the evolution of angiosperms [36]. In particular, it is known that most eudicot lineages share a hexaploidy event that occurred in their common ancestor [37, 38]. *Fagopyrum* is within Polygonaceae, an early branching family of Caryophyllales. A previous genome analysis of *F. tataricum* reported a WGD specific to *F. tataricum* 64.42-70.77 million years ago (mya) that is not shared with *Beta vulgaris*, a core Caryophyllales species [15]. Another recent study that analyzed the transcriptomes and genomes of 168 Caryophyllales species reported 26 WGDs in Caryophyllales, including one that is shared by all Polygonaceae species investigated including *Fagopyrum* and *Rumex* but not shared by other non-core Caryophyllales families [39].

Here, we first detected collinearity between the genomes of the three *Fagopyrum* species using MCScanX [40] (Main Fig. 2 and Supplementary

Fig. 16). High collinearity was maintained between the genomes of *F. esculentum* and *F. homotropicum*, although several rearrangements were observed between *F. tataricum* and the other two species. This is expected considering the much closer phylogenetic relationship between *F. esculentum* and *F. homotropicum* than with *F. tataricum* [41], and given that the *F. homotropicum* pseudomolecules were assembled based on collinearity with *F. esculentum* PL4. In addition to the orthologous regions, we detected several collinear duplicated blocks where one region of a species was collinear to multiple regions of the other species, consistent with a WGD shared by the three *Fagopyrum* species.

3.2 Number and timing of whole-genome duplications

If there were a single WGD in the common ancestor of *Fagopyrum* after its divergence from core Caryophyllales species as previously suggested, one would expect most regions of one species to be collinear to two regions of the other species even if a small number of regions may be collinear to more than two regions due to the ancestral hexaploidy event [37]. Instead, we found several regions collinear to more than two, and in many cases four other regions of the other species. A single WGD cannot explain this pattern of collinearity, suggesting that two WGDs may have occurred in the common ancestor of *Fagopyrum*. We therefore examined the collinearity between each *Fagopyrum* species against *B. vulgaris* [42], a reference species of core Caryophyllales, and against *Vitis vinifera* (grapevine) [37] and *Prunus persica* (peach) [43], two genomes that have undergone few rearrangements since the ancestral hexaploidy and are often used as reference genomes for comparative genomics [44]. Indeed, we found that several *Fagopyrum* regions are collinear to four regions of the genomes of *B. vulgaris*, *V. vinifera*, and *P. persica*, suggesting

that two WGDs likely occurred in *Fagopyrum* since the ancestral hexaploidy (Main Fig. 2, Supplementary Figs. 16 and 17).

To further determine the exact number and timing of WGDs in *Fagopyrum*, we constructed gene families with OrthoFinder [45] using the three *Fagopyrum* species, *Rumex hastatulus* [46] which also belongs to Polygonaceae, *Aldrovanda vesiculosa* which is another non-core Caryophyllales species of Droseraceae [47], and four core Caryophyllales species to serve as an outgroup, namely *Portulaca amilis* [48], *B. vulgaris* [42], *Chenopodium quinoa* [49], and *Amaranthus hypochondriacus* [50]. We then extracted gene families (i.e., hierarchical orthologous groups) containing at least two pairs (three genes) of collinear duplicates, representing two WGDs, in one of the *Fagopyrum* species and constructed phylogenetic trees. We were able to determine the timing of two or more duplications relative to the divergence between *Fagopyrum* and *Rumex*, *Aldrovanda*, and core Caryophyllales for 159 gene families, of which 116 were consistent with two duplications being shared with *Fagopyrum* and *Rumex* but not with *Aldrovanda* (Extended Data Fig. 5 and Supplementary Fig. 18). In addition, the mode of K_s between the WGD-derived paralogs of *F. esculentum* was slightly larger than the mode of K_s between the orthologs of *F. esculentum* and *R. hastatulus* and much smaller than that between the orthologs of *F. esculentum* and *A. vesiculosa* (Extended Data Fig. 5). While collinearity-based analysis with the genomes of *Rumex* and *Aldrovanda* would be desirable to confirm, we note that only limited collinearity could be detected between the genomes of *R. hastatulus* and *A. vesiculosa* and any of the other genomes, which may be due to many lineage-specific rearrangements and/or the quality of the assembly. Nevertheless, our results suggest that the most probable scenario is two WGDs having occurred in Polygonaceae in the common ancestor of *Fagopyrum* and *Rumex* after its divergence from the lineage of *Aldrovanda*.

3.3 Molecular dating of whole-genome duplications

We then performed phylogenetic dating of the 116 gene trees with nodes corresponding to two WGDs shared with *Fagopyrum* and *Rumex* but not with *Aldrovanda* following a previously described Bayesian framework [51, 52] using BEAST v2 [53]. As there were no fossil records suitable to apply as age constraints, we used the results of previous molecular dating studies. We applied a normal distribution prior with a mean of 100 mya and standard deviation (SD) of 1 million years to the divergence between *Fagopyrum* and *Aldrovanda* to use as a main age calibration following the results of a previous large-scale molecular dating study of angiosperms [54]. In addition, we applied a normal distribution with a mean of 60 mya and SD of 3 million years to the divergence between *Fagopyrum* and *Rumex*, and a normal distribution with a mean of 16 mya and SD of 5 million years to the divergence between *F. esculentum* and *F. tataricum* as secondary calibrations following a recent molecular dating study of Caryophyllales [55] (Supplementary Table 20). As a result, we estimated the ages of the two WGDs to be 71 mya (95% high posterior density: 59.8-86.6 mya) and 85.7 mya (71.3-96.9 mya) (Extended Data Fig. 5). Thus, the younger WGD is close to, albeit slightly older than the time period around the Cretaceous-Paleogene extinction event (~65 mya) in which many other angiosperm lineages have undergone independent WGDs [51, 56].

We also performed dating without either or both of the secondary calibrations to examine their effects (Supplementary Fig. 19). The ages of the two WGDs were 65.8 mya (39.2-87.5 mya) and 84.4 mya (59.2-97.7 mya) without both secondary calibrations, 71.7 mya (60.1-87.2 mya) and 86.5 mya (71.4-97.4 mya) without the constraints on the divergence between *F. esculentum* and *F. tataricum*, and 63.4 mya (37.8-86.4 mya) and 83.7 mya (58.1-97.1 mya) without the constraints on the divergence between *Fagopyrum* and *Rumex*

(Supplementary Tables 20 and 21). Thus, while the constraint applied to the divergence between *Fagopyrum* and *Rumex* does lead to a slightly older age estimates of the WGDs, the difference is at most ~7-8 mya.

4 Gene family evolution

4.1 Identification of gene families in *Fagopyrum*

We studied the evolution of gene families using the results of OrthoFinder described above based on nine Caryophyllales species including the three *Fagopyrum* species. Using the hierarchical orthologous groups reported by OrthoFinder of all nine species, we examined the shared and unique gene families in *Fagopyrum*, core Caryophyllales species, *R. hastatulus*, and *A. vesiculosa* (Supplementary Fig. 20a). We also determined shared and unique gene families in *F. esculentum*, *F. homotropicum*, *F. tataricum*, and *R. hastatulus* based on the hierarchical orthologous groups of these four species (Supplementary Fig. 20b). A larger number of gene families were shared between *F. esculentum* and *F. homotropicum* (18,209) than between *F. esculentum* and *F. tataricum* (16,580) or between *F. homotropicum* and *F. tataricum* (16,258), reflecting the closer relationship between *F. esculentum* and *F. homotropicum*.

As we were especially interested in genes responsible for phenotypic differences between *F. esculentum* and the other *Fagopyrum* species, we extracted 354 gene families (782 genes) unique to *F. esculentum*, and 1,849 gene families (2,529 genes in *F. esculentum*) shared between *F. esculentum* and *F. homotropicum* but not with *F. tataricum* or *R. hastatulus* based on the hierarchical orthologous groups of the four species (Supplementary Tables 22 and 23). We also extracted 208 gene families shared between *F. homotropicum*, *F. tataricum*, and *R. hastatulus* but not *F. esculentum*, which likely represent gene losses in *F. esculentum* (Supplementary Table 24).

We then searched for gene duplications that occurred in *F. esculentum* after its divergence from *F. homotropicum*, and those that occurred in the common ancestor of *F. esculentum* and *F. homotropicum* after its divergence from *F. tataricum*. Based on the duplications reported by OrthoFinder in the two branches of interest, we retained those within gene families shared with at least one core Caryophyllales species and either *R. hastatulus* or *A. vesiculosa*. In addition, only gene families containing orthologs of *F. esculentum* and *F. homotropicum* and orthologs of *F. esculentum* and *F. tataricum* as reported by OrthoFinder were retained. As a result, we identified 563 gene families with gene duplications in *F. esculentum* after its divergence from *F. homotropicum*, and 851 with gene duplications in the common ancestor of *F. esculentum* and *F. homotropicum* after its divergence from *F. tataricum*. Lists of *F. esculentum* genes in gene families described above (*F. homotropicum* for gene families not shared with *F. esculentum*) and their best hits by similarity searches against NR database of NCBI (DIAMOND search [57], more sensitive mode) and *A. thaliana* proteins (BLASTP search [12]) are provided in Supplementary Tables 22- 26. Although such differences in gene families between species may also result from problems in the assembly or annotation, these lists should serve as useful information in studying the genetic basis of the phenotypic variation in *Fagopyrum* species.

4.2 Rutin biosynthesis

Fagopyrum species are known for their very high content of rutin, which is a kind of flavonoid. In particular, previous studies have reported that the rutin content of *F. tataricum* is much higher than that of *F. esculentum* and *F. homotropicum* [58]. We examined the rutin content in PL4 and its hybrid parents, *F. esculentum* cv. Kitawasesoba and *F. homotropicum* C9255 and

confirmed that their rutin contents were much lower than that of *F. tataricum* (Supplementary Table 27).

The previous genome analysis of *F. tataricum* identified several genes in *F. tataricum* involved in the biosynthetic pathway of flavonoids such as rutin and reported that some of these gene families expanded by gene duplication in *F. tataricum* compared to *Arabidopsis* [15]. Here, we identified gene families involved in phenylpropanoid synthesis, namely phenylalanine ammonia-lyase (PAL), cinnamate-4-hydroxylase (C4H), 4-coumarate CoA ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone-3-hydroxylase (F3H), flavonoid-3'-hydroxylase (F3'H), flavonoid-3'-5'-hydroxylase (F3'5'H), flavonol synthase (FLS), and glycosyltransferase (GTR) based on *F. tataricum* genes in each gene family that were previously reported [15]. We found 15 gene families altogether, each containing one or more genes of all three *Fagopyrum* species (Supplementary Table 28). Note that here a gene family corresponds to a hierarchical orthologous group of Caryophyllales, meaning that duplications predating the origin of Caryophyllales would represent multiple gene families. As such, multiple gene families are reported for 4CL, CHS, F3H, and GTR.

Ten gene families encoding for PAL, C4H, CHS, CHI, FLS, GTR, F3'H and F3'5'H had two or more genes in each *Fagopyrum* species (Supplementary Table 28). We performed phylogenetic analyses for these 10 gene families and confirmed that all 10 underwent gene duplications that are shared with *Fagopyrum* and *Rumex*, and/or specific to *Fagopyrum* (Supplementary Figs. 21 - 23). In addition, WGDs shared by *Fagopyrum* and *Rumex* had contributed to the increase of genes in PAL, C4H, CHS, and FLS whereas tandem gene duplications specific to *Fagopyrum* had contributed to the expansion of PAL, C4H, CHS, CHI, GTR, and F3'5'H gene families. In particular, F3'5'H genes had greatly increased in *Fagopyrum* by several tandem duplications, with 13,

16, and 14 genes identified in *F. tataricum*, *F. esculentum*, and *F. homotrop-*
icum, compared to two in *R. hastatulus*, one in *A. vesiculosa*, and none in the
 core Caryophyllales species, although it is unclear if this enzyme is involved in
 the formation of rutin from quercetin. Likewise, the number of genes encoding
 for GTR, which is involved in the formation of flavonoid glycosides including
 rutin, had also greatly increased in *Fagopyrum*.

Thus, our study shows that genes involved in the pathway of rutin biosyn-
 thesis have increased in *Fagopyrum*, suggesting that some of these duplications
 may have contributed to the high rutin content of *Fagopyrum* species. On the
 other hand, it is worth noting that despite the much higher rutin content of
F. tataricum, the number of genes involved in rutin biosynthesis pathway is
 similar in the three *Fagopyrum* species, with many of the duplications having
 occurred in the common ancestor of the three species.

4.3 The first waxy buckwheat ever

Granule-bound starch synthase (GBSS) plays an important role in the pro-
 cess of amylose synthesis in plants, catalysing the elongation of the amylose
 molecule using ADP-glucose [59, 60]. In cereals, ratio of amylose to amy-
 lopectin in the seed affects the texture of the grain and their dough-based food
 products. The seed storage starch from mutants that have lost GBSS activity
 does not contain any linear polysaccharides (i.e., amylose), but only a highly
 branched polysaccharide (i.e., amylopectin), resulting in the so-called waxy
 phenotype [61–63]. This phenotype imparts a sticky (or glutinous) quality to
 the grain. The waxy and non-waxy phenotypes can be distinguished by iodine
 staining, in which waxy endosperms are stained brownish red, whereas non-
 waxy starch endosperms are stained dark blue. Although the waxy type has
 been used in various cereals and pseudo-cereals [61–63], no practically useful

waxy common or tartary buckwheat has been identified to date ([64], reviewed in [65]).

The previous draft genome of *F. esculentum* contained three *GBSS* genes [24]. Using one of these genes (Fes_sc0005258.1.g000004) as query in a similarity search against the predicted genes of *F. esculentum* PL4, we identified five homologous genes in the *F. esculentum* PL4 genome (Supplementary Table 29). These five genes were in two different gene families (hierarchical orthologous groups) reported by OrthoFinder with the nine Caryophyllales species, which included five genes each of *F. homotropicum* and *F. tataricum*. We merged these two gene families and constructed a phylogenetic tree. The *GBSS* genes of the three *Fagopyrum* species formed five phylogenetic groups, each containing one *GBSS* gene from each three species (Main Fig. 3a). Thus, the copy number of *GBSS* genes had already increased to five before the divergence of these three *Fagopyrum* species. While the oldest duplication predates the divergence of *Fagopyrum* and core Caryophyllales, the other three duplications correspond to the two WGDs in the lineage of *Fagopyrum*. RNA-Seq analysis showed that two genes, *FeGBSS1* and *FeGBSS2*, were highly expressed in the endosperm of buckwheat. The TPM (transcripts per million) values of these two genes accounted for 99.9% of the total TPM values of the five *FeGBSS* genes (Supplementary Table 30).

We therefore hypothesized that a double recessive homozygous mutant of these two genes should exhibit a waxy trait. To test this hypothesis, we first screened plants with knockout mutations in *FeGBSS1* and *FeGBSS2* from the EMS-induced mutant population using NGS-TILLING [66] and obtained two plants, one with a nonsense mutation in the third exon of *FeGBSS1* and the other with a nonsense mutation in the fourth exon of *FeGBSS2*. Although the wild type protein of *GBSS* has two large domains at the N- and C-termini,

which form a rigid structure shown by X-ray structural analysis [67], both mutants lacked the entire domain at the C-terminal (Supplementary Fig. 24).

As the two nonsense mutations were heterozygous, we obtained plants homozygous for the nonsense mutation in *FeGBSS1* and plants homozygous for the nonsense mutation in *FeGBSS2* by crossing. We further obtained plants carrying both knockout mutations as homozygous by crossbreeding. The seeds of each mutant were stained with iodine solution, and while the endosperms of the single recessive homozygous mutants of both *FeGBSS1* and *FeGBSS2* displayed a purple staining, similar to that of the wild type, the endosperm of the double recessive homozygous mutant displayed a brownish red staining (Main Fig. 3b). Because common buckwheat is an outcrossing plant, it is difficult to maintain double recessive homozygotes, which is probably one of the reasons why no waxy common buckwheat plants had been identified to date. As a sticky grain quality is often preferred in Asian countries [61], the first waxy common buckwheat in the world obtained in this study will be an important breeding material for the future.

4.4 2S albumins: key allergens triggering anaphylaxis

Allergies to common buckwheat pose serious health problems in East Asian countries, including Japan, Korea and buckwheat consuming areas of China [68], and are becoming an issue in Western countries [69] where buckwheat consumption has increased due to it being a gluten-free food. Several buckwheat allergens such as *Fag e 1* (β polypeptide of 13S globulin), *Fag e 2* (2S albumin) and *Fag e 3* (7S globulin) have been identified in common buckwheat (Reviewed in [70]). Among them, *Fag e 2* is considered to be the causative allergen of anaphylaxis of common buckwheat [71, 72].

We previously identified *Fag e 2* and three conserved homologues within a 108 kb genomic region in the draft genome of *F. esculentum* [24]. Using one of these genes (Fes_sc0000087.1.g000014) as query in a similarity search against the predicted genes of *F. esculentum* PL4, we identified four genes within a 123.4 kb genomic region on Chr 3 of *F. esculentum* PL4. These four genes were in the same gene family reported by OrthoFinder using the nine Caryophyllales species, which included four *F. homotropicum* genes in a 108.4 kb region on Chr 3 of *F. homotropicum*, and four *F. tataricum* genes in a 121.8 kb region on Chr 7 (Ft7) of *F. tataricum* (Supplementary Table 31, Extended Data Fig. 6a, and Supplementary Fig. 25). The physical order of these four genes was conserved among the three *Fagopyrum* species, indicating that the duplication events giving rise to these 2S albumin genes had occurred in the common ancestor of these species. 2S albumins are important seed storage proteins generally encoded by multigene families that have been reported as major food allergens in seeds of many plants [73, 74]. The duplication and retention of these genes in *Fagopyrum* may be related to the fact that albumin is one of the major storage proteins of buckwheat seeds, accounting for ~25% of their total seed storage proteins [75].

We performed phylogenetic analysis and found that *Fagopyrum* 2S albumins fall into three phylogenetic groups (Extended Data Fig. 6b). *Fag e 2* was included in Group I together with *Fag t 2*, a known allergen of *F. tataricum* [76]. The eight cysteine residues important for the conformation of the 2S albumin structure [77, 78] were conserved in all four sequences of the three species. However, the amino acid sequences were highly variable (Extended Data Fig. 6a, b). Notably, the epitope sequence (EGVRDLKE) [72], i.e., the part of the antigen recognized by the immune system, was highly variable

among the three groups (Extended Data Fig. 6c). Previous work on alanine-scanning mutagenesis [72] has shown that substitution of even one amino acid of the epitope sequence with alanine results in weak IgE-binding activity, suggesting that 2S albumins in Group II and III are associated with less severe anaphylaxis. Considering the importance of a non-functional *Fag e 2* in establishing non-anaphylactic buckwheat, we aimed to produce a non-functional *Fag e 2* by EMS mutagenesis and detected one nonsense mutant allele (*fag e 2_m1*) by NGS-TILLING [66]. *fag e 2_m1* has a stop codon at the 245th position (upstream of epitope sequence), and six of the eight conserved cysteine residues were not included in this mutant allele (Extended Data Fig. 6d). Given that the encoded protein does not contain the epitope sequence, this mutant should be an important material for developing non-/low-allergenic buckwheat in the future.

5 Genetic architecture of heterostyly and its transition to selfing

5.1 Background

Around 200 or more angiosperm genera including *Fagopyrum* exhibit heterostylous or heteromorphic self-incompatibility (SI), in which plants with “pin” (or *L*-morph) flowers which have long styles and short stamens and those with “thrum” (or *S*-morph) flowers which have short styles and long stamens coexist within the population and successful mating only occurs between thrum and pin plants (Main Fig. 1) (reviewed in [79]). The breakdown of heterostylous SI involves not only the transition of SI to selfing but also changes in flower morphology leading to the transition of heterostyly to homostyly, in which all plants within the population exhibit similar style lengths. For

instance, breakdown of heterostylous SI has occurred independently in many *Fagopyrum* species including *F. tataricum* and *F. homotropicum* [80, 81]. Both these species are self-compatible and the stigma and anther of all individuals occupy similar heights. In *Primula* species, it has been reported that almost all self-compatible plants have long homostyle flowers with long style and long stamens [82]. Likewise, *F. esculentum* PL4, a self-compatible accession originating from a cross between a self-incompatible *F. esculentum* accession and the self-compatible *F. homotropicum*, exhibits long homostyle flowers [1, 83]. Note that the term “homostyle” is also used for individuals within a heterostylous population in which heterostyly has broken down.

Heterostylous SI is thought to be regulated by a single genetic locus called *S* locus [79]. Under the classical model originally proposed in *Primula* [84], it used to be thought that thrum plants are heterozygous and possess the dominant *S* haplotype and recessive *s* haplotype, pin plants are homozygous of the recessive *s* haplotype, whereas the long homostyle self-compatible phenotype is conferred by the S^h haplotype that is dominant to the *s* haplotype and recessive to the *S* haplotype. Indeed, it has been proposed that *F. homotropicum* has the S^h haplotype [85] and that the long homostyle and self-compatible phenotype of *F. esculentum* PL4 is due to the introduction of the S^h haplotype from *F. homotropicum* [83]. On the basis that heterostyly involves the complete association of mating compatibility and morphology of both the female (i.e., style) and male (i.e., stamen) organs, the classical model has assumed that the *S* locus consists of five tightly linked genes regulating the style length (*G*), style incompatibility (I^S), pollen incompatibility (I^P), pollen size (*P*) (thrum plants have larger pollens), and anther height (*A*) [84].

However, recent studies in *Primula* [86–88], *Turnera* [89, 90], *Linum* [9], and *Fagopyrum* [81] have all shown that the dominant *S* haplotype consists of a

hemizygous region that is absent from the *s* haplotype. Moreover, it was shown in *Primula* that inactivation of a single gene converts short thrum styles into long styles with the same compatibility behavior as pin styles, suggesting that the transition of heterostyly to selfing long homostyle plants can be achieved by a single mutation, instead of recombination between the *S* and *s* haplotypes as suggested by the classical heterozygous model [82, 86].

In *F. esculentum*, a protein-coding gene *S-ELF3* was previously identified by screening genes differentially expressed in pistils of short and long-styled flowers [81]. The region containing this gene was specific to thrum plants and absent from pin plants, meaning that thrum plants are hemizygous and pin plants are null homozygous. Moreover, *S-ELF3* was disrupted in self-compatible species *F. tataricum* and *F. homotropicum*, indicating that it is likely involved in heterostyly [81]. As the self-compatible *F. esculentum* PL4 should not have the *S* haplotype and our previous draft assembly of a self-incompatible *F. esculentum* [24] was too fragmented, we generated a phased genome assembly of a self-incompatible *F. esculentum* line GS1 [91] to determine the *S* and *s* haplotypes and to understand the genetic architecture of heterostyly and its breakdown in PL4 by comparison with the PL4 genome.

5.2 Determining haplotypes associated with the different flower morphs

To determine the two different haplotypes, i.e. *S* haplotype and *s* haplotype, we assembled phased scaffolds of a thrum plant (*Ss*) of *F. esculentum* GS1. We generated a total of 581.1 Gb Illumina reads (480-fold genome coverage) and obtained phased genome sequences using DenovoMAGIC (Supplementary Table 1). This resulted in 42,256 scaffolds with an N50 of 2.83 Mb (Supplementary Table 3), which is much lower than that of PL4 (28.7 Mb), but much

higher than the two previously published draft genomes (25 kb [24] and 180 kb [25]). The total length of the scaffolds was 2.32 Gb, which is close to two times the length of the PL4 genome (1.27 Gb) (Supplementary Table 3). We then compared the assembled genome of GS1 and the PL4 pseudomolecules by minimap2 [92] and found that many regions of the PL4 chromosomes correspond to two regions of the GS1 assembly (Supplementary Fig. 26). BUSCO analysis [14] against the GS1 assembly showed that 97% of the embryophyta genes were covered including 84.7% duplicated (Supplementary Table 10). These results confirm that the GS1 assembly is largely a diploid assembly covering most of the *F. esculentum* genome, suggesting that it should be sufficient to examine the two haplotypes of *S* locus.

We therefore searched for the *S* haplotype region in the GS1 assembly by performing BLASTN using a previously identified sequence of *S-ELF3* (AB641416.1 [81]) as query. Only one scaffold, scaffold 2156 contained *S-ELF3*, indicating that this scaffold contains the *S* haplotype region. We then searched for allelic sequences of scaffold 2156 by performing minimap2 with scaffold 2156 as query against the GS1 assembly. A large part of both ends of scaffolds 2156 and 21180 was highly similar to each other, although a region containing *S-ELF3* in the middle of scaffold 2156 was absent from scaffold 21180 (Extended Data Fig. 7a).

Collinearity at the gene level between scaffolds 2156 and 21180 was confirmed by mapping the CDS sequences of the predicted protein-coding genes of *F. esculentum* PL4 to the GS1 assembly using GMAP [93]. Forty PL4 genes were mapped to scaffold 2156 and most PL4 genes that mapped to 1.3-11.4 Mb of scaffold 2156 also mapped to scaffold 21180, indicating that these two scaffolds are allelic (Supplementary Table 32). Furthermore, a genomic

region of ~ 2.8 Mb around 6.5-9.3 Mb of scaffold 2156 contained genes specific to scaffold 2156, including *S-ELF3* (Extended Data Fig. 7b). Thus, we concluded that scaffold 2156 and scaffold 21180 contain the *S* haplotype and *s* haplotype regions, respectively, in which the *S* haplotype has a ~ 2.8 Mb hemizygous region. No genes expressed specifically in the flowers of pin plants were detected in the *s* haplotype region of scaffold 21180 by RNA-Seq analysis and the expression pattern of this region was similar between pin and thrum flowers (Supplementary Fig. 27). Although it is difficult to entirely rule out the possibility that there are unidentified genes with pin-specific functions in the *s* haplotype, these results suggest that the *s* haplotype does not contain any genes controlling the floral morphology and compatibility of pin plants. 89% of both scaffolds were comprised of TEs, of which *Gypsy*-type LTR retrotransposons made up of $\sim 80\%$. The TE proportion (92%) of the hemizygous region was similar to the rest of scaffold 2156 (88%) and scaffold 21180 (89%) (Supplementary Table 33).

5.3 Knockout of *S-ELF3* results in long homostyle plants

To determine the exact role of *S-ELF3* in heterostylous SI, we screened for a dysfunctional *S-ELF3* gene from the EMS-induced mutant population of 5,801 plants and identified a mutant with the splice acceptor site of the fourth intron disrupted (AG to AA) (Supplementary Fig. 28b). Abnormal splicing of *S-ELF3* was confirmed by electrophoresis of the RT-PCR products and subsequent Sanger sequencing (Supplementary Fig. 28b). M_1 plants displayed long homostyle flowers with long styles and long stamens, i.e., the same morphology as the flowers of PL4 (Supplementary Fig. 29a,b).

To minimize the possibility that the long homostyle phenotype of the knockout mutant is due to off-target effects of other mutations elsewhere in

the genome, we back-crossed the long homostyle mutant (*s-elf3-ψ2/s*) with pin plants (*s/s*) of *F. esculentum* cv. Harunoibuki. A long homostyle F₁ plant (*s-elf3-ψ2/s*) obtained by this cross was back-crossed with pin plants (*s/s*) and a long homostyle F₂ plant (*s-elf3-ψ2/s*) obtained was back-crossed again with four pin plants (*s/s*), which resulted in 13 long homostyle and 28 pin BC₃ plants. We confirmed by PCR that all long homostyle plants have the *s-elf3-ψ2* allele but the pin plants do not.

These results indicate that *S-ELF3* controls the style length, whereas genes controlling the stamen are likely present elsewhere in the hemizygous region. Seven other predicted protein-coding genes of *F. esculentum* PL4 mapped to the hemizygous region (Supplementary Table 32). One of these genes or other unidentified protein-coding genes with short ORFs or RNA genes in this region may be involved in the regulation of the male functions.

5.4 Long homostyle *S-ELF3* knockout mutants are self-compatible

To test the mating compatibility of the knockout mutants with long homostyle flowers, we self-pollinated the BC₃ long homostyle plants obtained by back-cross using wild type pin plants as recurrent parents. 56 BC₃S₁ seeds were obtained, which yielded 37 long homostyle and 19 pin plants. Self-pollination of these 37 long homostyle plants yielded 10.54 seeds on average per plant, whereas self-pollination of the pin plants yielded only 0.53 seeds on average per plant ($p=8.23 \times 10^{-9}$, two-sided permutation test using the perm.test function of the R exactRankTests package, <https://cran.r-project.org/web/packages/exactRankTests/>). This demonstrates that the *S-ELF3* knockout mutants exhibiting long homostyle flowers are self-compatible (Supplementary Fig. 29a).

We then investigated the cross compatibility of the long homostyle mutants with wild type pin and thrum plants to determine whether the self-fertility of these mutants is due to the conversion of the pistil incompatibility trait from the thrum to pin type as observed in long homostyle mutants of *Primula* [82]. We first hand-pollinated the pollen of BC₃ long homostyle mutants to the styles of *F. esculentum* cv. Harunoibuki pin and thrum plants. Results showed that the pollen of long homostyle mutants is compatible with the style of pin plants but incompatible with the style of thrum plants (Supplementary Fig. 29b). This is expected considering that the pollen of long homostyle mutants is of the thrum-type and that the knockout of *S-ELF3* should only affect the style. We next hand-pollinated the pollen of *F. esculentum* cv. Harunoibuki pin and thrum plants to the styles of BC₃ long homostyle mutants. Results showed that the style of long homostyle mutants is compatible with the pollen of thrum plants but incompatible with the pollen of pin plants (Supplementary Fig. 29c). Furthermore, we confirmed that the growth of pin pollen tubes was arrested in the long styles of the mutant, whereas the thrum pollen tubes reached the ovule, indicating that the long pistils of the long homostyle mutants exhibit the incompatibility trait of pin plants (Extended Data Fig. 8b). These results demonstrate that the knockout of *S-ELF3* did not abolish the incompatibility trait of the thrum style such that it is compatible with both pin-type and thrum-type pollen. Instead, it resulted in the conversion of a thrum-type short style to a pin-type long style. Thus, the long homostyle mutants are self-compatible, i.e., are able to self-fertilize, because they have pin-type styles and thrum-type stamens (Supplementary Fig. 29d). This cross compatibility behavior is consistent with that of *F. esculentum* PL4 [1] and the long homostyle mutants of *Primula* [82].

5.5 Analyses of *F. esculentum* PL4 and *F.*

homotropicum genomes

We then determined regions homologous to the *S* and *s* haplotype in *F. esculentum* PL4 and *F. homotropicum* based on the results of minimap2 between *F. esculentum* PL4 and GS1 and the gene-based collinearity detected by MCScanX between *F. esculentum* PL4 and *F. homotropicum*. While *S-ELF3* and other neighboring genes on the *S* haplotype of scaffold 2156 were homologous to genes on Chr 1 of *F. esculentum* PL4, genes further up- and downstream of scaffold 2156 and genes on scaffold 21180 were homologous to Chr 6 of *F. esculentum* PL4, indicating a translocation involving the *S* haplotype region between Chr 6 and Chr 1 in PL4 (Extended Data Fig. 7b). This is consistent with the comparison of PL4 pseudomolecules and the previously developed linkage map (Extended Data Fig. 1). Considering that the alleles of a few genes on PL4 Chr 1 that mapped to the region flanking the hemizygous region on scaffold 2156 were present on scaffold 21180, translocation from Chr 6 to Chr 1 seems to be the more parsimonious scenario (Extended Data Fig. 7c).

Gene-based collinearity between *F. esculentum* PL4 and *F. homotropicum* detected by MCScanX showed that the *s* haplotype homologous region of PL4 Chr 6 is collinear to Chr 6 of *F. homotropicum*, whereas the *S* haplotype homologous region of PL4 Chr 1 is collinear to Chr 1 of *F. homotropicum* (Supplementary Fig. 30). This region collinear to *S* haplotype on PL4 Chr 1 was within a 31 Mb region whose sequence is almost identical to *F. homotropicum* (Supplementary Fig. 30). Moreover, homologues of *S-ELF3* in this region of *F. homotropicum* and PL4 both had identical frameshift mutations (Supplementary Fig. 28a). Thus, *F. esculentum* PL4 has inherited the *S* haplotype homologous region from *F. homotropicum* in which *S-ELF3* is disrupted, i.e.,

the S^h haplotype. These results indicate that the translocation involving the S haplotype region did not occur during the tissue-culture mediated breeding procedure in which PL4 was established, but instead occurred in nature.

To further confirm that the S^h haplotype in *F. esculentum* PL4 is independent of S locus and that it confers the long homostyle phenotype, we conducted allelism test crosses. F_1 thrum plants were obtained by a cross between *F. esculentum* PL4 and thrum plants of *F. esculentum* cv. Harunoibuki (S/s). These F_1 thrum plants were back-crossed with pin plants of Harunoibuki (s/s) and the flower morphology of the descendants was observed. If the S^h and S haplotypes are allelic and of a single locus, the segregation ration in BC_1F_1 should be “thrum:long homostyle:pin = 1:1:0”. However, the observed segregation ratio was “thrum:long homostyle:pin = 44:18:20 ($\approx 2:1:1$, $p = 0.765$; chi-square test, $df = 2$)”, indicating that the S^h haplotype is independent of the S haplotype (Supplementary Fig. 31). These results are also consistent with the S^h haplotype being recessive to the S haplotype but dominant to the s haplotype and confirm that the S^h haplotype is responsible for the long homostyle phenotype of *F. esculentum* PL4.

5.6 Evolutionary origin of *S-ELF3*

Many *Fagopyrum* species other than *F. esculentum* exhibit heterostylous SI [41]. It was previously shown that *S-ELF3* is likely associated with heterostylous SI in *F. cymosum*, a close relative of *F. tataricum*, and *F. urophyllum*, a distantly related species that is sister to *F. esculentum* and *F. tataricum* [81]. Thus, heterostyly and the role of *S-ELF3* in heterostylous SI likely evolved in the common ancestor of most, if not all *Fagopyrum* species. Although

heterostyly is observed in species of other genera in Polygonaceae such as *Polygonum* [94] and *Persicaria* [95], it is not clear whether they share a common origin or evolved independently.

It was also previously reported that *S-ELF3* is a homologue of early flowering 3 (*ELF3*) of *A. thaliana* and that *F. esculentum* contains another homologue that is more closely related to *ELF3* of *A. thaliana* compared to *S-ELF3* [81]. Here, we found that *F. esculentum* PL4 contains two additional paralogs of *S-ELF3* that were duplicated by the Polygonaceae WGD (Main Fig. 2d), although it is not clear whether it corresponds to the younger or older WGD. No orthologs of *S-ELF3* or the other two paralogs were identified by OrthoFinder in the other Caryophyllales species used for gene family analyses including *R. hastatulus*, another Polygonaceae species.

We then performed molecular evolutionary analysis using the *S-ELF3* genes previously identified in *F. esculentum*, *F. cymosum*, *F. urophyllum*, and the paralogs identified in this study. Phylogenetic analysis suggested that *S-ELF3* likely originated by duplication after the WGD, but before the divergence of *F. esculentum* and *F. urophyllum*, although we cannot determine whether this was a WGD or a single-gene duplication (Extended Data Fig. 9). The K_s between *S-ELF3* and its closest paralog in *F. esculentum* (FesPL4.r1.1-Chr4.g236220.1) was ~ 0.64 - 0.67 , which is similar to the peak of orthologs between *F. esculentum* and *R. hastatulus* (~ 0.7) (Extended Data Fig. 5), and considerably larger than the K_s between *S-ELF3* of *F. esculentum* and *F. urophyllum* (0.36) (Extended Data Fig. 9). Thus, although orthologs of *S-ELF3* could not be identified in *R. hastatulus*, the timing of the duplication should be close to the divergence of *Fagopyrum* and *Rumex*, a genus that likely diverged from *Fagopyrum* ~ 60 mya [55] and contains hermaphroditic, polygamous, gynodioecious, monoecious, and dioecious species [96]. It should

therefore be worth investigating the role of this duplication in the evolution of heterostyly and its timing in relation to the divergence of other Polygonaceae genera containing heterostylous species.

5.7 Genetic architecture of heterostyly and its breakdown

Based on our findings, the genetic architecture of heterostyly in *F. esculentum* can be summarized as follows (see also Main Fig. 4). First, thrum plants have a ~2.8 Mb hemizygous region containing a locus required for the control of heterostylous SI (Main Fig. 4 and Extended Data Fig. 7). Plants that have this locus as a single copy, i.e., hemizygous, exhibit thrum flowers with short styles and long stamens, whereas plants that do not have this locus, i.e., null homozygous, exhibit pin flowers with long styles and short stamens. This locus contains a protein-coding gene *S-ELF3*, which has a dual role in regulating the style length and female compatibility. Additional gene(s) controlling the male morphology and compatibility may be present within the hemizygous region that we have not identified yet. We note that other genes outside of the *S* locus, such as the polygalacturonase protein (FePG1) that is expressed specifically in short styles (i.e., thrum-specific), may also be involved in heterostylous SI by functioning downstream of the *S* locus [97].

Plants with an intact *S-ELF3* gene therefore exhibit short styles that are compatible with short stamens of pin flowers, whereas plants with putative male gene(s) exhibit long stamens that are compatible with long styles of pin flowers. Thus, plants with both an intact *S-ELF3* and male gene(s) exhibit short styles and long stamens, whereas plants with neither an intact *S-ELF3* or male gene(s) exhibit long styles and short stamens. On the other hand, plants that do not have an intact *S-ELF3* but have the male gene(s) exhibit long styles of pin flowers and long stamens of thrum flowers, i.e., self-compatible

long homostyle flowers as observed in *F. esculentum* PL4 and the EMS mutant with a nonsense mutation in *S-ELF3*. We note that *F. esculentum* PL4 and the EMS mutant both show the same phenotype even though the region containing the putative male gene(s) is homozygous on Chr 1 of *F. esculentum* PL4 and hemizygous on Chr 6 of the EMS mutant of *F. esculentum* GS1.

Our findings highlight interesting parallels with plant genera in which heterostyly evolved independently. As previously reported, the *S* locus is hemizygous in *Primula* [86–88], *Turnera* [89, 90], and *Linum* [9], although the hemizygous region in *F. esculentum* is ~2.8 Mb and much larger than that of *P. vulgaris* (~280 kb) [86], *T. subulata* (~250 kb) [89], and *L. tenue* (~260 kb) [9]. In addition, similar to the dual role of *S-ELF3* in *F. esculentum*, the style length and female compatibility have been suggested to be jointly regulated by a single gene in both *Primula* (*CYP734A50*, also referred as *CYP^T*) [86, 87] and *Turnera* (*TsBAHD*) [90, 98]. These results show that nonhomologous genes have been recruited to perform similar dual functions in three independently evolved heterostylous systems. Moreover, a single nonsense mutation in this gene appears to be a common pathway to the breakdown of heterostyly to self-compatible long homostyle plants in *Fagopyrum* and *Primula* [82]. It is also interesting to note that *CYP734A50* in *Primula* [99, 100] and *S-ELF3* in *Fagopyrum* both originated by duplication of a gene outside the *S* locus before the divergence of the different heterostylous species in each genera. A gene likely regulating the anther position in *Primula* (*GLOBOSA2*, also referred to as *GLO^T*) also originated by duplication of a gene outside the *S* locus [86, 99, 100]. It is therefore tempting to speculate that the acquisition of the dual female functions in these genes following gene duplication may have facilitated the evolution of heterostylus SI in both groups of species.

6 Population genomic analyses and the domestication of common buckwheat

6.1 Background

Although the geographical origin and the wild progenitor of cultivated common buckwheat remained unknown until the 1980s, numerous sampling efforts and genetic analyses by Ohnishi and colleagues established that a wild species (*F. esculentum* ssp. *ancestrale*) currently distributed in southwest China, i.e., Yunnan and Sichuan provinces and east Tibet, is the wild progenitor [101, 102]. Nevertheless, questions remain regarding the process of domestication including the exact timing and geographical location. Genetic studies based on AFLP, SSR, and allozyme markers have shown that wild populations from around the border of Yunnan and Sichuan are most closely related to the cultivated population [103–105]. On the other hand, analysis of archaeobotanical data in China has indicated that the earliest archaeobotanical evidence of common buckwheat cultivation may come from the north and northwest of China and thus outside the current distribution of the wild progenitor species [106]. Limitations of previous genetic studies such as (i) failure to sample the cultivated taxon from outside southwest China, and (ii) use of genetic markers that are now outdated have been cited as possible reasons that may explain this potential discrepancy between the genetic and archaeobotanical data [106]. Thus, whole-genome sequencing of cultivated and wild accessions utilizing the pseudomolecules constructed in this study as a reference genome should allow us to better understand the domestication process of common buckwheat.

6.2 Resequencing and identifying single nucleotide polymorphisms

Here, we sequenced the genomes of 47 wild and 57 cultivated accessions. All wild accessions were sampled from sites in southwest China around Yunnan, Sichuan, and Tibet provinces covering the distribution range of the natural population of this species. The cultivated accessions were mainly from southwest China where the wild species is distributed, and also some from neighboring countries such as Nepal, India, and Bhutan (Supplementary Fig. 32 and Supplementary Table 34). We obtained on average 61 million 150 bp Illumina paired-end read pairs per sample. After cleaning the raw reads with Trimmomatic [107], the remaining paired reads were mapped to the *F. esculentum* PL4 genome assembly using BWA [108]. On average, 89% of the read pairs were mapped as properly paired, resulting in 86% of the genome being covered on average per sample. We then performed variant calling using bcftools [109] and identified single nucleotide polymorphisms (SNPs). After filtering regions identified as Transposable Elements or simple repeats and sites with ambiguous calls (see Methods), we identified ~ 47 million called sites on average per sample with an average depth of $\times 10.84$ that were used for further population genetic analyses (Supplementary Table 35).

The proportion of called sites was higher in some of the distal ends of the chromosome which are gene-rich and TE-poor (Supplementary Fig. 33). We note that while the *F. esculentum* PL4 genome assembly contains a few regions derived from *F. homotropicum*, the proportion of called sites in these regions was similar to those of the surrounding regions, suggesting that variant calling can also be performed for these regions. Thus, while the linkage around this region may be disrupted, other analyses should not be affected by including these regions. We therefore excluded these regions when performing phasing

and estimating the decay of linkage disequilibrium (LD), but included them in the remaining analyses. In total, we obtained 858,604 single nucleotide polymorphisms (SNPs) of high confidence with a minor allele frequency (MAF) of $\geq 5\%$ in all accessions, 615,106 SNPs with $>5\%$ MAF in all cultivated accessions, and 821,638 SNPs with $>5\%$ MAF in all wild accessions (Supplementary Table 36). Note that as common buckwheat is an outcrossing species, heterozygous calls were retained, meaning that the actual number of haploid genomes analyzed is twice the number of accessions.

6.3 Genetic relationship between the samples and pattern of polymorphism

We then examined the genetic relationship between the different accessions. We used 49,526 unlinked SNPs to perform principal component analysis (PCA) [110, 111], ADMIXTURE [112], and phylogenetic analysis using the Neighbor-Joining method [113]. Each result showed that the cultivated accessions are separated from the wild accessions, and that the wild accessions form four different clusters which roughly correspond to their geographic locations, which we refer to as Wild Shg, Wild Sic, Wild Tib, and Wild Yun (Main Fig. 5a-d and Supplementary Figs. 34 - 36). In particular, the Wild Tib population close to the border of Yunnan, Sichuan, and Tibet (referred to as southeast Tibet) was most closely related to the cultivated accessions, confirming previous results that were based on AFLP, SSR, and allozyme markers [103–105] using genome-wide SNP data with a much wider sampling of cultivated accessions. These results address concerns that were raised previously [106] and suggest that the cited discrepancy between the genetic and archaeobotanical data is unlikely to be due to limitations of genetic data.

All the cultivated accessions, which include accessions from various regions of Yunnan, Sichuan, Tibet, India, Nepal, and Bhutan clustered together and were distinguished from the southeast Tibet wild population by each analysis. These results suggest that the close relationship between the southeast Tibet wild population and the cultivated population cannot be explained simply by recent gene flow or some cultivated accessions having escaped to the wild in the recent past, which should lead to the southeast Tibet wild population clustering together with specific cultivated accessions close to southeast Tibet. The cultivated accessions of Nepal were separated from the remaining cultivated accessions at $K=6$ of ADMIXTURE (Supplementary Fig. 35).

We estimated various summary statistics to better understand the population structure of the cultivated and wild accessions. The nucleotide diversity π [114] of the cultivated population was lower than the wild population when all the wild accessions were grouped together, but slightly higher than the Wild Tib population and similar to the Wild Sic population (Supplementary Table 36). F_{ST} [115] between the cultivated population and the Wild Tib population was much lower than the F_{ST} between the cultivated population and the other wild populations (Supplementary Table 38), consistent with the results of PCA, ADMIXTURE, and phylogenetic analyses.

Minor allele frequency spectrum of synonymous and nonsynonymous mutations of each population showed a typical heavy-tailed distribution, with nonsynonymous mutations showing a higher proportion of singletons compared to those of synonymous mutations, as expected (Supplementary Fig. 37). We also examined the pattern of the decay of linkage disequilibrium against the physical distance in each population (Supplementary Fig. 38). The linkage measured by r^2 [116] of the cultivated population between sites within ~ 25 kb was higher than the Wild Shg and Wild Yun population but slightly lower

than the Wild Sic population and much lower than the Wild Tib population. Linkage equilibrium was reached by ~ 50 kb in each population.

We then examined the spatial pattern of polymorphisms across each chromosome. The nucleotide diversity π of each population tended to be lower around the pericentromeric regions compared to the distal regions of each chromosome (Main Fig. 2, Supplementary Figs. 39 and 40). By contrast, F_{ST} between each population was high around the pericentromeric regions and low in the distal regions of most chromosomes (Main Fig. 2 and Supplementary Figs. 41 - 43). These suggest that the differentiation between each population is high around the pericentromeric regions although the diversity within each population is low. As a result, when π is calculated for all wild accessions as one population, π around the pericentromeric region ends up being much higher than π of each wild population. We also calculated the genetic divergence d_{xy} between the cultivated and each wild population across each chromosome. The pattern was similar to that of F_{ST} and the divergence between the cultivated and the Wild Tib population was lower than between the cultivated and the other wild populations across each chromosome as expected (Supplementary Fig. 44).

6.4 Targets of artificial selection

We next searched for signatures of selection in the cultivated population. Population genetics theory predicts that selection during domestication should leave signatures around the target genomic region such as reduced nucleotide diversity, change in allele frequency, or extended haplotype homozygosity compared to the wild population [117]. On the other hand, factors including the population structure can also produce similar misleading patterns of polymorphisms. Indeed, while π of the cultivated population is much lower than that of the

entire wild population around some pericentromeric regions, this is likely to be due to high differentiation between the wild populations rather than selective sweep in the cultivated population (Main Fig. 2a, Supplementary Figs. 42 and 43).

In particular, we must consider that the cultivated population is much more closely related to the Wild Tib population compared to the other wild populations. For instance, comparing the cultivated population against the entire wild population would fail to take into account the genetic distance between the Wild Tib population and the other wild populations. This may result in false positives where the cultivated population is highly differentiated from the other wild populations but similar to the Wild Tib population, as observed around some of the pericentromeric regions (Main Fig. 2a, Supplementary Figs. 39 and 41). On the other hand, comparing the cultivated population only against the Wild Tib population cannot distinguish between changes in allele frequencies that occurred in the cultivated population and in the Wild Tib population.

We therefore used Population Branch Statistic (PBS) as a summary statistic to detect signatures of selection within the cultivated population. PBS uses an F_{ST} -based distance between three populations to estimate the branch length leading to the population of interest [118]. Regions showing exceptionally large PBS values, which indicate much longer branches leading to the population of interest compared to other genomic regions, can be interpreted as candidate target regions of selection. We first calculated the F_{ST} -based distance between the cultivated population, Wild Tib population, and the remaining wild population for each genomic region, and then estimated PBS for the branch leading to the cultivated population. This method allows us to

detect regions showing high differentiation between the cultivated and Wild Tib population due to changes in allele frequency in the cultivated population.

Clustering of large PBS values was observed around 50-60 Mb of Chr 1, indicating that this region contains an allele that was targeted by selection during the domestication process (Main Fig. 5e). This region in fact contained two peaks, one around 52-53 Mb and another around 56-57 Mb in which π of the cultivated population was much lower than that of the Wild Tib population and the remaining wild populations (Extended Data Fig. 10). Further inspection of a region within 52-53 Mb where almost no variation is present within the cultivated population revealed that many alleles present at low frequency in the Wild Tib population but absent from the other wild populations have reached fixation in the cultivated population (Main Fig. 5f).

Examining the phased dataset showed that a haplotype identical to the reference PL4 genome shared by all cultivated accessions is present at low frequency in the Wild Tib population but not in the other wild populations. We note that if the length of this haplotype in the Wild Tib population is longer than or of similar length to the cultivated population, a recent introgression of the haplotype from the cultivated population to the Wild Tib population may be inferred. Instead, the length of haplotype homozygosity is much longer in the cultivated population (Extended Data Fig. 10). Thus, the most likely scenario to explain the pattern of polymorphisms in this region should be a selective sweep in the cultivated population of a haplotype that originated in the ancestral Wild Tib population. It is also worth pointing out that many sites of the Wild Tib population in this genomic region contained alleles that are shared with the other wild populations but not present in the cultivated population, consistent with the Wild Tib population being a genuine wild population. These results, together with the close genetic relationship of the Wild

Tib population and the cultivated population, suggest that the wild population around southeast Tibet made a significant contribution to cultivated common buckwheat.

We report the top 10 regions showing the highest PBS values together with the predicted protein-coding genes in each region (Supplementary Table 39). Note that the p -values of each top 10 window following a permutation test were $p < 0.0001$ and that neighboring overlapping windows with $p < 0.0001$ were merged. The 500 kb region (52.3-52.8 Mb of Chr 1) showing the highest PBS value contained four predicted protein-coding genes (Supplementary Table 39). One of these genes, FesPL4.r1.1-Chr1.g091000.1, is a homologue of SNF1-related protein kinase 2 (*SnRK2*). SnRK2s play an important role in osmotic stress and abscisic acid signalling and some genes in *Arabidopsis thaliana* are essential for the control of seed development and dormancy [119, 120]. Nevertheless, the amino acid sequences of the SnRK2-like gene of *F. esculentum* PL4 is identical to the ortholog of *F. homotropicum* (Fho.r1.0.Chr1.g089790.1) and none of the SNPs identified overlapping this gene was nonsynonymous, although a few SNPs were present in the upstream region and introns. Thus, the target could be a mutation affecting the expression of this gene, or one of the other three genes in this 500 kb region or in a neighboring linked region. The functional annotation based on SnpEff of each SNP in this 500 kb region is shown in Supplementary Table 40. Further experiments will be required to determine the actual target of selection. Other regions with high PBS values may also contain targets of artificial selection, although careful inspection of the inferred candidate regions should be required as other factors such as the population structure and genetic linkage may also result in high PBS values. Nevertheless, this list should aid further efforts in identifying domestication-related genes.

References

- [1] Matsui, K., Tetsuka, T., Nishio, T. & Hara, T. Heteromorphic incompatibility retained in self-compatible plants produced by a cross between common and wild buckwheat. *New Phytol.* **159**, 701–708 (2003).
- [2] Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
- [3] Fierst, J. L. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front. Genet.* **6**, 220 (2015).
- [4] Tang, H. *et al.* ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).
- [5] Catchen, J., Amores, A. & Bassham, S. Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. *G3: Genes Genom. Genet.* **10**, 4115–4128 (2020).
- [6] Rastas, P. Lep-Anchor: automated construction of linkage map anchored haploid genomes. *Bioinformatics* **36**, 2359–2364 (2020).
- [7] Akagi, T. *et al.* Recurrent neo-sex chromosome evolution in kiwifruit. *Nat. Plants* **9**, 393–402 (2023).
- [8] Healey, A. L. *et al.* Newly identified sex chromosomes in the *Sphagnum* (peat moss) genome alter carbon sequestration and ecosystem dynamics. *Nat. Plants* **9**, 238–254 (2023).

- [9] Gutiérrez-Valencia, J. *et al.* Genomic analyses of the *Linum* distyly supergene reveal convergent evolution at the molecular level. *Curr. Biol.* **32**, 4360–4371 (2022).
- [10] Iwata, H. & Ninomiya, S. AntMap: constructing genetic linkage maps using an ant colony optimization algorithm. *Breed. Sci.* **56**, 371–377 (2006).
- [11] Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018).
- [12] Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
- [13] Nagano, M., Aii, J., Campbell, C., Kawasaki, S. & Adachi, T. Genome size analysis of the genus *Fagopyrum*. *Fagopyrum* **17**, 35–39 (2000).
- [14] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
- [15] Zhang, L. *et al.* The tartary buckwheat genome provides insights into rutin biosynthesis and abiotic stress tolerance. *Mol. Plant* **10**, 1224–1237 (2017).
- [16] Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).

- 1292 [17] Shearer, L. A. *et al.* Fluorescence *in situ* hybridization and optical map-
 1293 ping to correct scaffold arrangement in the tomato genome. *G3: Genes*
 1294 *Genom. Genet.* **4**, 1395–1405 (2014).
- 1295 [18] Rai, A. *et al.* Chromosome-level genome assembly of *Ophiorrhiza pumila*
 1296 reveals the evolution of camptothecin biosynthesis. *Nat. Commun.* **12**,
 1297 405 (2021).
- 1298 [19] Han, Y., Zhang, T., Thammapichai, P., Weng, Y. & Jiang, J.
 1299 Chromosome-specific painting in *Cucumis* species using bulked oligonu-
 1300 cleotides. *Genetics* **200**, 771–779 (2015).
- 1301 [20] Jiang, J. Fluorescence *in situ* hybridization in plants: recent devel-
 1302 opments and future applications. *Chromosome Res.* **27**, 153–165
 1303 (2019).
- 1304 [21] Yabe, S. *et al.* Rapid genotyping with DNA micro-arrays for high-density
 1305 linkage mapping and QTL mapping in common buckwheat (*Fagopyrum*
 1306 *esculentum* Moench). *Breed. Sci.* **64**, 291–299 (2014).
- 1307 [22] Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M.
 1308 BRAKER2: automatic eukaryotic genome annotation with GeneMark-
 1309 EP+ and AUGUSTUS supported by a protein database. *NAR Genom.*
 1310 *Bioinform.* **3**, lqaa108 (2021).
- 1311 [23] Kidwell, M. G. Transposable elements and the evolution of genome size
 1312 in eukaryotes. *Genetica* **115**, 49–63 (2002).
- 1313 [24] Yasui, Y. *et al.* Assembly of the draft genome of buckwheat and its
 1314 applications in identifying agronomically useful genes. *DNA Res.* **23**,
 1315 215–224 (2016).

- [25] Penin, A. A. *et al.* High-resolution transcriptome atlas and improved genome assembly of common buckwheat, *Fagopyrum esculentum*. *Front. Plant Sci.* **12**, 612382 (2021).
- [26] Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181–1184 (2014).
- [27] Badouin, H. *et al.* The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
- [28] Teh, B. T. *et al.* The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat. Genet.* **49**, 1633–1641 (2017).
- [29] Shi, J. *et al.* Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nat. Commun.* **10**, 464 (2019).
- [30] SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
- [31] Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
- [32] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
- [33] Neumann, P., Novák, P., Hošťáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their

- polyprotein domains and provides a reference for element classification.
Mob. DNA **10**, 1 (2019).
- [34] Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
- [35] Neumann, P. *et al.* Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA* **2**, 4 (2011).
- [36] Fawcett, J. A., Van de Peer, Y. & Maere, S. Significance and biological consequences of polyploidization in land plant evolution. *Plant Genome Diversity, Springer-Verlag Wien* **2**, 277–293 (2013).
- [37] Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
- [38] Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).
- [39] Yang, Y. *et al.* Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *New Phytol.* **217**, 855–870 (2018).
- [40] Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
- [41] Ohsako, T. & Li, C. Classification and systematics of the *Fagopyrum* species. *Breed. Sci.* **70**, 93–100 (2020).

- [42] Dohm, J. C. *et al.* The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* **505**, 546–549 (2014).
- [43] Verde, I. *et al.* The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–494 (2013).
- [44] Albert, V. A. *et al.* The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
- [45] Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
- [46] Rifkin, J. L. *et al.* Recombination landscape dimorphism and sex chromosome evolution in the dioecious plant *Rumex hastatulus*. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **377**, 20210226 (2022).
- [47] Palfalvi, G. *et al.* Genomes of the venus flytrap and close relatives unveil the roots of plant carnivory. *Curr. Biol.* **30**, 2312–2320 (2020).
- [48] Gilman, I. S., Moreno-Villena, J. J., Lewis, Z. R., Goolsby, E. W. & Edwards, E. J. Gene co-expression reveals the modularity and integration of C₄ and CAM in *Portulaca*. *Plant Physiol.* **189**, 735–753 (2022).
- [49] Jarvis, D. E. *et al.* The genome of *Chenopodium quinoa*. *Nature* **542**, 307–312 (2017).
- [50] Lightfoot, D. *et al.* Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biol.* **15**, 74

(2017).

- [51] Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).

- [52] Ng, K. K. S. *et al.* The genome of *Shorea leprosula* (Dipterocarpaceae) highlights the ecological relevance of drought in aseasonal tropical rainforests. *Commun. Biol.* **4**, 1166 (2021).

- [53] Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).

- [54] Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 437–453 (2015).

- [55] Yao, G. *et al.* Plastid phylogenomic insights into the evolution of Caryophyllales. *Mol. Phylogenet. Evol.* **134**, 74–86 (2019).

- [56] Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl. Acad. Sci. U S A* **106**, 5737–5742 (2009).

- [57] Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

- [58] Jiang, P. *et al.* Rutin and flavonoid contents in three buckwheat species *Fagopyrum esculentum*, *F. tataricum*, and *F. homotropicum* and their

- 1407 protective effects against lipid peroxidation. *Food Res. Int.* **40**, 356–364
 1408 (2007).
- 1409 [59] Jobling, S. Improving starch for food and industrial applications. *Curr.*
 1410 *Opin. Plant Biol.* **7**, 210–218 (2004).
- 1411 [60] Seung, D. Amylose in starch: Towards an understanding of biosynthesis,
 1412 structure and function. *New Phytol.* **228**, 1490–1504 (2020).
- 1413 [61] Sakamoto, S. *Glutinous-endosperm starch food culture specific to Eastern*
 1414 *and Southeastern Asia*, 215–231 (Oxford: Berg, 1996).
- 1415 [62] Hunt, H. V., Denyer, K., Packman, L. C., Jones, M. K. & Howe, C. J.
 1416 Molecular basis of the waxy endosperm starch phenotype in broomcorn
 1417 millet (*Panicum miliaceum* L.). *Mol. Biol. Evol.* **27**, 1478–1494 (2010).
- 1418 [63] Hunt, H. V. *et al.* Waxy phenotype evolution in the allotetraploid cereal
 1419 broomcorn millet: mutations at the *GBSSI* locus in their functional and
 1420 phylogenetic context. *Mol. Biol. Evol.* **30**, 109–122 (2013).
- 1421 [64] Gregori, M. & Kreft, I. Breakable starch granules in a low-amylose
 1422 buckwheat (*Fagopyrum esculentum* Moench) mutant. *J. Food Agric.*
 1423 *Environ* **10**, 258–262 (2012).
- 1424 [65] Kreft, I. *et al.* Breeding buckwheat for nutritional quality. *Breed. Sci.*
 1425 **70**, 67–73 (2020).
- 1426 [66] Tsai, H. *et al.* Discovery of rare mutations in populations: TILLING by
 1427 sequencing. *Plant Phys.* **156**, 1257–1268 (2011).
- 1428 [67] Momma, M. & Fujimoto, Z. Interdomain disulfide bridge in the rice
 1429 granule bound starch synthase I catalytic domain as elucidated by X-ray

- 1430 structure analysis. *Biosci. Biotechnol. Biochem.* **76**, 1591–1595 (2012).
- 1431 [68] Norbäck, D. & Wieslander, G. A review on epidemiological and clinical
1432 studies on buckwheat allergy. *Plants* **10**, 607 (2021).
- 1433 [69] Heffler, E., Pizzimenti, S., Badiu, I., Guida, G. & Rolla, G. Buckwheat
1434 allergy: an emerging clinical problem in Europe. *J. Allergy Ther.* (2014).
- 1435 [70] Satoh, R., Jensen-Jarolim, E. & Teshima, R. Understanding buckwheat
1436 allergies for the management of allergic reactions in humans and animals.
1437 *Breed. Sci.* **70**, 85–92 (2020).
- 1438 [71] Tanaka, K. *et al.* Pepsin-resistant 16-kD buckwheat protein is associated
1439 with immediate hypersensitivity reaction in patients with buckwheat
1440 allergy. *Int. Arch. Allergy Immunol.* **129**, 49–56 (2002).
- 1441 [72] Satoh, R., Koyano, S., Takagi, K., Nakamura, R. & Teshima, R. Identifi-
1442 cation of an IgE-binding epitope of a major buckwheat allergen, BWp16,
1443 by SPOTs assay and mimotope screening. *Int. Arch. Allergy Immunol.*
1444 **153**, 133–140 (2010).
- 1445 [73] Shewry, P. R., Napier, J. A. & Tatham, A. S. Seed storage proteins:
1446 structures and biosynthesis. *Plant Cell* **7**, 945 (1995).
- 1447 [74] Moreno, F. J. & Clemente, A. 2S albumin storage proteins: what makes
1448 them food allergens? *Open Biochem. J.* **2**, 16 (2008).
- 1449 [75] Radovic, R., Maksimovic, R., Brkljacic, M., Varkonji Gasic, I. & Savic,
1450 P. 2S albumin from buckwheat (*Fagopyrum esculentum* Moench) seeds.
1451 *J. Agric. Food Chem.* **47**, 1467–1470 (1999).

- [76] Chen, P., Guo, Y., Yan, Q. & Li, Y. Molecular cloning and characterization of Fag t 2: a 16-kDa major allergen from Tartary buckwheat seeds. *Allergy* **66**, 1393–1395 (2011).
- [77] Rico, M., Bruix, M., González, C., Monsalve, R. I. & Rodríguez, R. ^1H NMR assignment and global fold of napin BnIb, a representative 2S albumin seed protein. *Biochemistry* **35**, 15672–15682 (1996).
- [78] Satoh, R. *et al.* Immunological characterization and mutational analysis of the recombinant protein BWp16, a major allergen in buckwheat. *Biol. Pharm. Bull.* **31**, 1079–1085 (2008).
- [79] Barrett, S. C. ‘A most complex marriage arrangement’: recent advances on heterostyly and unresolved questions. *New Phytol.* **224**, 1051–1067 (2019).
- [80] Nishimoto, Y., Ohnishi, O. & Hasegawa, M. Topological incongruence between nuclear and chloroplast DNA trees suggesting hybridization in the urophyllum group of the genus *Fagopyrum* (Polygonaceae). *Genes Genet. Sys.* **78**, 139–153 (2003).
- [81] Yasui, Y. *et al.* *S-LOCUS EARLY FLOWERING 3* is exclusively present in the genomes of short-styled buckwheat plants that exhibit heteromorphic self-incompatibility. *PLoS One* **7**, e31264 (2012).
- [82] Huu, C. N., Plaschil, S., Himmelbach, A., Kappel, C. & Lenhard, M. Female self-incompatibility type in heterostylous *Primula* is determined by the brassinosteroid-inactivating cytochrome P450 CYP734A50. *Curr. Biol.* **32**, 671–676 (2022).

- [83] Matsui, K. & Yasui, Y. Buckwheat heteromorphic self-incompatibility: genetics, genomics and application to breeding. *Breed. Sci.* **70**, 32–38 (2020).
- [84] Pamela, V. & Dowrick, J. Heterostyly and homostyly in *Primula obconica*. *Heredity* **10**, 219–236 (1956).
- [85] Woo, S. H., Adachi, T., Jong, S. K. & Campbell, C. G. Inheritance of self-compatibility and flower morphology in an inter-specific buckwheat hybrid. *Can. J. Plant Sci.* **79**, 483–490 (1999).
- [86] Li, J. *et al.* Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nat. Plants* **2**, 16188 (2016).
- [87] Huu, C. N. *et al.* Presence versus absence of CYP734A50 underlies the style-length dimorphism in primroses. *eLife* **5**, e17956 (2016).
- [88] Cocker, J. M. *et al.* *Primula vulgaris* (primrose) genome assembly, annotation and gene expression, with comparative genomics on the heterostyly supergene. *Sci. Rep.* **8**, 17942 (2018).
- [89] Shore, J. S. *et al.* The long and short of the *S*-locus in *Turnera* (Passifloraceae). *New Phytol.* **224**, 1316–1329 (2019).
- [90] Matzke, C. M. *et al.* Pistil mating type and morphology are mediated by the brassinosteroid inactivating activity of the *S*-locus gene *BAHD* in heterostylous *Turnera* species. *Int. J. Mol. Sci.* **22**, 10603 (2021).
- [91] Yabe, S. *et al.* Potential of genomic selection in mass selection breeding of an allogamous crop: an empirical study to increase yield of common buckwheat. *Front. Plant Sci.* **9**, 276 (2018).

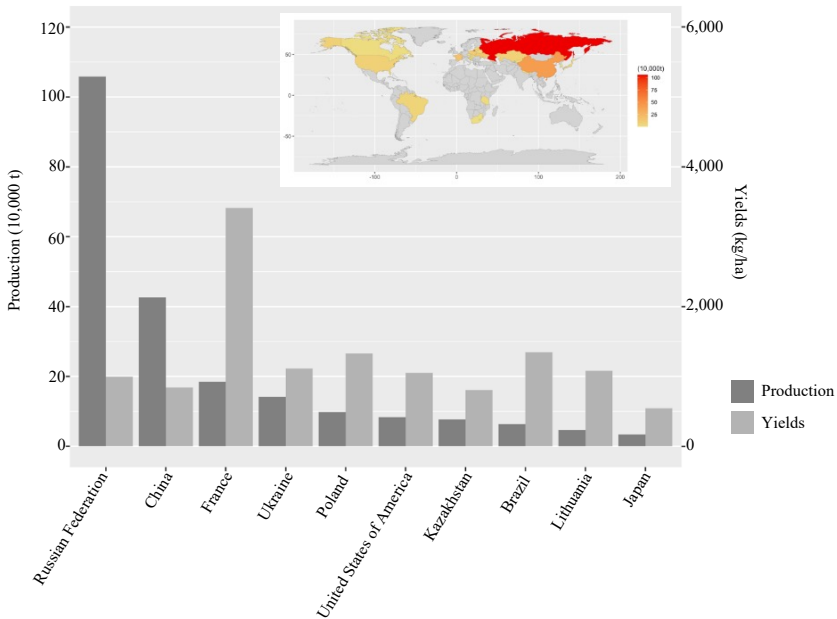
- 1498 [92] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioin-*
1499 *formatics* **34**, 3094–3100 (2018).
- 1500 [93] Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment
1501 program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875
1502 (2005).
- 1503 [94] Chen, M.-L., Qi, M.-Y., Bai, B.-B. & Han, X. Reproductive ecology of
1504 distylous shoreside *Polygonum criopolitanum* Hance. *Diversity* **14**, 222
1505 (2022).
- 1506 [95] Nishihira, J. & Washitani, I. Patterns and consequences of self-pollen
1507 deposition on stigmas in heterostylous *Persicaria japonica* (Polygo-
1508 naceae). *Am. J. Bot.* **85**, 352–359 (1998).
- 1509 [96] Navajas-Pérez, R. *et al.* The evolution of reproductive systems and sex-
1510 determining mechanisms within *Rumex* (Polygonaceae) inferred from
1511 nuclear and chloroplastidial sequence data. *Mol. Biol. Evol.* **22**, 1929–
1512 1939 (2005).
- 1513 [97] Takeshima, R., Nishio, T., Komatsu, S., Kurauchi, N. & Matsui, K.
1514 Identification of a gene encoding polygalacturonase expressed specifically
1515 in short styles in distylous common buckwheat (*Fagopyrum esculentum*).
1516 *Heredity* **123**, 492–502 (2019).
- 1517 [98] Matzke, C. M., Shore, J. S., Neff, M. M. & McCubbin, A. G. The *Turn-*
1518 *era* style *S*-locus gene *TsBAHD* possesses brassinosteroid-inactivating
1519 activity when expressed in *Arabidopsis thaliana*. *Plants* **9**, 1566 (2020).

- 1520 [99] Huu, C. N., Keller, B., Conti, E., Kappel, C. & Lenhard, M. Supergene
1521 evolution via stepwise duplications and neofunctionalization of a floral-
1522 organ identity gene. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 23148–23157
1523 (2020).
- 1524 [100] Potente, G. *et al.* Comparative genomics elucidates the origin of a super-
1525 gene controlling floral heteromorphism. *Mol. Biol. Evol.* **39**, msac035
1526 (2022).
- 1527 [101] Ohnishi, O. Discovery of the wild ancestor of common buckwheat.
1528 *Fagopyrum* **11**, 5–10 (1991).
- 1529 [102] Ohnishi, O. Search for the wild ancestor of buckwheat III. the wild
1530 ancestor of cultivated common buckwheat, and of tatar buckwheat.
1531 *Econ. Bot.* **52**, 123–133 (1998).
- 1532 [103] Konishi, T., Yasui, Y. & Ohnishi, O. Original birthplace of cultivated
1533 common buckwheat inferred from genetic relationships among cultivated
1534 populations and natural populations of wild common buckwheat revealed
1535 by AFLP analysis. *Genes Genet. Sys.* **80**, 113–119 (2005).
- 1536 [104] Konishi, T. & Ohnishi, O. Close genetic relationship between culti-
1537 vated and natural populations of common buckwheat in the Sanjiang
1538 area is not due to recent gene flow between them - an analysis using
1539 microsatellite markers. *Genes Genet Sys* **82**, 53–64 (2007).
- 1540 [105] Ohnishi, O. On the origin of cultivated common buckwheat based
1541 on allozyme analyses of cultivated and wild populations of common
1542 buckwheat. *Fagopyrum* **26**, 3–9 (2009).

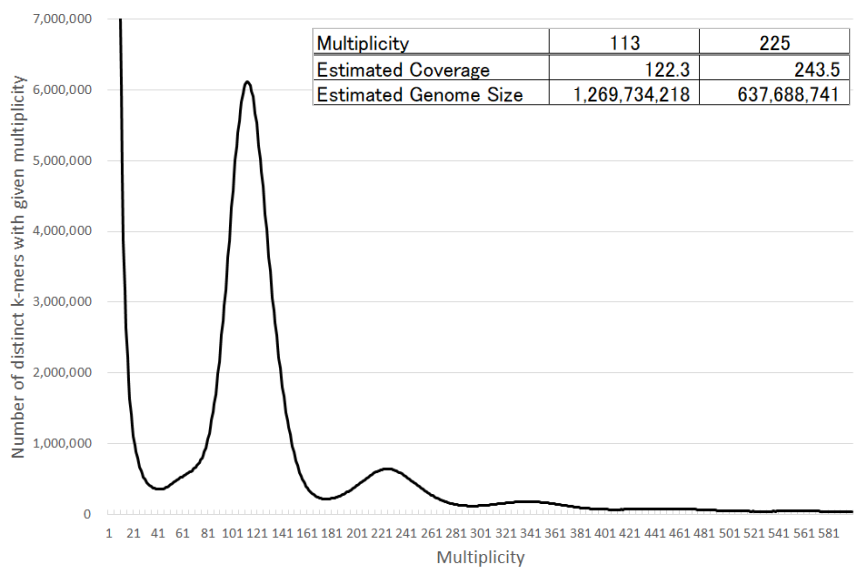
- [106] Hunt, H. V., Shang, X. & Jones, M. K. Buckwheat: a crop from outside
the major chinese domestication centres? a review of the archaeobotan-
ical, palynological and genetic evidence. *Veg. Hist. Archaeobotany* **27**,
493–506 (2018).
- [107] Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer
for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- [108] Li, H. & Durbin, R. Fast and accurate short read alignment with
Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- [109] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience*
10, gigab008 (2021).
- [110] Patterson, N., Price, A. L. & Reich, D. Population structure and
eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
- [111] Novembre, J. & Stephens, M. Interpreting principal component analyses
of spatial population genetic variation. *Nat. Genet.* **40**, 646–649 (2008).
- [112] Alexander, D., Novembre, J. & Lange, K. Fast model-based estimation
of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- [113] Saitou, N. & Nei, M. The neighbor-joining method: a new method for
reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
- [114] Tajima, F. Evolutionary relationship of DNA sequences in finite
populations. *Genetics* **105**, 437–460 (1983).
- [115] Weir, B. S. & Cockerham, C. C. Estimating F -statistics for the analysis
of population structure. *Evolution* **38**, 1358–1370 (1984).

- 1565 [116] Hill, W. & Robertson, A. Linkage disequilibrium in finite populations.
1566 *Theor. Appl. Genet.* **38**, 226–231 (1968).
- 1567 [117] Shi, J. & Lai, J. Patterns of genomic changes with crop domestication
1568 and breeding. *Curr. Opin. Plant Biol.* **24**, 47–53 (2015).
- 1569 [118] Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high
1570 altitude. *Science* **329**, 75–78 (2010).
- 1571 [119] Nakashima, K. *et al.* Three *Arabidopsis* SnRK2 protein kinases,
1572 SRK2D/SnRK2.2, SRK2E/SnRK2.6/OST1 and SRK2I/SnRK2.3,
1573 involved in ABA signaling are essential for the control of seed
1574 development and dormancy. *Plant Cell Physiol.* **50**, 1345–1363 (2009).
- 1575 [120] Shinozawa, A. *et al.* SnRK2 protein kinases represent an ancient system
1576 in plants for adaptation to a terrestrial environment. *Comm. Biol.* **2**, 30
1577 (2019).
- 1578 [121] Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient
1579 parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770
1580 (2011).
- 1581 [122] Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for
1582 phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534
1583 (2020).
- 1584 [123] Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Research* **51**,
1585 D418–D427 (2023).
- 1586 [124] Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L.
1587 Graph-based genome alignment and genotyping with HISAT2 and

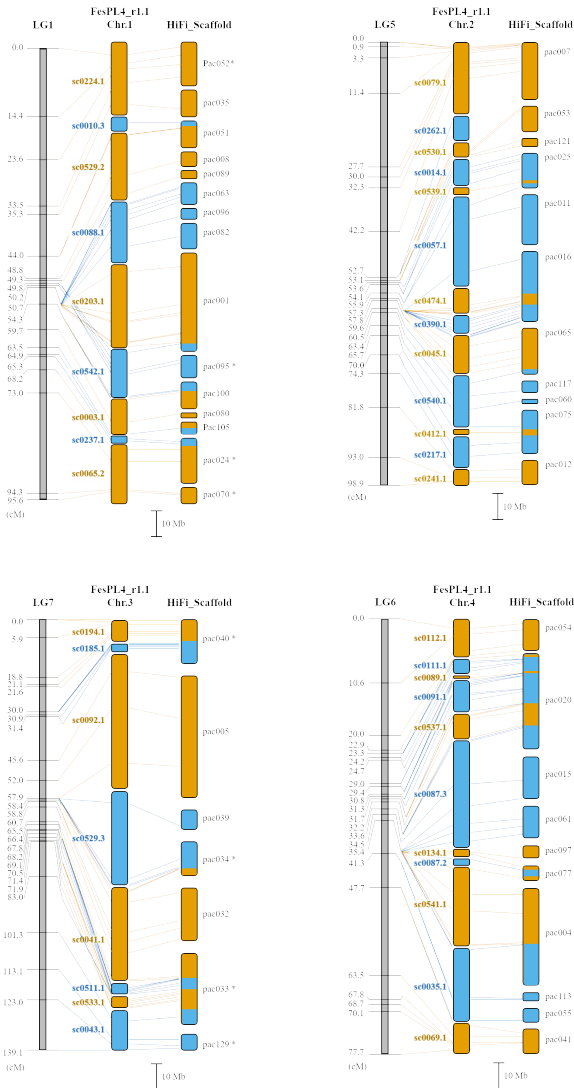
- 1588 HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
- 1589 [125] Pertea, M. *et al.* Stringtie enables improved reconstruction of a
1590 transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).



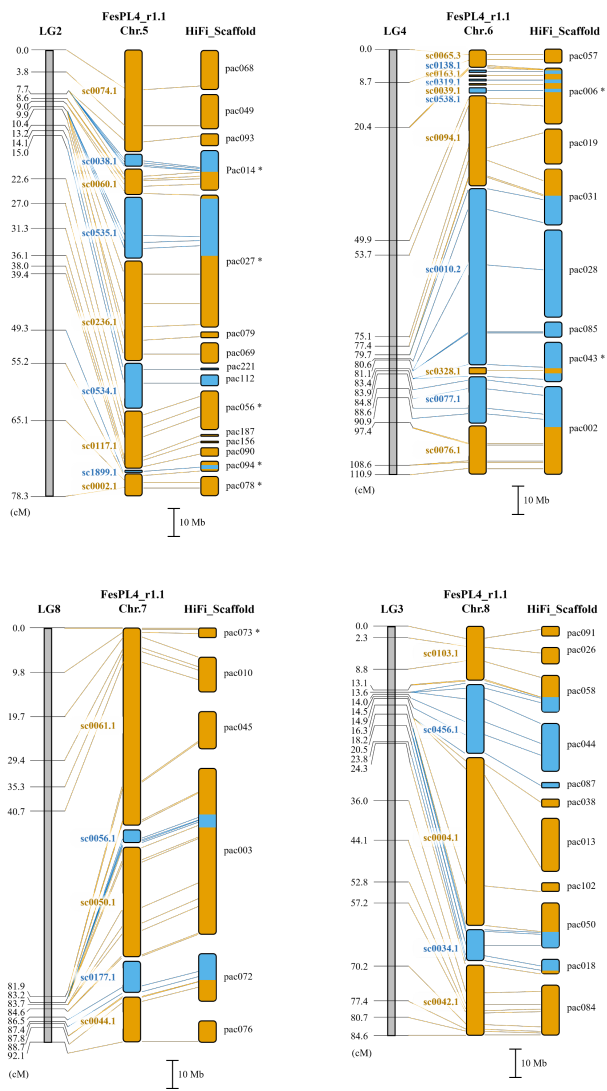
Supplementary Fig. 1 World production and yield of common buckwheat. The bar chart shows the world's top 10 producing countries of common buckwheat, arranged in order of decreasing total production from left to right. The total production and yield (i.e., production per unit area of land) are indicated by dark and light grey bars, respectively. The production and yield values are averages over 2015 to 2019 obtained from FAOSTAT (<http://www.fao.org/faostat/en/#home>). Values for France are averages over three years (2017 to 2019) as data for 2015 and 2016 were not available. Inset shows heat map of all producing countries of common buckwheat, coloured according to their production levels (Countries with no confirmed production are in gray).



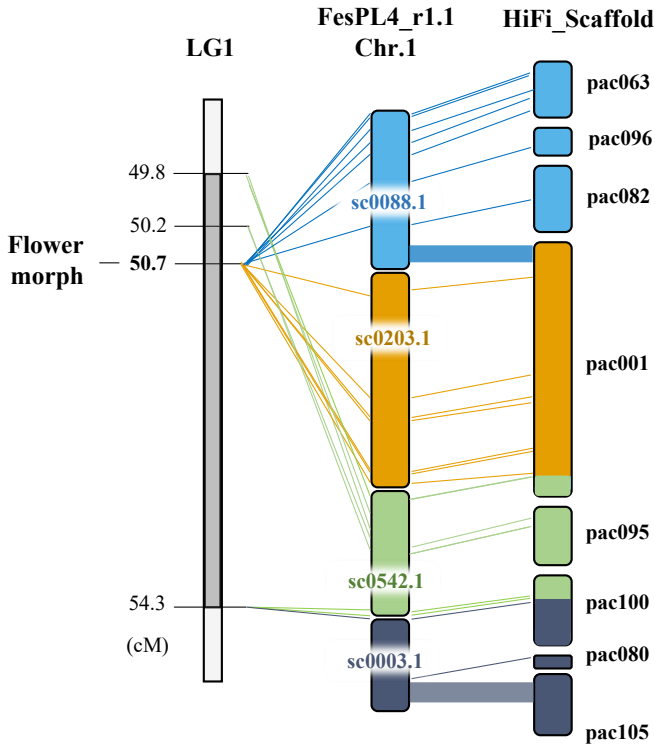
Supplementary Fig. 2 *k*-mer frequency distribution of *F. esculentum* PL4. The distribution of *k*-mer values was obtained from the paired-end short reads. The genome size and coverage were estimated by Jellyfish [121] using the peak value of the *k*-mer distribution corresponding to a multiplicity of 113.



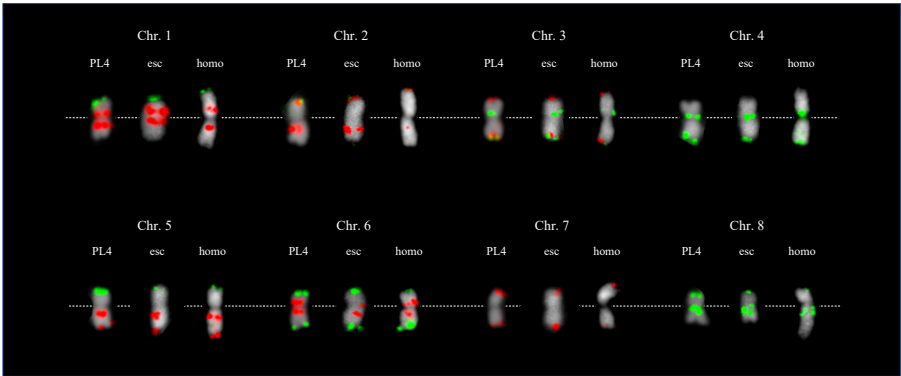
Supplementary Fig. 3 Correspondence between the linkage maps, pseudomolecules, and HiFi contigs of *F. esculentum* PL4 chromosomes 1 to 4. The left bar shows the linkage map of *F. esculentum* chromosomes constructed using AmpliSeq markers. The map positions are indicated on the left of the bar. Pseudomolecules based on the NRG scaffolds are shown in the middle and HiFi contigs are shown in the right. The colours of each NRG scaffold are alternated and the HiFi contigs are coloured to match the corresponding NRG scaffolds. Asterisks indicate scaffolds in reverse orientation.



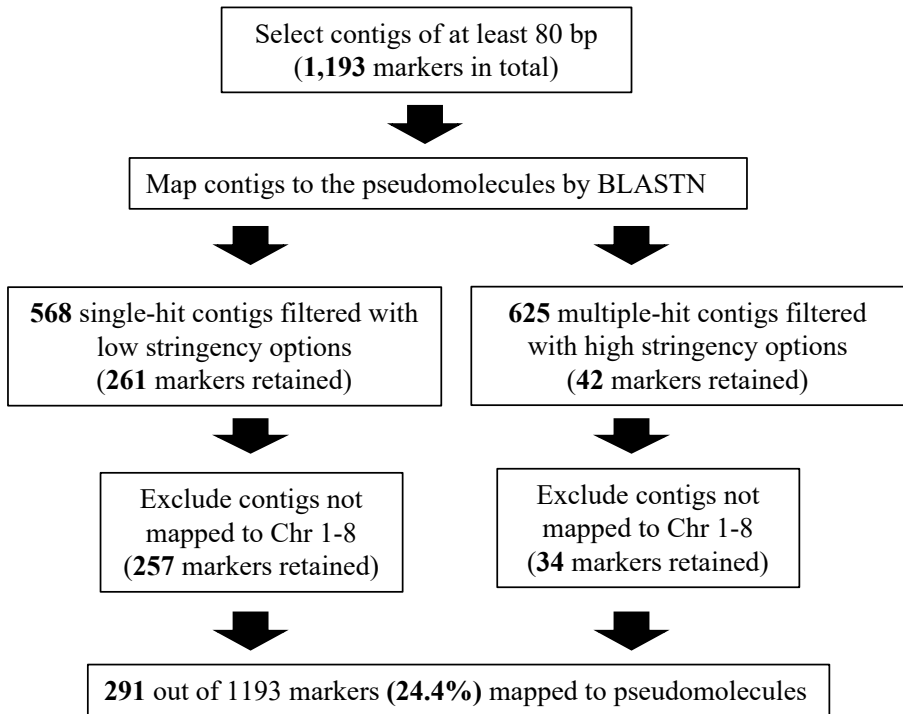
Supplementary Fig. 4 Correspondence between the linkage maps, pseudo-molecules, and HiFi contigs of *F. esculentum* PL4 chromosomes 5 to 8. The left bar shows the linkage map of *F. esculentum* chromosomes constructed using AmpliSeq markers. The map positions are indicated on the left of the bar. Pseudomolecules based on the NRG scaffolds are shown in the middle and HiFi contigs are shown in the right. The colours of each NRG scaffolds are alternated and the HiFi contigs are coloured to match the corresponding NRG scaffolds. Asterisks indicate scaffolds in reverse orientation.



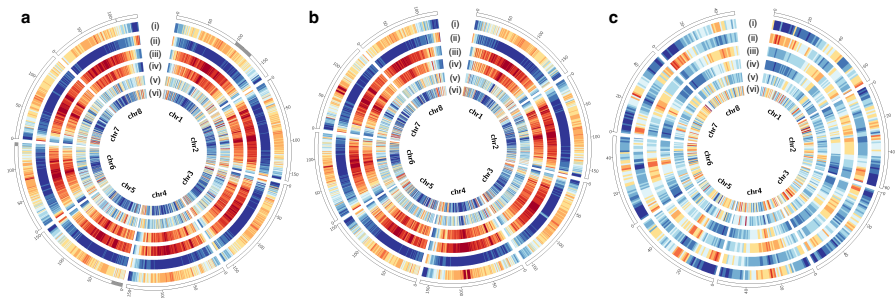
Supplementary Fig. 5 Example of a region on Chr 1 showing how NRG scaffolds were anchored to linkage groups by utilizing HiFi contigs. FesPL4_r1.1 indicates the pseudomolecule constructed by assembling the NRG scaffolds (e.g. sc0088.1). AmpliSeq markers mapped to NRG and HiFi contigs are connected by thin coloured lines. Regions with >99% similarity between the NRG scaffolds and HiFi contigs that were used to determine the orientation of the NRG scaffolds are connected by thick coloured lines. Note the discrepancy between the linkage order and physical order of the AmpliSeq markers on sc0542.1. The physical order was judged to be more likely because the physical order was consistent between the NRG and HiFi contigs, whereas the bootstrap support of the linkage order was low (Supplementary Table 6). See also Supplementary Fig. 3.



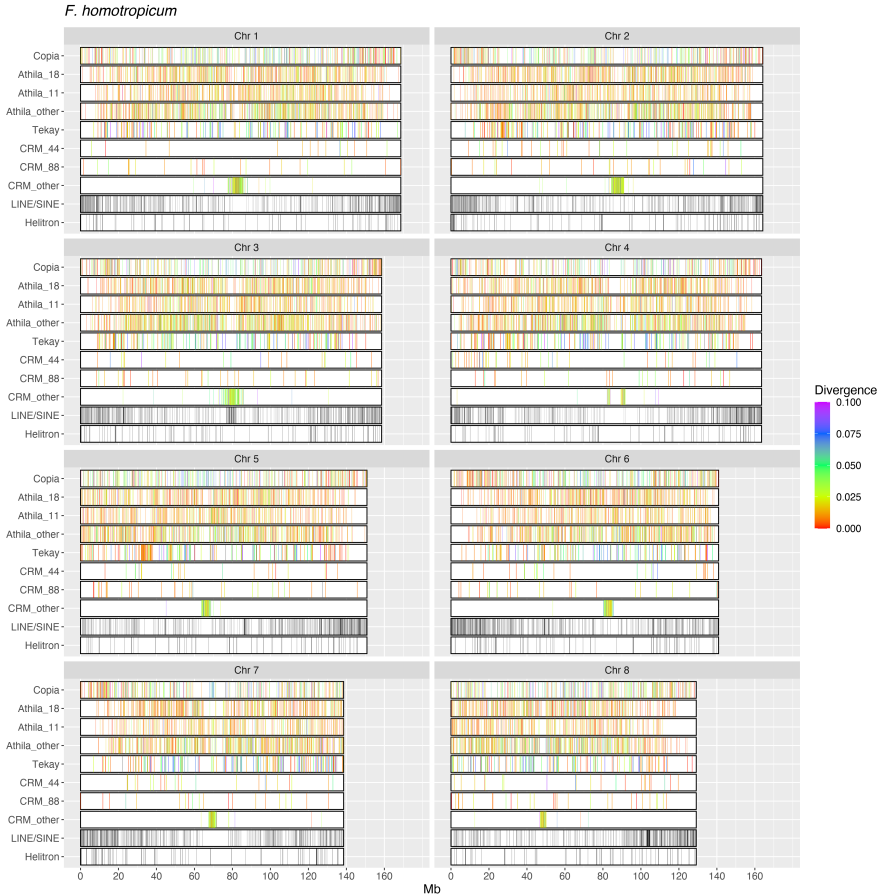
Supplementary Fig. 6 Comparative karyotyping using oligo-FISH probes. PL4, esc, and homo correspond to chromosomes of *F. esculentum* PL4, *F. esculentum* cv. Harunoibuki, and *F. homotropicum*, respectively.



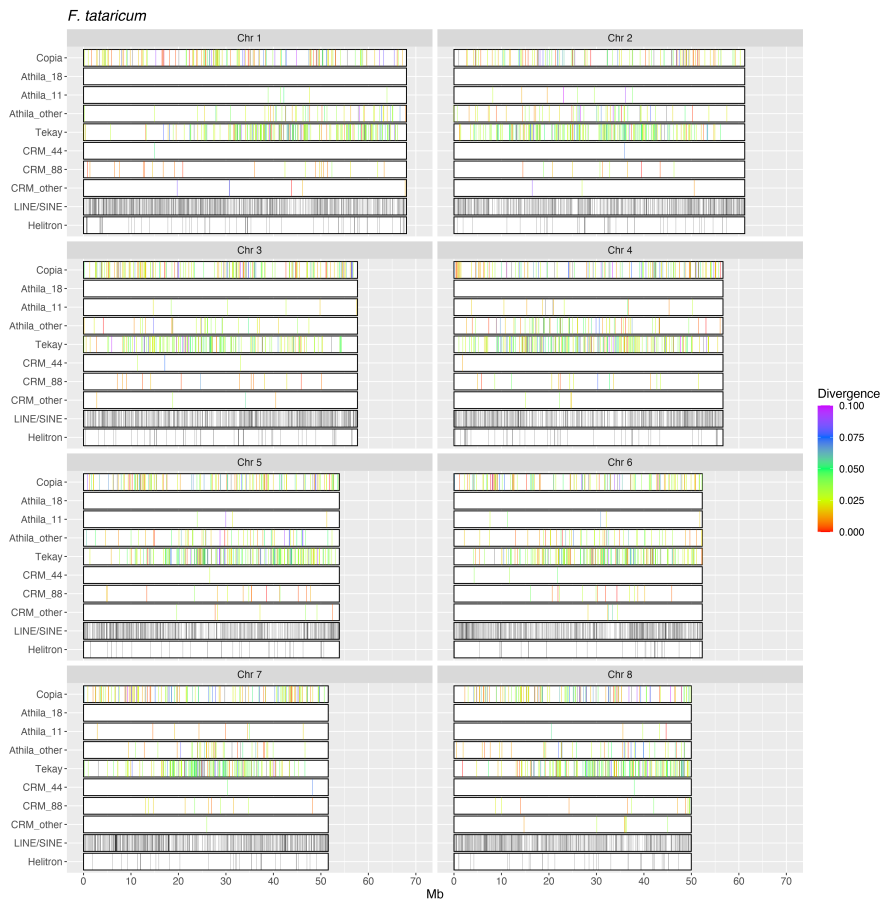
Supplementary Fig. 7 Procedure for mapping previously published linkage markers to *F. esculentum* PL4 pseudomolecules. Linkage markers are from [21].



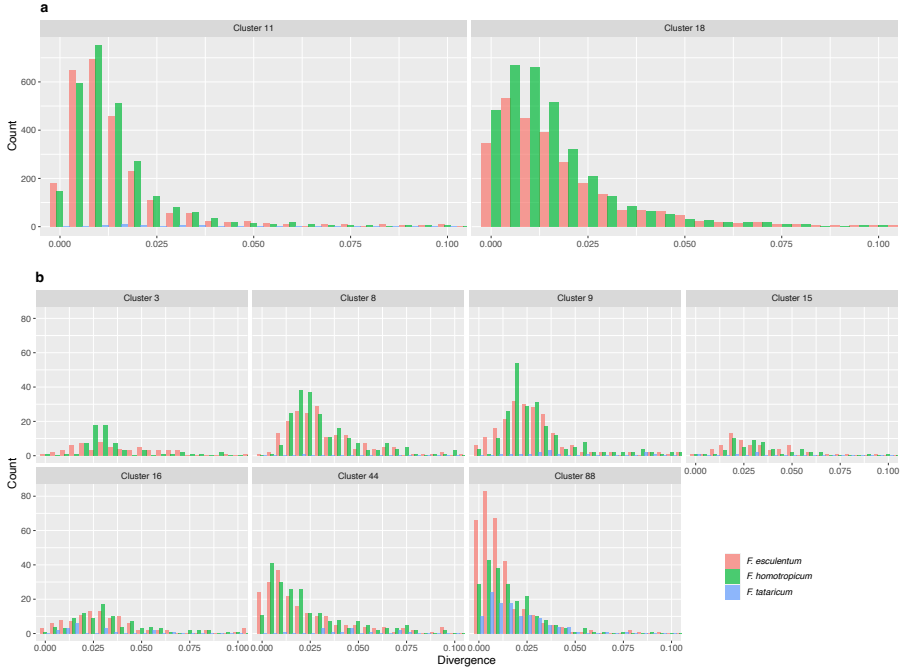
Supplementary Fig. 8 Circos plots of various components of the genomes of (a) *F. esculentum*, (b) *F. homotropicum*, and (c) *F. tataricum*. The heatmaps represents densities per Mb with the colour scale ranging from blue (low) to red (high). Grey blocks in a indicate regions derived from *F. homotropicum*. (i) GC content (35-45%) (ii) Gene density (0-180 genes). (iii) TE density (0-100%). (iv) Density of *Gypsy*-type LTR retrotransposons (0-100%). (v) Density of *Copia*-type LTR retrotransposons (0-20%). (vi) Density of DNA transposons (0-4%).



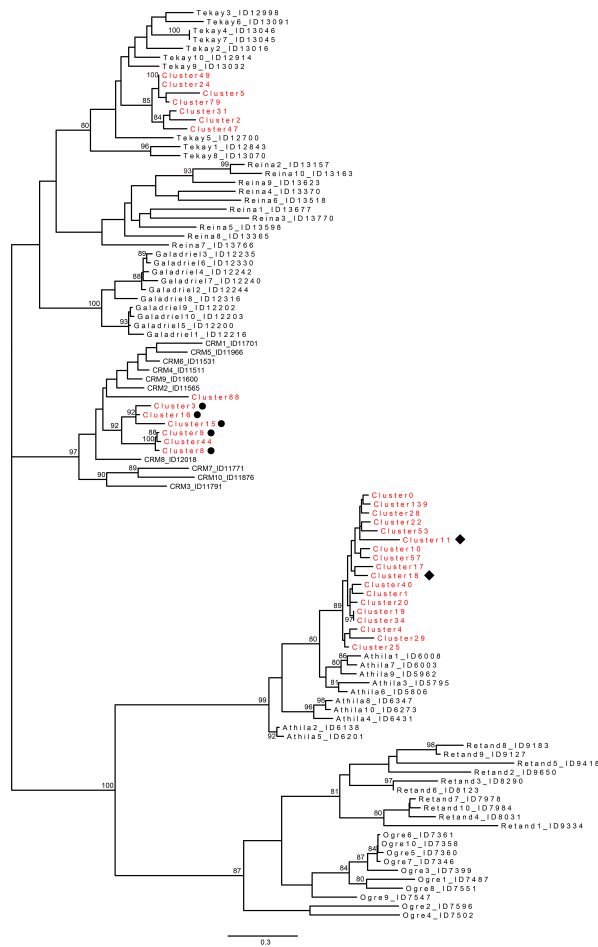
Supplementary Fig. 9 Distribution of various types of Transposable Elements across the *F. homotropicum* genome. For LTR retrotransposons, only full-length LTR retrotransposons whose nucleotide divergence could be calculated with an alignment length of ≥ 100 nucleotides are shown and the colour represents the nucleotide divergence between the flanking LTRs. The divergence corresponds to those shown in Extended Data Fig. 3 and Supplementary Fig. 11. Those with divergence > 0.1 are shown as 0.1. Athila.11 and Athila.18 are the two largest *Gypsy*-type subfamilies. CRM.44 and CRM.88 are *Gypsy*-type subfamilies of the *CRM* clade not associated with centromeric regions in *F. esculentum* and *F. homotropicum*, whereas the remaining *CRM* subfamilies (i.e. other) are associated with centromeric regions in *F. esculentum* and *F. homotropicum*. The numbers 11, 18, 44, and 88 correspond to the ClusterIDs of Supplementary Table 17 (see also Supplementary Fig. 12). Number of elements plotted are as follows - Copia: $n = 2,700$, Athila.18: $n = 3,399$, Athila.11: $n = 2,722$, Athila.other: $n = 4,268$, Tekay: $n = 1,197$, CRM.44: $n = 206$, CRM.88: $n = 215$, CRM.other: $n = 652$, LINE/SINE: $n = 5,525$, Helitron: $n = 801$.



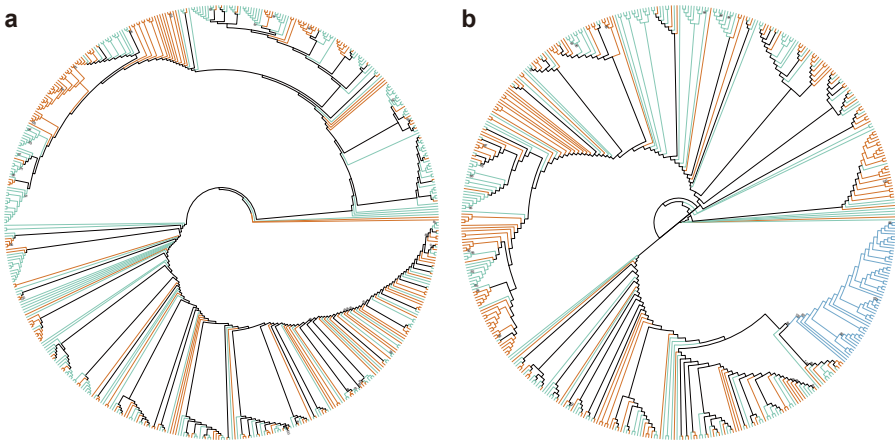
Supplementary Fig. 10 Distribution of various types of Transposable Elements across the *F. tataricum* genome. For LTR retrotransposons, only full-length LTR retrotransposons whose nucleotide divergence could be calculated with an alignment length of ≥ 100 nucleotides are shown and the colour represents the nucleotide divergence between the flanking LTRs. The divergence corresponds to those shown in Extended Data Fig. 3 and Supplementary Fig. 11. Those with divergence >0.1 are shown as 0.1. Athila.11 and Athila.18 are the two largest *Gypsy*-type subfamilies. CRM.44 and CRM.88 are *Gypsy*-type subfamilies of the *CRM* clade not associated with centromeric regions in *F. esculentum* and *F. homotropicum*, whereas the remaining *CRM* subfamilies (i.e. other) are associated with centromeric regions in *F. esculentum* and *F. homotropicum*. The numbers 11, 18, 44, and 88 correspond to the ClusterIDs of Supplementary Table 17 (see also Supplementary Fig. 12). Number of elements plotted are as follows - Copia: $n = 1,028$, Athila.18: $n = 0$, Athila.11: $n = 53$, Athila.other: $n = 372$, Tekay: $n = 1,427$, CRM.44: $n = 13$, CRM.88: $n = 116$, CRM.other: $n = 37$, LINE/SINE: $n = 6,930$, Helitron: $n = 370$.



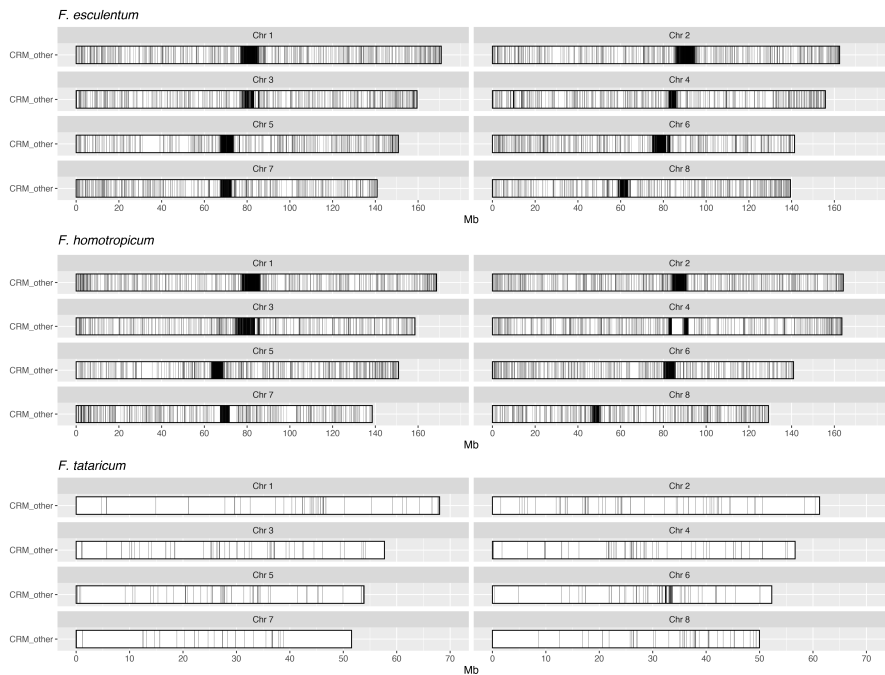
Supplementary Fig. 11 Nucleotide divergence between the flanking LTRs of subfamilies of full-length *Gypsy*-type LTR retrotransposons of the three *Fagopyrum* species. **a**, Two largest subfamilies in *F. esculentum* and *F. homotropicum* that have amplified recently in the two species, Cluster 11 (*F. esculentum*: $n = 2,603$, *F. homotropicum*: $n = 2,722$, *F. tataricum*: $n = 53$) and Cluster 18 (*F. esculentum*: $n = 2,712$, *F. homotropicum*: $n = 3,399$, *F. tataricum*: $n = 0$). **b**, *CRM* subfamilies, Cluster 3 (*F. esculentum*: $n = 62$, *F. homotropicum*: $n = 68$, *F. tataricum*: $n = 0$), Cluster 8 (*F. esculentum*: $n = 186$, *F. homotropicum*: $n = 208$, *F. tataricum*: $n = 1$), Cluster 9 (*F. esculentum*: $n = 216$, *F. homotropicum*: $n = 224$, *F. tataricum*: $n = 14$), Cluster 15 (*F. esculentum*: $n = 68$, *F. homotropicum*: $n = 55$, *F. tataricum*: $n = 4$), Cluster 16 (*F. esculentum*: $n = 100$, *F. homotropicum*: $n = 97$, *F. tataricum*: $n = 18$), Cluster 44 (*F. esculentum*: $n = 201$, *F. homotropicum*: $n = 206$, *F. tataricum*: $n = 13$), and Cluster 88 (*F. esculentum*: $n = 321$, *F. homotropicum*: $n = 215$, *F. tataricum*: $n = 116$). Clusters 3, 8, 9, 15, and 16 (referred to as *FeCEN*) are associated with centromeric regions in *F. esculentum* and *F. homotropicum* and their insertion ages are generally older compared to Clusters 44 and 88 that are of the same *CRM* clade but not associated with centromeric regions in all three species. Only those whose alignment length was ≥ 100 nucleotides are shown.



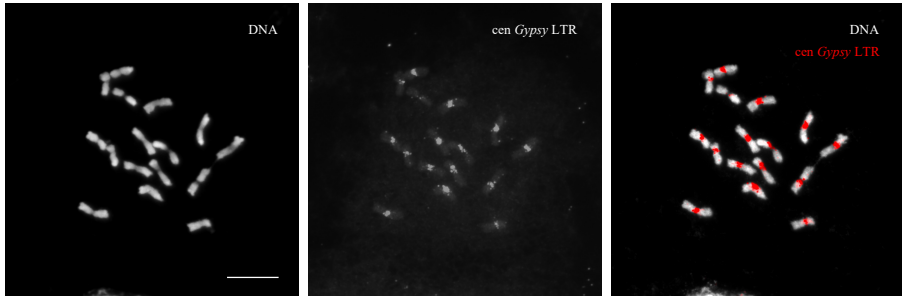
Supplementary Fig. 12 Phylogenetic relationship of *Gypsy* elements in *Fagopyrum* and *Gypsy* elements from other eudicot species. Representative sequence of each *Fagopyrum* subfamily with 100 or more sequences are shown in red. The sequence IDs correspond to the ClusterIDs in Supplementary Table 17. Black diamonds indicate the two largest subfamilies shown in Supplementary Fig. 11, and the black dots indicate subfamilies associated with centromeric regions in *F. esculentum* and *F. homotropicum* (*FeCEN*). Ten randomly selected sequences of each clade from various other eudicot species present in REXdb are shown in black. The relationship between each clade and the placement of the *Fagopyrum* species were consistent across trees constructed using different sets of randomly selected sequences. Phylogenetic tree was inferred by maximum likelihood with IQ-TREE [122] using the amino acid sequences of the reverse transcriptase domain and is unrooted. Bootstrap values of nodes with $\geq 80\%$ bootstrap support are indicated. Scale bar indicates branch length.



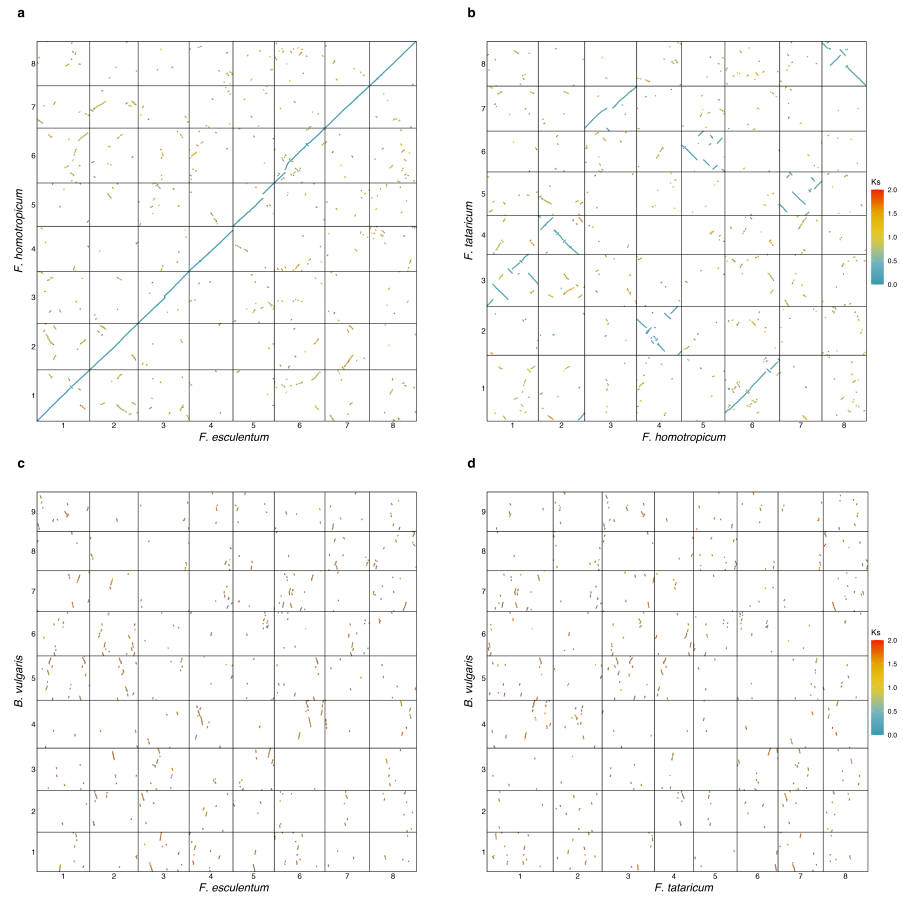
Supplementary Fig. 13 Phylogenetic tree of the two largest subfamilies of *Gypsy*-type LTR retrotransposons in *F. esculentum* and *F. homotropicum*. **a**, Phylogenetic tree of 500 randomly chosen sequences of the largest subfamily (Cluster 18). **b**, Phylogenetic tree of all 54 sequences of *F. tataricum* and 446 randomly chosen sequences of the remaining two species of the second largest subfamily (Cluster 11). Clusters 18 and 11 correspond to those shown in Supplementary Table 17. Sequences of *F. esculentum*, *F. homotropicum*, and *F. tataricum* (only present in **b**) are shown in red, green, and blue, respectively. Phylogenetic trees were inferred by maximum likelihood with IQ-TREE using the nucleotide sequences of the reverse transcriptase domain and is unrooted. Bootstrap values of nodes with $\geq 80\%$ bootstrap support are indicated.



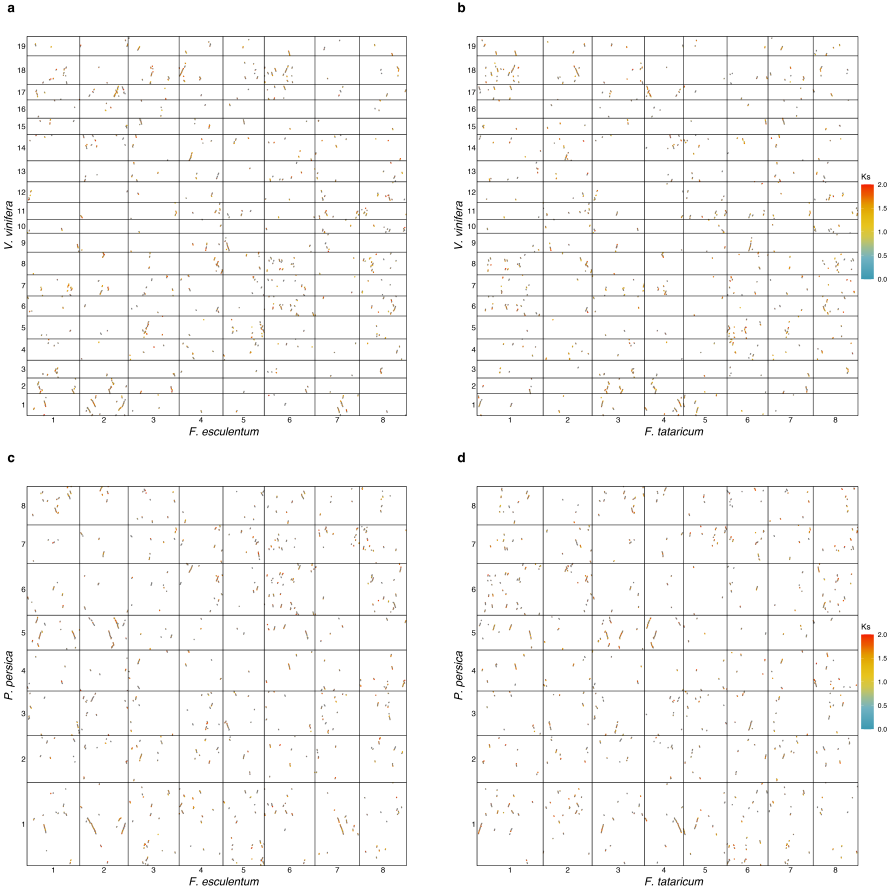
Supplementary Fig. 14 Distribution of non full-length *FeCEN* LTR retrotransposons across the *F. esculentum*, *F. homotropicum*, and *F. tataricum* genomes. 7,451, 7,075, and 455 copies of *F. esculentum*, *F. homotropicum*, and *F. tataricum*, respectively, are plotted.



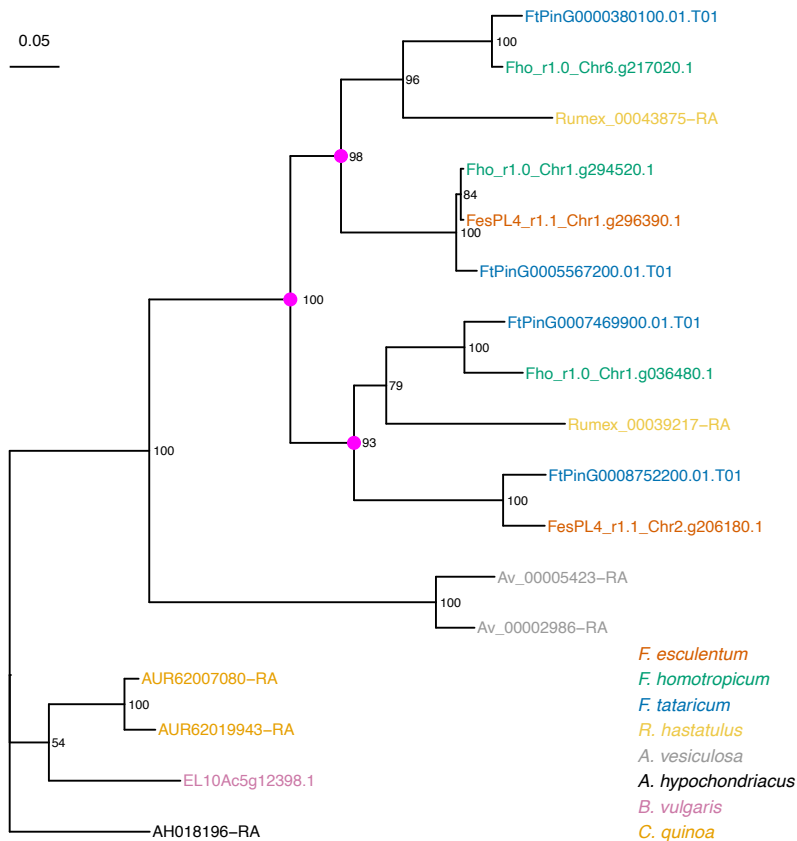
Supplementary Fig. 15 FISH analysis probed with *FeCEN* LTR retrotransposon to the mitotic chromosomes of *F. homotropicum*. Signals derived from the probe of *FeCEN* can be observed at the centromeric regions (including the primary constriction) of all *F. homotropicum* chromosomes. Scale bar = 5 μ m.



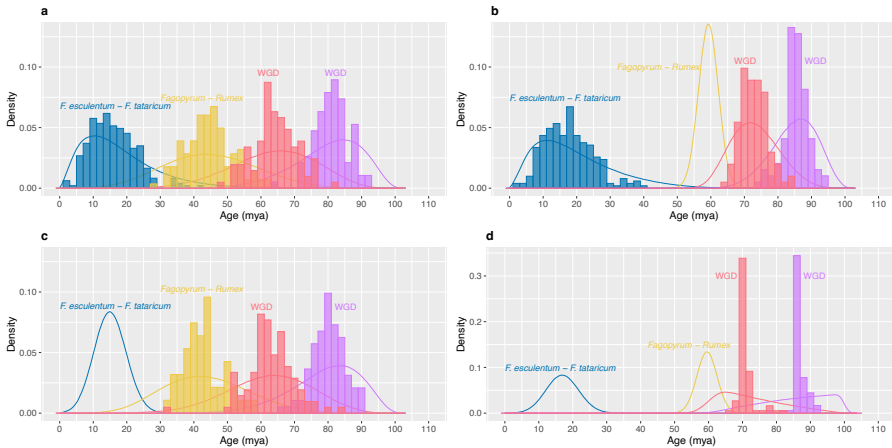
Supplementary Fig. 16 Genome-wide collinear dot plots between each *Fagopyrum* species and between *Fagopyrum* and sugarbeet. **a**, *F. esculentum* vs *F. homotropicum* (26,227 collinear gene pairs). **b**, *F. homotropicum* vs *F. tataricum* (22,789 collinear gene pairs). **c**, *F. esculentum* vs *B. vulgaris* (7,545 collinear gene pairs). **d**, *F. tataricum* vs *B. vulgaris* (7,207 collinear gene pairs). Each dot represents a collinear gene pair identified by MCSanX and the colour represents the K_s between the gene pair. K_s of >2 are coloured grey.



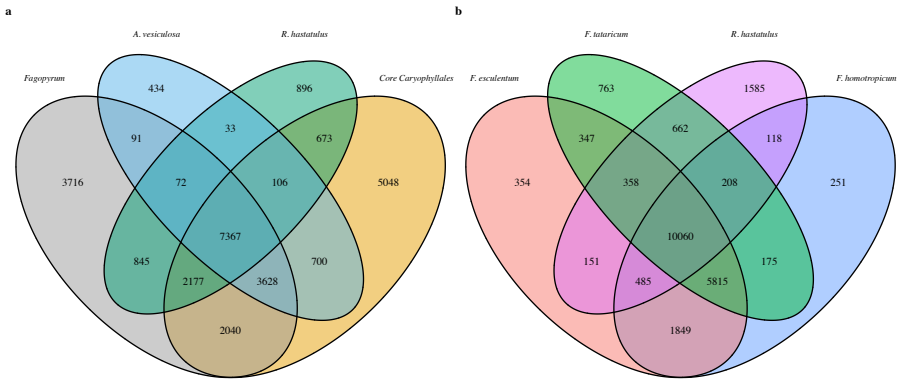
Supplementary Fig. 17 Genome-wide collinear dot plots between *Fagopyrum* and grapevine and peach. a, *F. esculentum* vs *V. vinifera* (8,994 collinear gene pairs). b, *F. tataricum* vs *V. vinifera* (8,503 collinear gene pairs). c, *F. esculentum* vs *P. persica* (9,804 collinear gene pairs). d, *F. tataricum* vs *P. persica* (9,407 collinear gene pairs). Each dot represents a collinear gene pair identified by MCScanX and the colour represents the K_s between the gene pair. K_s of >2 are coloured grey.



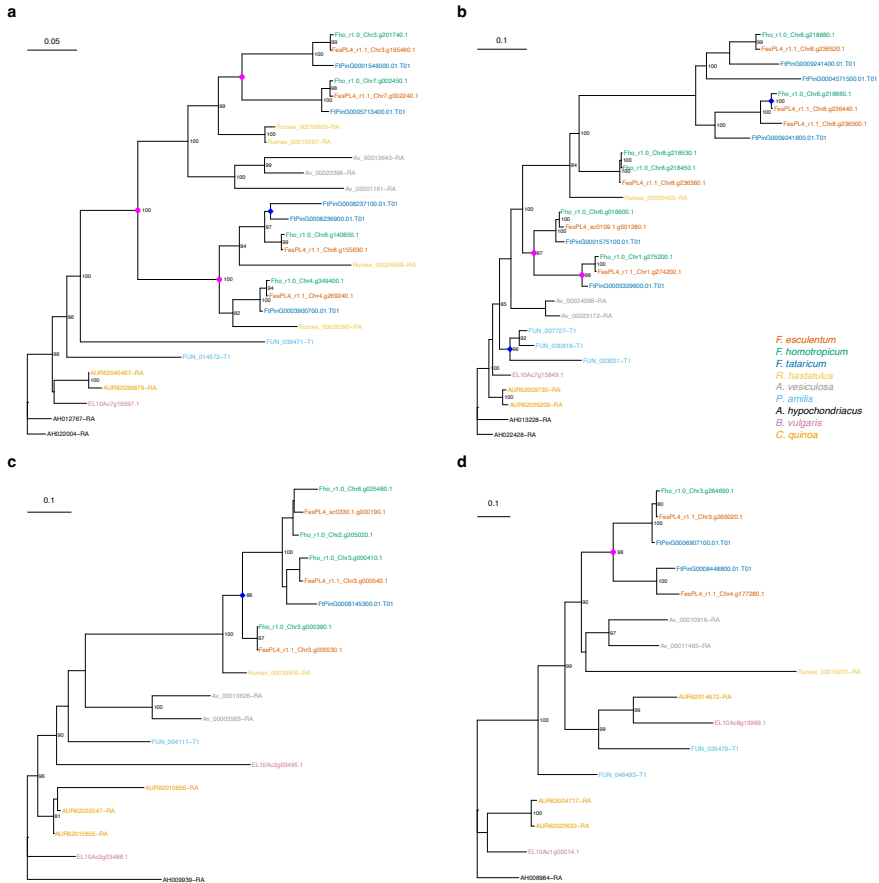
Supplementary Fig. 18 Example of a phylogenetic tree of a gene family with two whole-genome duplications (WGDs) in *Fagopyrum* shared with *Rumex* but not with *Aldrovanda*. Phylogenetic tree was inferred by maximum likelihood with IQ-TREE. Tree is unrooted and bootstrap values supporting each node are shown. Nodes corresponding to WGDs containing at least one collinear gene pair are indicated by magenta dots. Scale bar indicates branch length.



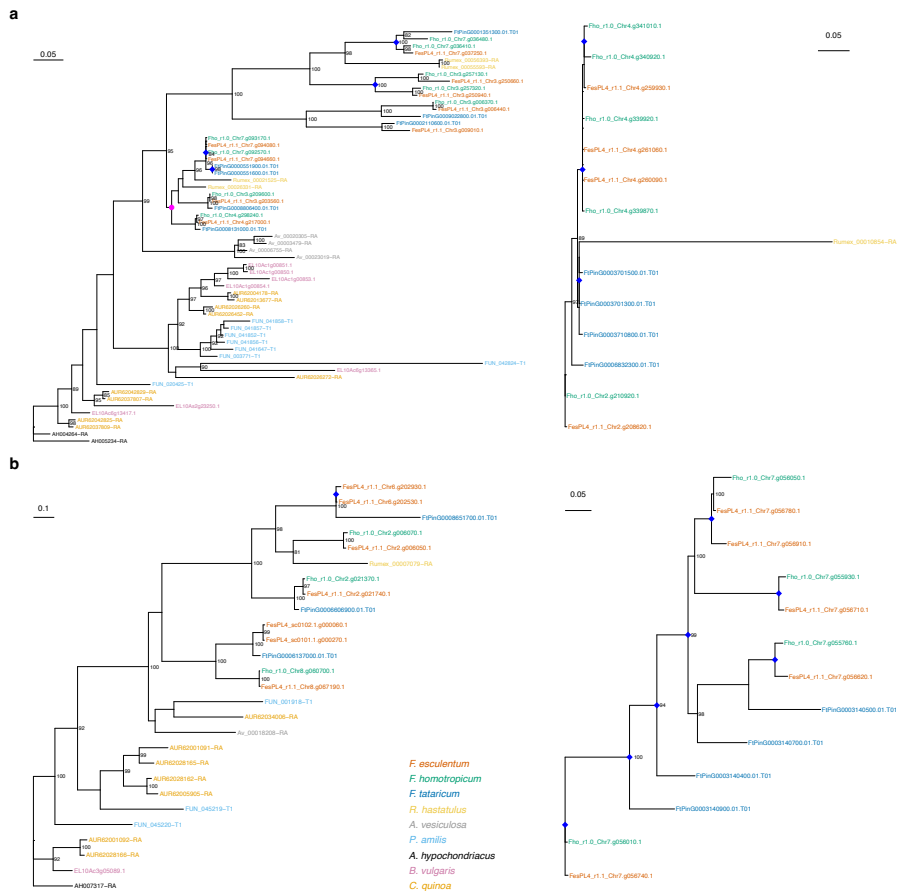
Supplementary Fig. 19 Effect of prior age constraints on phylogenetic dating results. Age estimates based on phylogenetic dating analysis of gene families with two WGDs in *Fagopyrum* shared with *Rumex* but not with *Aldroranda*. Bar plots indicate the median age estimates of each gene family and the line plots are based on all ages of the MCMC analyses combined. Bar plots are shown only for nodes where prior age constraints were not applied. Prior and posterior age estimates are shown in Supplementary Table 20. **a**, Results without prior age constraints on the nodes *F. esculentum*-*F. tataricum* and *Fagopyrum*-*Rumex* (Setting 2 of Supplementary Table 20). Number of data points for bar plots are - *F. esculentum*-*F. tataricum*: $n = 243$, *Fagopyrum*-*Rumex*: $n = 141$, younger WGD: $n = 103$, older WGD: $n = 95$. Number of data points for line plots are - *F. esculentum*-*F. tataricum*: $n = 2,187,243$, *Fagopyrum*-*Rumex*: $n = 1,269,141$, younger WGD: $n = 927,103$, older WGD: $n = 855,095$. **b**, Results without prior age constraints on the node *F. esculentum*-*F. tataricum* (Setting 3 of Supplementary Table 20). Number of data points for bar plots are - *F. esculentum*-*F. tataricum*: $n = 253$, younger WGD: $n = 101$, older WGD: $n = 98$. Number of data points for line plots are - *F. esculentum*-*F. tataricum*: $n = 2,277,253$, *Fagopyrum*-*Rumex*: $n = 1,332,148$, younger WGD: $n = 909,101$, older WGD: $n = 882,098$. **c**, Results without prior age constraints on the node *Fagopyrum*-*Rumex* (Setting 4 of Supplementary Table 20). Number of data points for bar plots are - *Fagopyrum*-*Rumex*: $n = 146$, younger WGD: $n = 104$, older WGD: $n = 96$. Number of data points for line plots are - *F. esculentum*-*F. tataricum*: $n = 2,268,252$, *Fagopyrum*-*Rumex*: $n = 1,314,146$, younger WGD: $n = 936,104$, older WGD: $n = 864,096$. **d**, Results run with only the priors and no sequence data (Setting 1 (no data) of Supplementary Table 20). Number of data points for bar plots are - younger WGD: $n = 93$, older WGD: $n = 90$. Number of data points for line plots are - *F. esculentum*-*F. tataricum*: $n = 2,682,298$, *Fagopyrum*-*Rumex*: $n = 1,593,177$, younger WGD: $n = 837,093$, older WGD: $n = 810,090$.



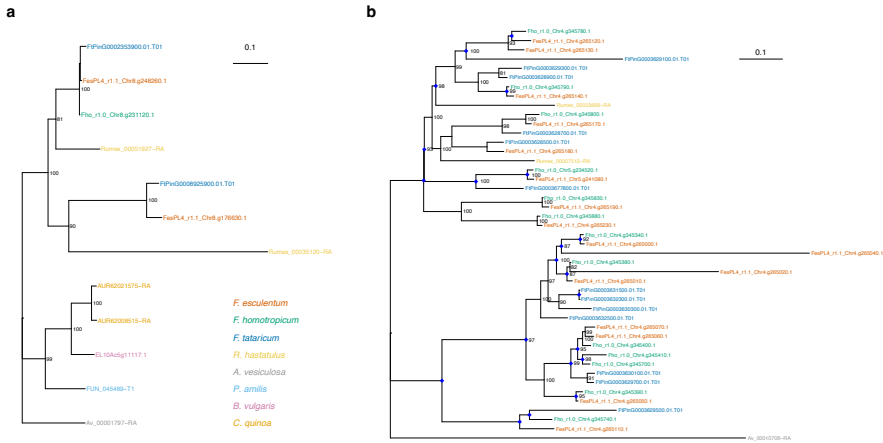
Supplementary Fig. 20 **Number of gene families identified by OrthoFinder.**
a, Venn diagram showing shared and unique gene families in *Fagopyrum* (*F. esculentum*, *F. homotropicum*, and *F. tataricum*), *R. hastatulus*, *A. vesiculosa*, core Caryophyllales (*P. amilis*, *B. vulgaris*, *C. quinoa*, and *A. hypochondriacus*) based on the hierarchical orthologous groups of all nine species identified by OrthoFinder. **b**, Venn diagram showing shared and unique gene families in *F. esculentum*, *F. homotropicum*, *F. tataricum*, and *R. hastatulus* based on the hierarchical orthologous groups of these four species identified by OrthoFinder.



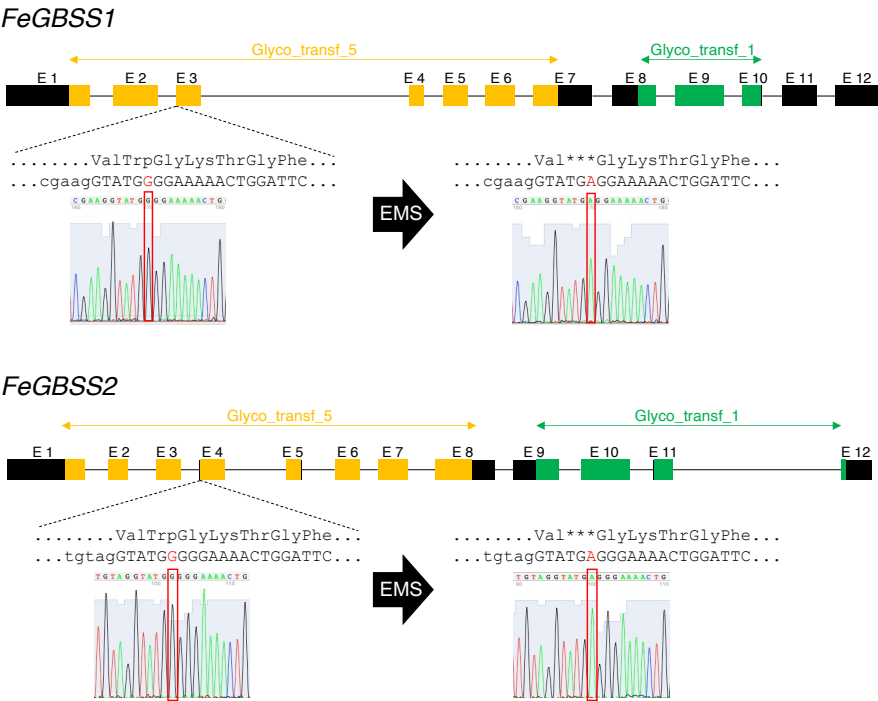
Supplementary Fig. 21 Phylogenetic trees of gene families related to rutin biosynthesis (PAL, C4H, CHI, and FLS). **a**, PAL (phenylalanine ammonia-lyase) gene family. **b**, C4H (cinnamate-4-hydroxylase) gene family. **c**, CHI (chalcone isomerase) gene family. **d**, FLS (flavonol synthase) gene family. Phylogenetic trees were inferred by maximum likelihood with IQ-TREE. Trees are unrooted and bootstrap values for nodes with $\geq 80\%$ bootstrap support are shown. Scale bars indicate branch length. Nodes corresponding to WGDs containing at least one collinear gene pair are indicated by magenta dots, and nodes corresponding to tandem duplications with at least one gene pair with five or less intervening genes are indicated by blue diamonds. Each gene family corresponds to the gene families shown in Supplementary Table 28.



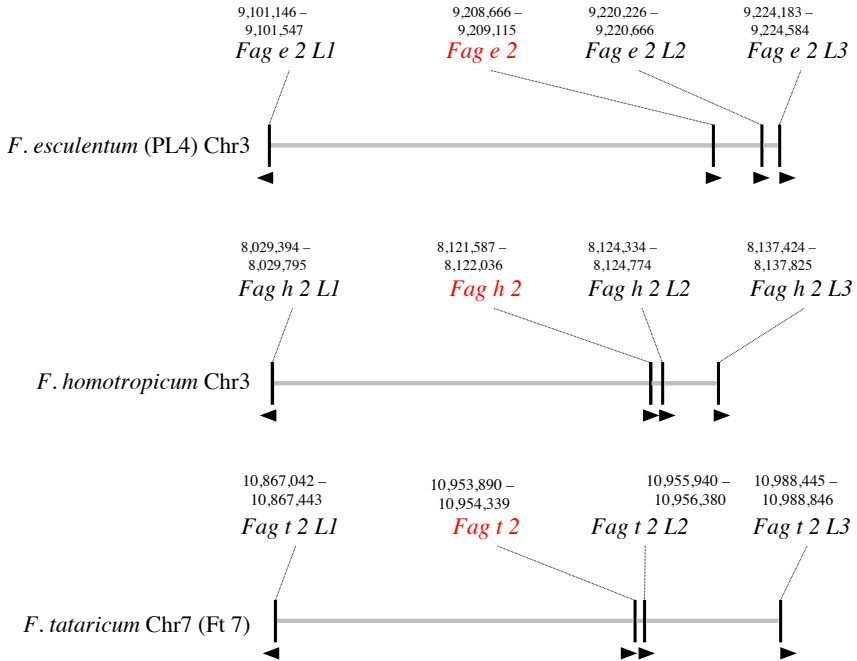
Supplementary Fig. 22 Phylogenetic trees of gene families related to rutin biosynthesis (CHS and GTR). a, CHS (chalcone synthase) gene families. Trees to the left and right correspond to CHS-1 and CHS-2 of Supplementary Table 28, respectively. b, GTR (glycosyltransferase) gene families. Trees to the left and right correspond to GTR-2 and GTR-3 of Supplementary Table 28, respectively. Phylogenetic trees were inferred by maximum likelihood with IQ-TREE. Trees are unrooted and bootstrap values for nodes with $\geq 80\%$ bootstrap support are shown. Scale bars indicate branch length. Nodes corresponding to WGDs containing at least one collinear gene pair are indicated by magenta dots, and nodes corresponding to tandem duplications with at least one gene pair with five or less intervening genes are indicated by blue diamonds.



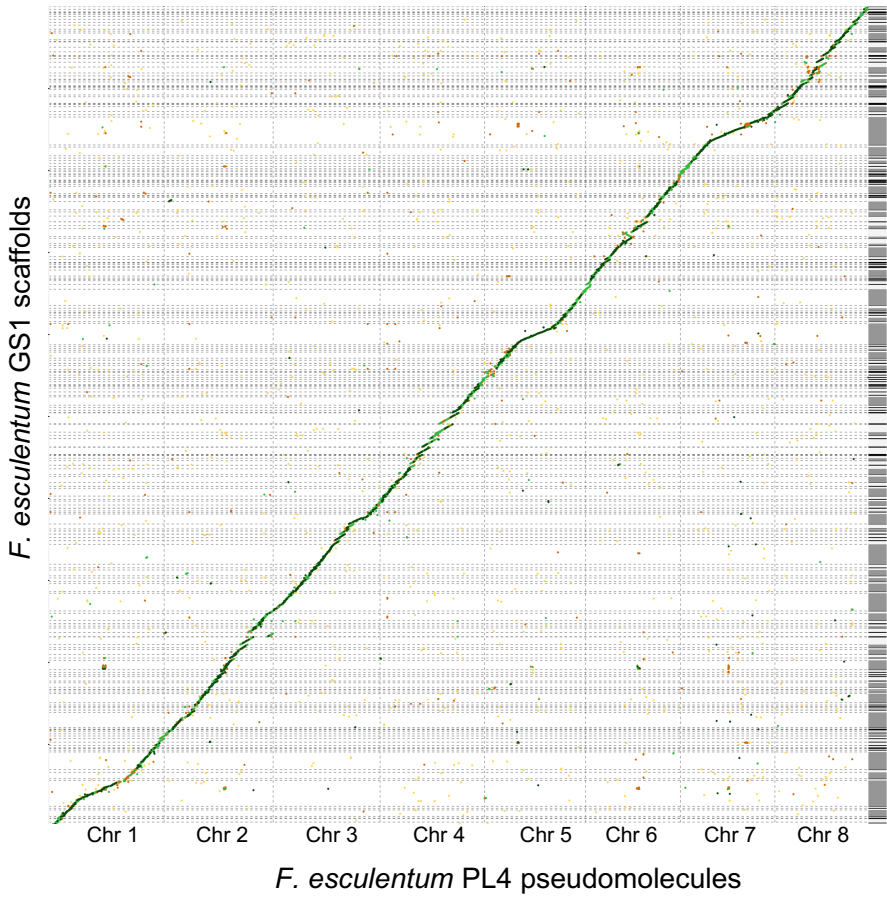
Supplementary Fig. 23 Phylogenetic trees of gene families related to rutin biosynthesis (F3'H and F3'5'H). **a**, F3'H (flavonoid-3'-hydroxylase) gene family. **b**, F3'5'H (flavonoid-3'-5'-hydroxylase) gene family. Phylogenetic trees were inferred by maximum likelihood with IQ-TREE. Trees are unrooted and bootstrap values for nodes with $\geq 80\%$ bootstrap support are shown. Scale bars indicate branch length. Nodes corresponding to tandem duplications with at least one gene pair with five or less intervening genes are indicated by blue diamonds. Both gene families corresponds to those shown in Supplementary Table 28.



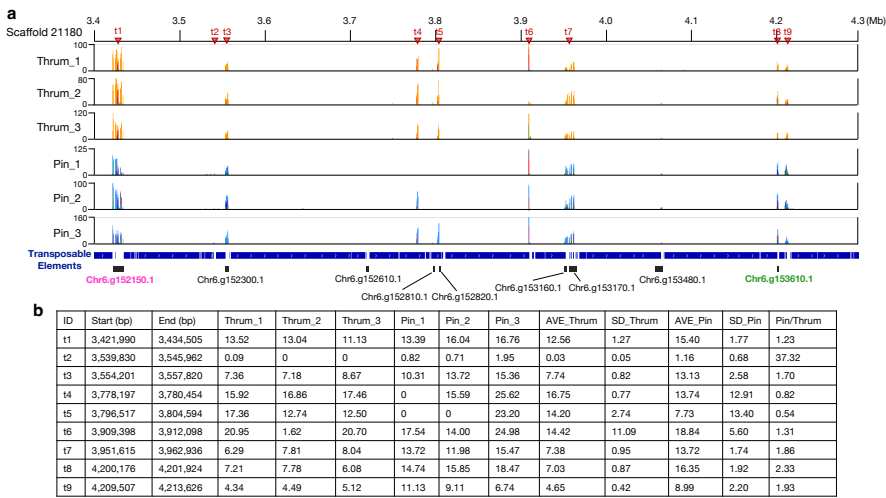
Supplementary Fig. 24 Diagrams of *FeGBSS1* and *FeGBSS2* genes with EMS-induced nonsense mutations. Both genes consist of 12 exons (thick black lines of E1-E12). Two domains identified by searching InterPro (<https://www.ebi.ac.uk/interpro/>) [123] using the full-length amino acid sequences of the two genes as queries are displayed in orange (Glyco_transf_5) and green (Glyco_transf_1). The electropherogram traces from Sanger sequencing and the nucleotide and amino acid sequences are also presented for the wild type of *FeGBSS1* and *FeGBSS2*.



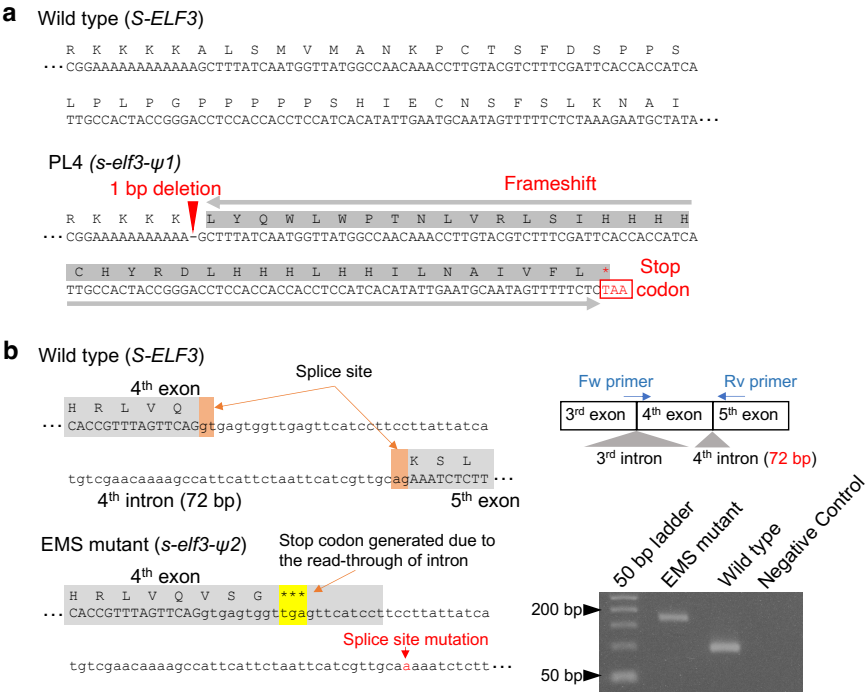
Supplementary Fig. 25 Genomic locations of *Fag e 2* (2S albumin) gene and its homologues in the three *Fagopyrum* species. *Fag e 2* and its orthologous genes (*Fag h 2* and *Fag t 2*) are shown in red. Arrows indicate the orientation of the genes.



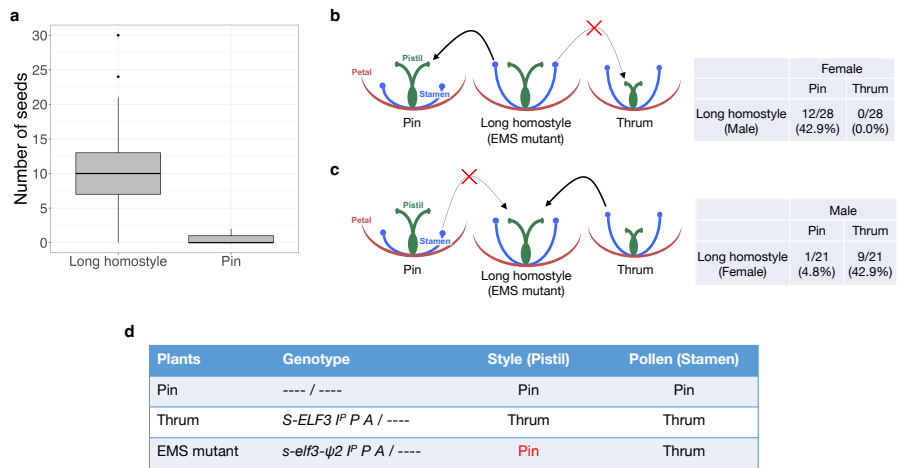
Supplementary Fig. 26 Dotplot between *F. esculentum* PL4 pseudomolecules and *F. esculentum* GS1 scaffolds. Dotplot analysis was conducted by minimap with the options ‘-H -k 27’. The paf-formatted result was visualized by the D-GENIES platform (<https://dgenies.toulouse.inra.fr>). Only *F. esculentum* GS1 scaffolds of over 100 kb were used. Scaffolds and chromosomes are separated by dashed lines. The colours correspond to the identity calculated by D-GENIES.



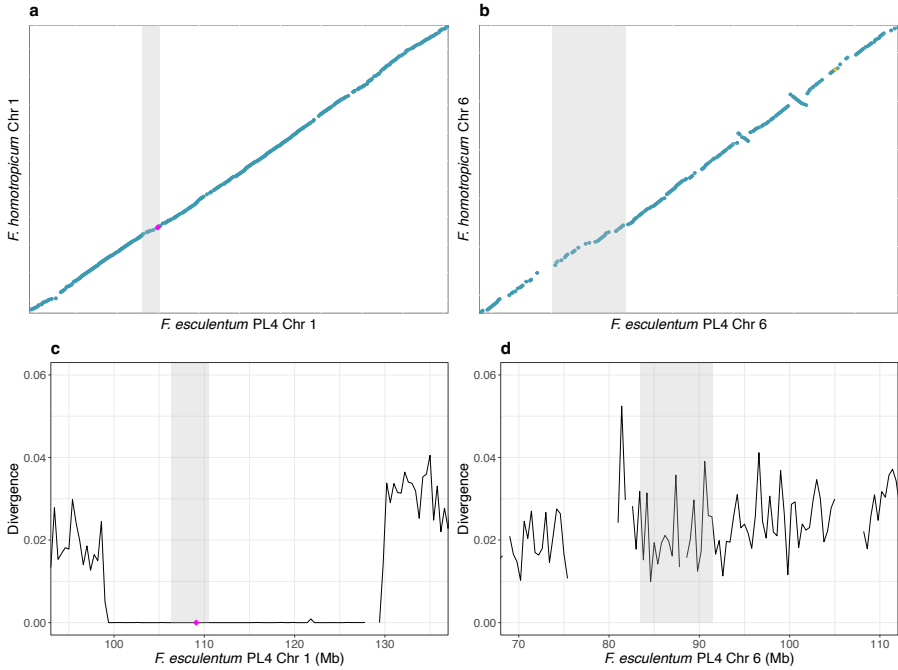
Supplementary Fig. 27 Genomic structure and RNA-Seq analysis of the *s* haplotype region. **a**, Depth of RNA-Seq reads of the flowers of three pin and three thrum plants mapped to the *s* haplotype region of scaffold 21180 by HISAT2 [124]. No expression specific to pin flowers is observed and the expression pattern is similar between pin and thrum flowers, suggesting that the *s* haplotype does not contain any genes with pin-specific functions. The dark blue horizontal bars represent TEs detected by RepeatMasker (www.repeatmasker.org) using the library of TEs constructed in *F. esculentum* PL4. Red arrowheads labelled t1-t9 indicate positions of transcripts identified by Stringtie [125] and the IDs correspond to those shown in **b**. The two genes Chr6.g152150.1 and Chr6.g153610.1 correspond to the border genes squared by the same colours shown in Extended Data Fig. 7b and were identified by mapping the CDS sequences of the genes predicted in *F. esculentum* PL4 by GMAP. As identifying genes based on GMAP may be too conservative, we performed BLASTN with the CDS sequences of the seven genes predicted between the two border genes in *F. esculentum* PL4 Chr 6 as queries against scaffold 21180. Homologous sequences of all seven genes could be identified in scaffold 21180 with >99% nucleotide identity. **b**, Transcript per million (TPM) values of each transcript in pin and thrum plants calculated by Stringtie. The average (AVE) and standard deviation (SD) of the three pin flowers and three thrum flowers, and the ratio of the average of pin and thrum flowers are also shown.



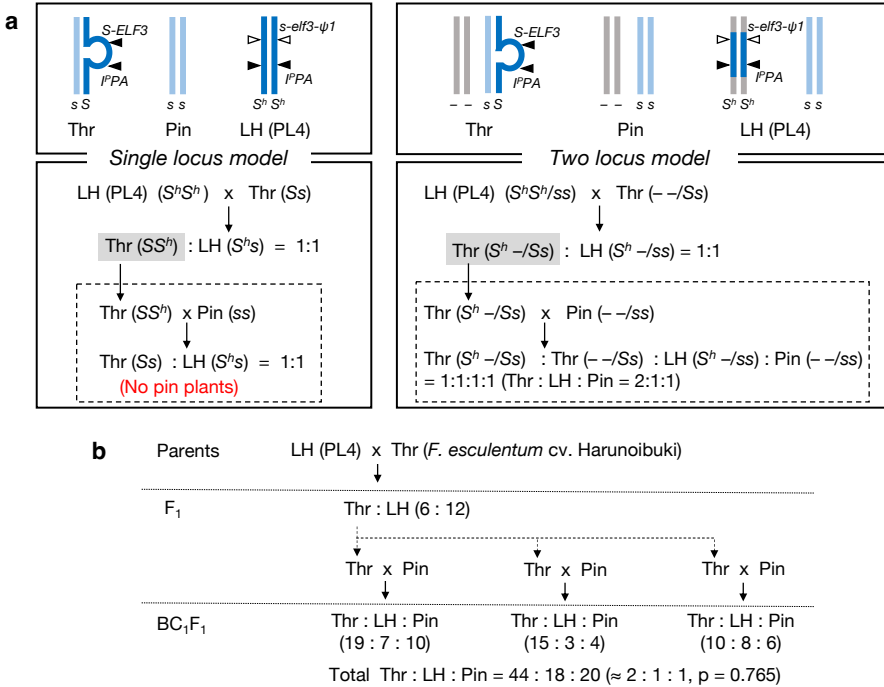
Supplementary Fig. 28 Mutants of *S-ELF3*. **a**, Nucleotide and amino acid sequences of *S-ELF3* of *F. esculentum* GS1 and *F. esculentum* PL4 around the 5th exon in which PL4 contains a 1 bp deletion leading to a frameshift. *F. homotropicum* has an identical sequence including the 1 bp deletion and stop codon to that of *F. esculentum* PL4. **b**, Left panel shows the nucleotide and amino acid sequences of *S-ELF3* of *F. esculentum* GS1 and the EMS mutant around the 4th and 5th exon. The EMS mutant contains a splice site mutation which generates a nonsense mutation due to the read-through of intron. Right panel shows the results of RT-PCR and the positions of primers used to confirm the presence of aberrant transcripts in which the 4th intron is not spliced out.



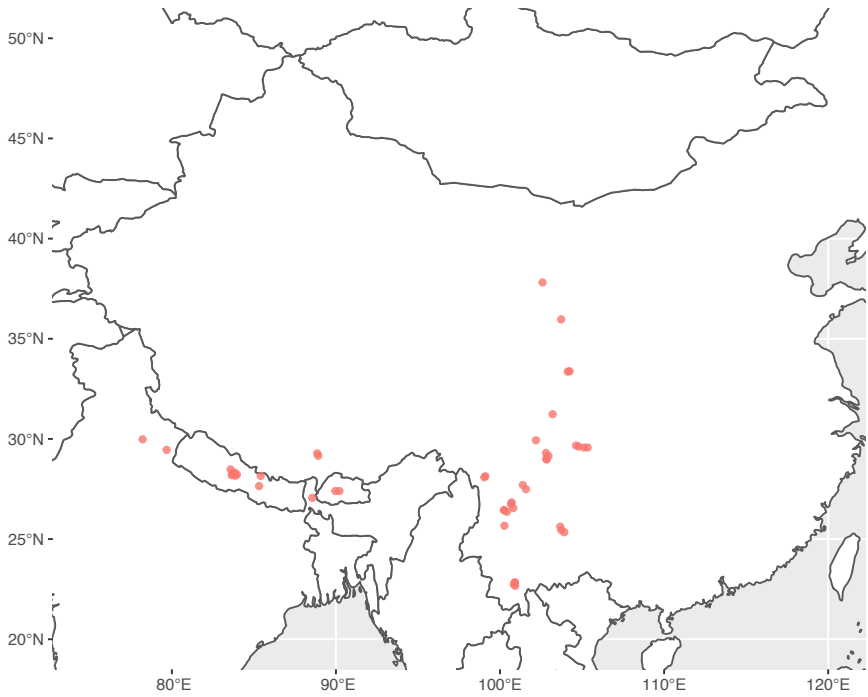
Supplementary Fig. 29 Self and cross compatibility of long homostyle flowers of *S-ELF3* knockout plants. **a**, Self compatibility between the pollen and pistil of long homostyle mutants and the pollen and pistil of pin plants. The box plot shows the number of seeds yielded per plant for 37 long homostyle plants and 19 pin plants after self-pollination. The box covers the first and third quartiles with the line corresponding to the median and the whiskers extending to 1.5 times the inter-quartile range. **b**, Compatibility between the pollen of long homostyle mutants and the pistils of wild type thrum and pin plants. The pollen of long homostyle flowers are incompatible with the pistil of thrum flowers (0/28 = 0%) but compatible with the pistil of pin flowers (12/28 = 42.9%). **c**, Cross compatibility between the pollen of wild type thrum and pin plants and the pistil of long homostyle mutants. The pollen of thrum flowers are compatible with the pistil of long homostyle flowers (9/21 = 42.9%) whereas the pollen of pin flowers are incompatible with the pistil of long homostyle flowers (1/21 = 4.8%). All wild type plants used here are *F. esculentum* cv. Harunoibuki. **d**, Expected genotypes of pin plants, thrum plants, and the long homostyle EMS mutants under the model assumed in Main Fig. 4a,c and whether the style (pistil) and pollen (stamen) are of pin-type or thrum-type plants.



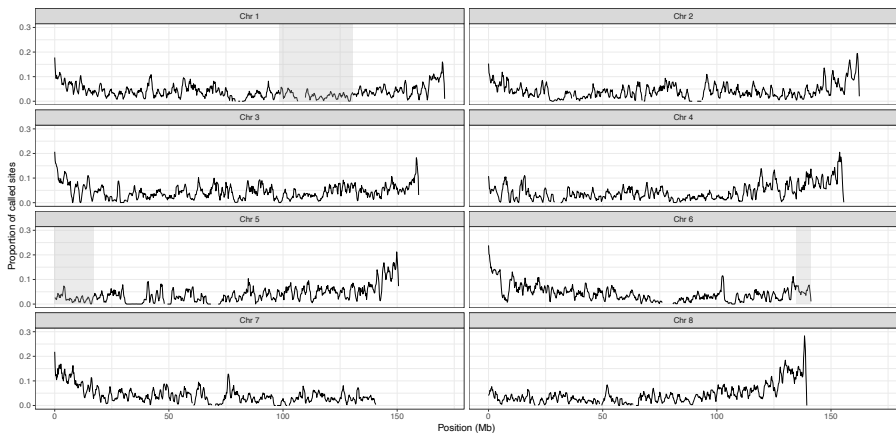
Supplementary Fig. 30 Characterization of regions homologous to *S* and *s* haplotypes in *F. esculentum* PL4. **a**, Gene-based collinearity between *F. esculentum* PL4 and *F. homotropicum* detected by MCScanX around 95-135 Mb of *F. esculentum* PL4 Chr 1. **b**, Gene-based collinearity between *F. esculentum* PL4 and *F. homotropicum* around 70-110 Mb of *F. esculentum* PL4 Chr 6. **c**, Nucleotide divergence, i.e., the average number of differences per site, between *F. esculentum* PL4 and *F. homotropicum* around 95-135 Mb of *F. esculentum* PL4 Chr 1. **d**, Nucleotide divergence between *F. esculentum* PL4 and *F. homotropicum* around 70-110 Mb of *F. esculentum* PL4 Chr 6. The genome-wide plot of the gene-based collinearity is shown and described in Supplementary Fig. 16 and the genome-wide plot of the divergence between *F. esculentum* PL4 and *F. homotropicum* is shown and described in Extended Data Fig. 2. Homologue of *S-ELF3* in *F. esculentum* PL4 which has a nonsense mutation is indicated by magenta diamonds in **a** and **c**. The region in *F. esculentum* PL4 Chr 1 that is collinear to scaffold 2156 (*S* haplotype) of *F. esculentum* GS1 and the region in *F. esculentum* PL4 Chr 6 that is collinear to scaffold 21180 (*s* haplotype) as depicted in Extended Data Fig. 7b are indicated by grey shades in **a** and **c**, and **b** and **d**, respectively.



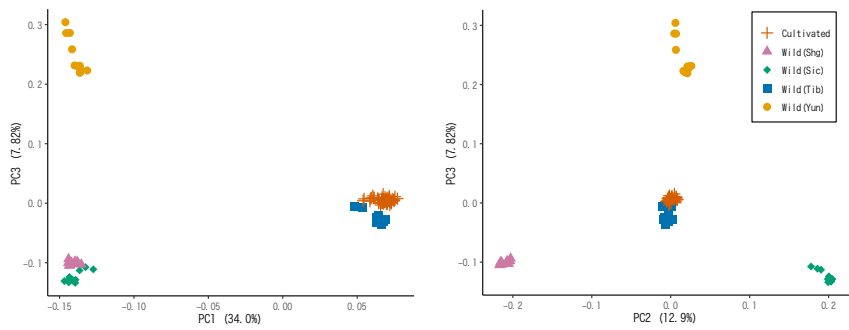
Supplementary Fig. 31 Expected segregation of flower types in the single locus and two locus models and results of allelism test crosses. a, Upper squares illustrate the allelic relationship of S , S^h , s in the two models. S , s , and S^h are alleles in the single locus model, whereas S^h is not an allele of S or s and located on a different chromosome in the two locus model. The lower squares show the crossing scheme for the allelism test cross. The expected segregation ratio of the phenotypes following the allelism test cross is shown in the dashed squares. Thr, Pin, and LH indicate plants with thrum, pin, and long homostyle flowers, respectively. A cross between thrum and pin plants is expected to produce thrum and long homostyle plants at a 1:1 ratio (no pin plants) under the single locus model and thrum, long homostyle, and pin plants at a 2:1:1 ratio under the two locus model. **b,** Results of the allelism test cross. 18 F_1 plants were obtained, of which three thrum flower plants were back-crossed with pin flower plants of cv. Harunoibuki. The ratio in parentheses indicate the number of observed phenotypes.



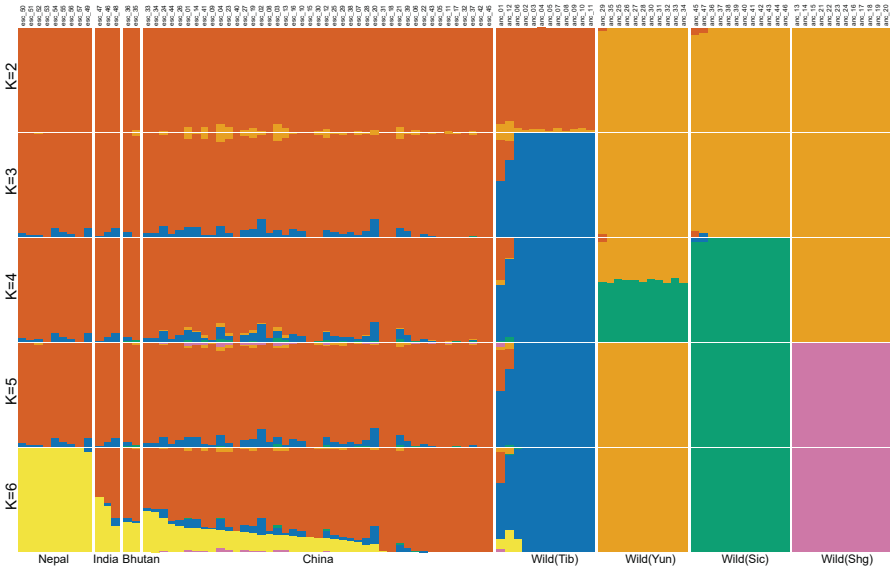
Supplementary Fig. 32 Geographic origin of cultivated accessions used for population genetic analyses. Red dots indicate the geographic origin of each cultivated accession shown in Supplementary Table 34. The map was drawn using the R rnatualearth package (<https://github.com/ropensci/rnatualearth>).



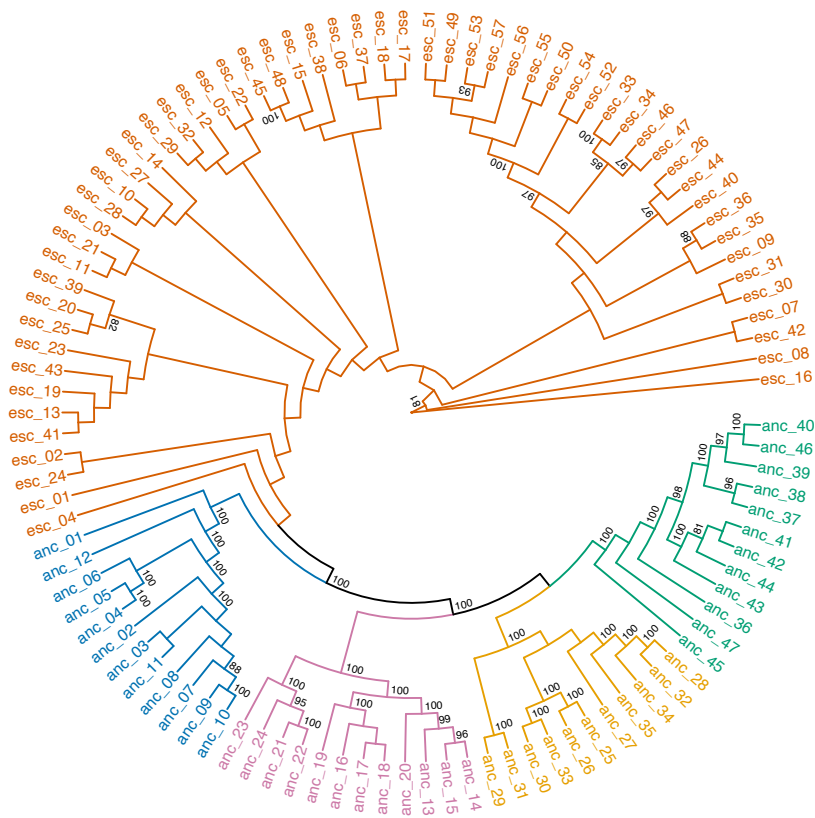
Supplementary Fig. 33 Proportion of high-confidence called sites per region retained for population genetic analyses. Called sites consist of sites where the genotypes of $\geq 80\%$ of the accessions were called. Distribution is based on a sliding window of 1 Mb with a step of 200 kb and only windows with $\geq 1,000$ called sites are plotted. Regions likely derived from *F. homotropicum* are indicated by grey shades.



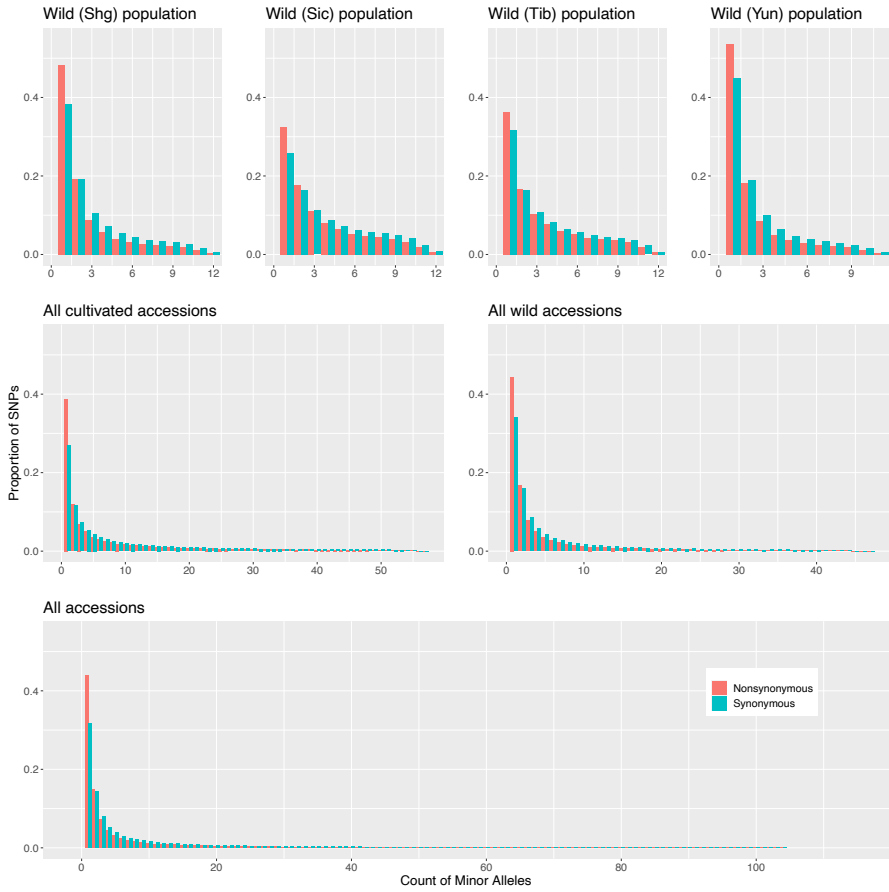
Supplementary Fig. 34 PC1 vs PC3 and PC2 vs PC3 of Principal Component Analysis. Results of PC1 vs PC2 are shown in Main Fig. 5.



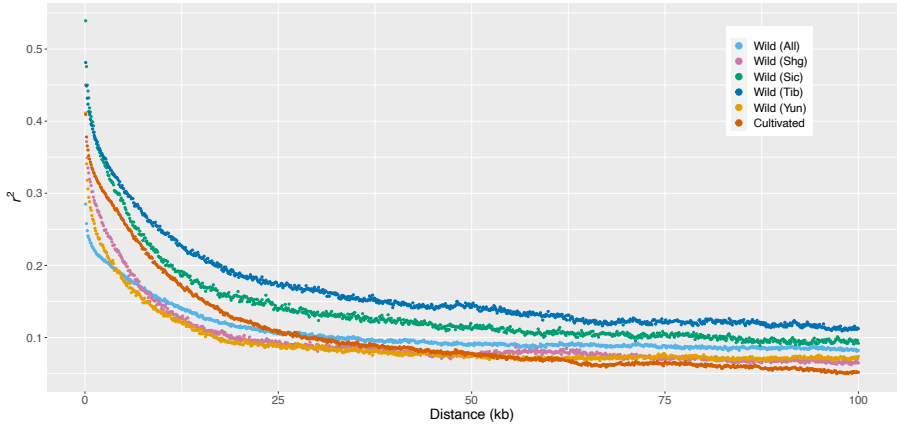
Supplementary Fig. 35 Results of ADMIXTURE analysis of $K=2\sim6$. Detailed figure based on the same results shown in Main Fig. 5 with the ID and derived region/country of each accession shown. Nepal, India, Bhutan, and China represent cultivated samples from each country. IDs of each accession correspond to those in Supplementary Table 34. Cross validation error for each K is shown in Supplementary Table 37.



Supplementary Fig. 36 Consensus phylogenetic tree constructed with all cultivated and wild accessions. Consensus tree of 1,000 bootstrap replicates of the tree shown in Main Fig. 5. Tree is unrooted and was inferred using the Neighbor-Joining method with the neighbor program of PHYLIP (<https://evolution.genetics.washington.edu/phylip>). Bootstrap replicates were generated with the SEQBOOT program of PHYLIP and the consensus tree was generated with the consense program of PHYLIP. Bootstrap values of $\geq 80\%$ are indicated by each node. IDs of each accession correspond to those in Supplementary Table 34. Colours of each population follow Main Fig. 5.



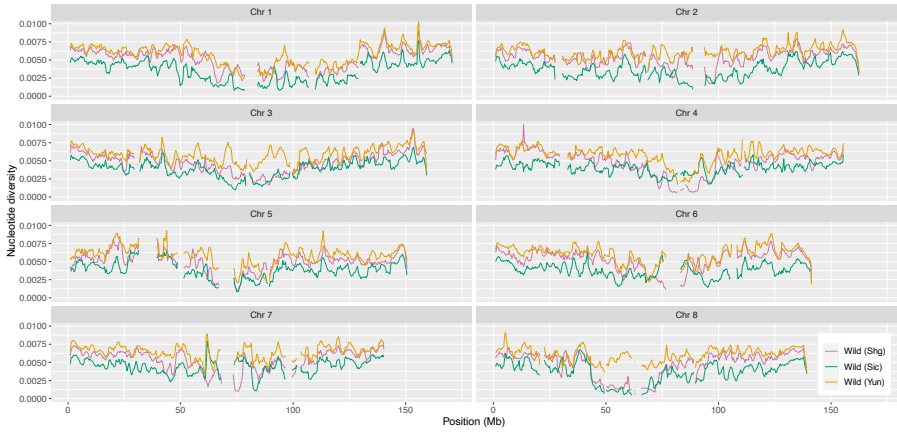
Supplementary Fig. 37 Minor allele frequency spectrum of synonymous and nonsynonymous sites of each population. Only codons where all three positions were retained and one of the three positions was bi-allelic were used.



Supplementary Fig. 38 Decay of linkage disequilibrium of each population plotted against physical distance. Average r^2 values of each non-overlapping 100 bp window are plotted.



Supplementary Fig. 39 Distribution of nucleotide diversity π of the cultivated and Wild Tib populations across each chromosome. Average π was calculated for each sliding window of 2 Mb with a step of 400 kb. Windows with average values based on <10,000 sites are not shown.



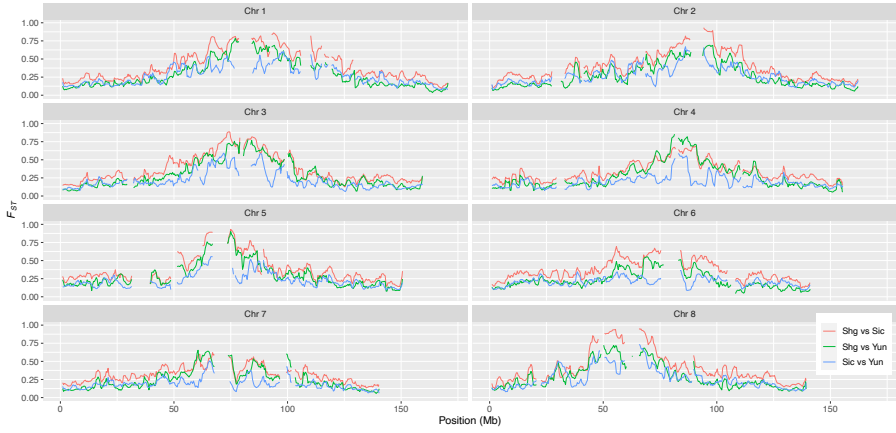
Supplementary Fig. 40 Distribution of nucleotide diversity π of the Wild Shg, Sic, and Yun populations across each chromosome. Average π was calculated for each sliding window of 2 Mb with a step of 400 kb. Windows with average values based on <10,000 sites are not shown.



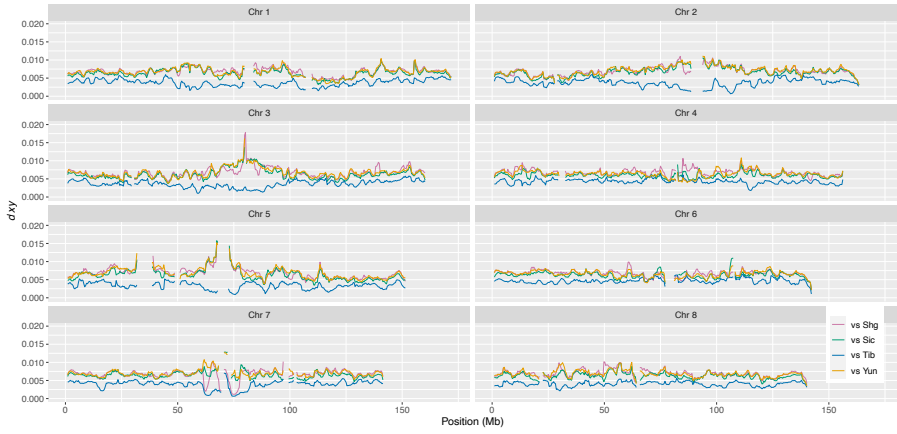
Supplementary Fig. 41 Distribution of F_{ST} between the cultivated and each wild population across each chromosome. Average F_{ST} was calculated for each sliding window of 2 Mb with a step of 400 kb. Windows with average values based on <100 sites are not shown.



Supplementary Fig. 42 Distribution of F_{ST} between the Wild Tib and the other wild populations across each chromosome. Average F_{ST} was calculated for each sliding window of 2 Mb with a step of 400 kb. Windows with average values based on <100 sites are not shown.



Supplementary Fig. 43 Distribution of F_{ST} between each wild populations apart from the Wild Tib population across each chromosome. Average F_{ST} was calculated for each sliding window of 2 Mb with a step of 400 kb. Windows with average values based on <100 sites are not shown.



Supplementary Fig. 44 Distribution of genetic divergence d_{xy} between the cultivated and each wild population across each chromosome. Average d_{xy} was calculated for each sliding window of 2 Mb with a step of 400 kb. Windows with average values based on $<3,000$ sites are not shown.