BMC Biology

RESEARCH ARTICLE

Open Access

Origin and diversity of *Capsella bursa-pastoris* from the genomic point of view

Aleksey A. Penin[1*†], Artem S. Kasianov[1†], Anna V. Klepikova[1], Denis O. Omelchenko[1], Maksim S. Makarenko[1] and Maria D. Logacheva[1,2]

Abstract

Background *Capsella bursa-pastoris*, a cosmopolitan weed of hybrid origin, is an emerging model object for the study of early consequences of polyploidy, being a fast growing annual and a close relative of *Arabidopsis thaliana*. The development of this model is hampered by the absence of a reference genome sequence.

Results We present here a subgenome-resolved chromosome-scale assembly and a genetic map of the genome of *Capsella bursa-pastoris*. It shows that the subgenomes are mostly colinear, with no massive deletions, insertions, or rearrangements in any of them. A subgenome-aware annotation reveals the lack of genome dominance—both subgenomes carry similar number of genes. While most chromosomes can be unambiguously recognized as derived from either paternal or maternal parent, we also found homeologous exchange between two chromosomes. It led to an emergence of two hybrid chromosomes; this event is shared between distant populations of *C. bursa-pastoris*. The whole-genome analysis of 119 samples belonging to *C. bursa-pastoris* and its parental species *C. grandiflora/rubella* and *C. orientalis* reveals introgression from *C. orientalis* but not from *C. grandiflora/rubella*.

Conclusions *C. bursa-pastoris* does not show genome dominance. In the earliest stages of evolution of this species, a homeologous exchange occurred; its presence in all present-day populations of *C. bursa-pastoris* indicates on a single origin of this species. The evidence coming from whole-genome analysis challenges the current view that *C. grandiflora/rubella* was a direct progenitor of *C. bursa-pastoris*; we hypothesize that it was an extinct (or undiscovered) species sister to *C. grandiflora/rubella*.

Keywords *Capsella bursa-pastoris*, Genome assembly, Genome evolution, Genetic diversity, Chromosome-scale assembly, Allopolyploid

Background

Polyploidization, or whole-genome duplication, is a recurrent trend in the evolution of eukaryotic genomes [1]. It is especially prevalent in plants—in this lineage it greatly contributed to the morphological diversity and ecological adaptations [2]. Polyploidy can also influence such critical traits as mating system (self-incompatible vs self-compatible), tolerance to abiotic stresses and growth rates. It was also shown that polyploidy is a key factor of plant domestication [3]. It is thus thoroughly studied; however, many important points concerning polyploids are not well understood. While the main model object of plant genetics, *Arabidopsis thaliana*, is not a polyploid, much attention has focused on the relatives of *A. thaliana* that display auto- or allopolyploidy [4–6]. Their close relationships to *A. thaliana* enhance the characterization of genomes and the application of modern functional genetics tools such as gene silencing and genetic

†Aleksey A. Penin and Artem S. Kasianov contributed equally to this work.

*Correspondence:
Aleksey A. Penin
alekseypenin@gmail.com
[1] Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia
[2] Skolkovo Institute of Science and Technology, Moscow, Russia

modification. Among these species, the most intriguing is *Capsella bursa-pastoris*. This is an allopolyploid; after a long discussion about its origin, its parental species were found to be *Capsella rubella/grandiflora* (these two species diverged very recently, many years after the emergence of *C. bursa-pastoris*) and *Capsella orientalis* [7], two species that diverged about 1 Mya [7]. Each parent has relatively small habitat regions—*C. grandiflora* is confined to Mediterranean region [8] and *C. orientalis*—to Eastern and Southern Russia, Kazakhstan, Mongolia, and China [9]. Surprisingly, their hybrid is a cosmopolitan weed growing around the globe—from Arctica on the North and Kerguelen islands on the South [10, 11]. It is one of the most widespread plants on Earth, the ecological niches that it inhabits range from African deserts to polar lands. Such plasticity has been at the focus of the studies in ecology and ecological genetics [12, 13]. It is presumably mediated by polyploidy; earlier some genetic cues that might be involved in environmental adaptation were found, they include differential splicing and asymmetry of regulatory elements between subgenomes [14, 15] though this aspect is very far from being well understood. *C. bursa-pastoris* also has considerable morphological variation, including the floral phenotypes that are not found in *A. thaliana* and rarely—in other Brassicaceae [16, 17], thus being promising object for the study of development [18]. The main obstacle to using *Capsella bursa-pastoris* as a model object is the lack of a well-assembled genome. Despite the significant progress in sequencing, there is still no chromosome-scale assembly of *C. bursa-pastoris* genome; most analyses are done using mapping on a reference genome of *C. rubella*, one of the parental species, and further phasing into subgenomes [19]. This leads to biases due to the unequal efficiency of mapping of the reads belonging to the O (the one derived from *C. orientalis* parent) and R (derived from *C. rubella*/*grandiflora*) subgenomes.

This paper presents the assembly of the *Capsella bursa-pastoris* genome down to the chromosome level, obtained by a combination of HiFi Pacbio reads, HiC, and high-resolution genetic mapping. The genome-wide analysis showed that this species has single origin based on the presence of homeologous exchange between chromosomes from O and R subgenomes shared between three isolated accessions.

Results

Our approach to the construction and correction of the assembly is a multistep procedure that includes the integration of chromosome conformation capture (HiC) data and the genetic map (Fig. 1). At first step, we obtained contigs from Pacbio CCS reads. Total length of this initial assembly was 363,941 kb with N50 ~ 1251 kb. Since the subgenomes have very high sequence similarity, we may expect the chimeric contigs combining regions from homeologous chromosomes. Thus, we performed an assembly check using the data on segregation of markers in F2 population derived from the cross of two polymorphic accessions—lel and msu-wt (Additional file 1: Fig. S1) [14, 17]. These data, obtained using WGS, contained ~ 254,000 SNP markers with average distance between markers of 864 bp. We analyzed the character states for these markers in 50 plants from F2 population (see example on Additional file 1: Fig. S2). If more than two recombinations were observed between two adjacent markers, the sequence containing these markers was treated as misassembly (see example on Additional file 1: Fig. S3). The contig/scaffold was then split and the sequence between markers was removed from the assembly. We found 71 such cases; in most of the cases, the number of observed (false) recombinations was more than 25%—that supports the markers are in fact not linked and are inherited independently. As a result, N50 reduced to 1063 kb. Then we proceeded to scaffolding using HiC data. The HiC reads are short and thus frequently cannot be unambiguously assigned to one of the subgenomes for many genome regions. In order to overcome this limitation, we again used genetic map data and assigned contigs to linkage groups. Four hundred ninety contigs were assigned to 16 linkage groups corresponding to 16 chromosomes typical for *C. bursa-pastoris*. N50 of contigs assigned to linkage groups was 1312 kb and total length of the assembly ~ 319 Mb. After that, we performed scaffolding on each linkage group separately. Most contigs were scaffolded; for only four short contigs together making 0.5 Mb, the scaffolding was not successful. After this we mapped HiC data on the chromosome scaffolds and performed visual examination and check of the assembly quality. Based on this check, we removed the regions which showed irregular or ambiguous position based on HiC map (Additional file 1: Fig. S4). These regions mostly correspond to the repeat-rich fraction of the genome. At this step, we removed 66 Mb from the assembly. The resulting assembly consisted of 16 chromosomes with N50 16.6 Mb and total length ~ 253 Mb. This is less than the size estimated based on cytofluorometry [20]; however, as we show below, it contains the overwhelming majority of genes.

At the next step, we assigned chromosomes to subgenomes. To do this, we applied approach based on the mapping of the reads from parental species *Capsella rubella* and *Capsella orientalis*. In order to assess its validity, we performed a simulation of *C. bursa-pastoris* genome using the sequences of parental genomes: we combined sequences of *C. rubella* and *C. orientalis* genomes into single reference and then mapped the reads

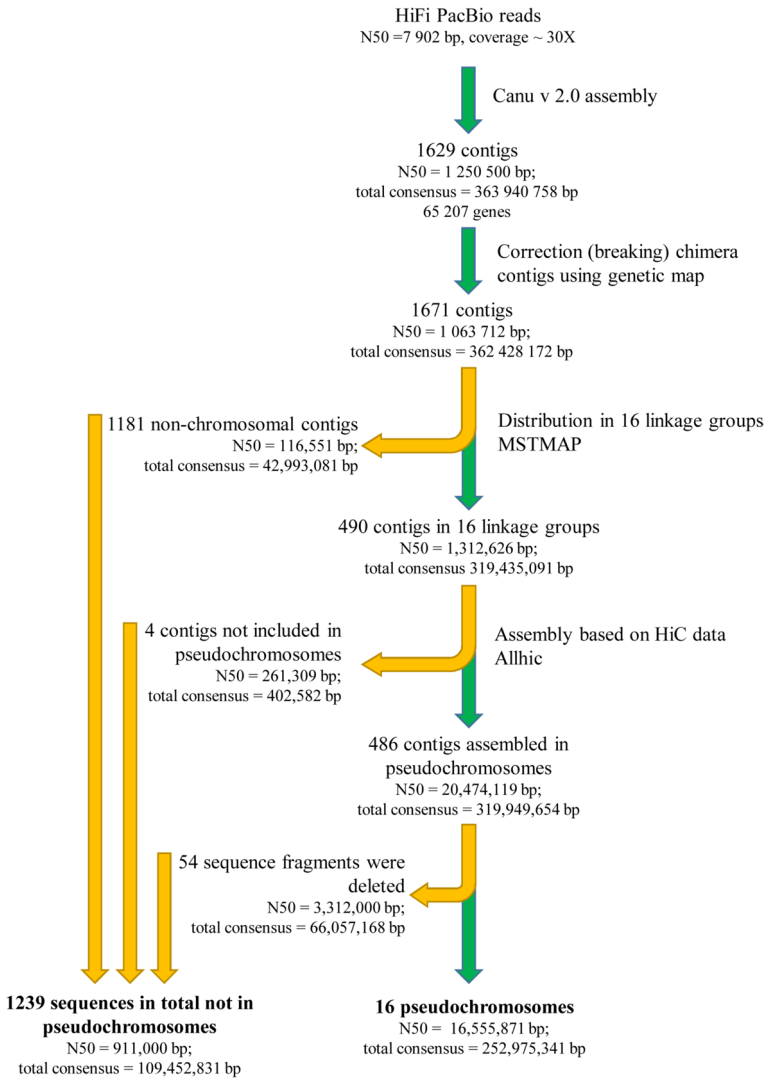

Fig. 1. Assembly pipeline for the *Capsella bursa-pastoris* genome

of each of the species. The analysis of mapping rates showed that for the most part the reads from each species are mapped on the subgenome where they belong to, not on the subgenome from other parent, despite their high similarity (Additional file 1: Fig. S5). This allows using mapping rate as a proxy for the determination of the origin of chromosomes from either R or O parent. Thus we mapped the reads of *C. rubella* and *C. orientalis* on *C. bursa-pastoris* chromosomes and calculated coverage on 10-kb windows (Fig. 2a). This showed that most of the chromosomes are derived from one parent exclusively (i.e., there were no recombination between R and O subgenome after the formation of a hybrid) with exception of two chromosomes that carry an evidence of homeologous exchange between R and O subgenomes chromosomes (these two chromosomes are termed here and further O7_R7 and R7_O7). The genetic map also supports the hybrid nature of these chromosomes.

In order to perform the quality control of the assembly in terms of completeness of gene set, we estimated the occurrence of all annotated genes and the genes from BUSCO set in different fractions of the assembly: the initial assembly, the subgenomes of the final assembly and the sequences that were deleted at different steps of the curation (for example, the misassembled regions and repeat-rich regions). This analysis demonstrated that the fraction of contigs not included in the final (pseudochromosomes) assembly contained less than 5% of BUSCO genes and less than 2% of all genes (Fig. 2b, Additional file 2: Table S1). This means that the fragments not

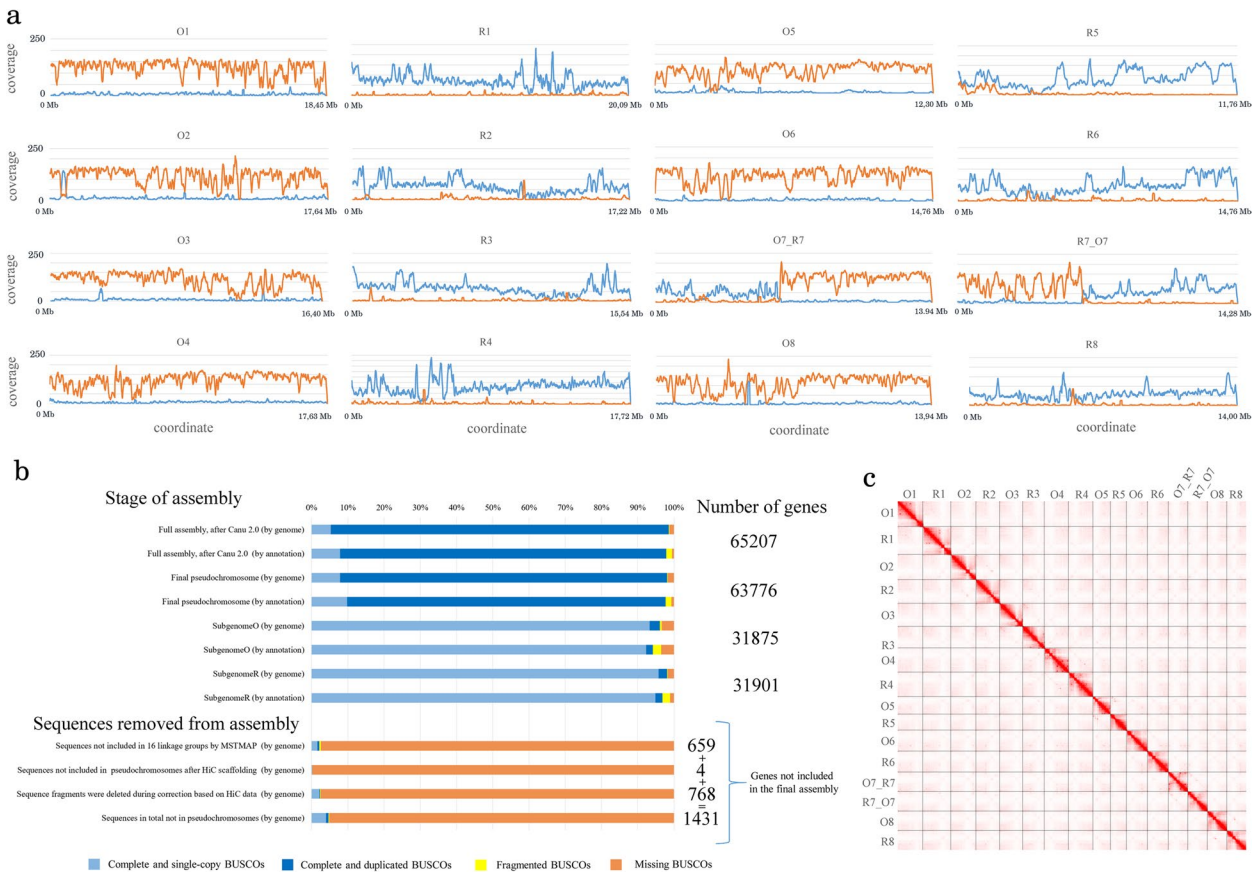

Fig. 2. *Capsella bursa-pastoris* genome assembly. **a** Coverage of *C. bursa-pastoris* chromosomes by the reads of *C. orientalis* (orange) and *C. rubella* (blue). **b** Analysis of the representation of BUSCO genes and all protein-coding genes in chromosomes and in the fragments of genome not included in chromosomes. **c** HiC map of *C. bursa-pastoris* chromosomes

included in the final assembly correspond to the gene-poor fraction of the genome (presumably pericentromeric regions) and do not affect any further analysis unless it is focused on repetitive elements. Another approach to quality control, aimed at the inference of the correctness of the assembly is based on the HiC data. Visual exploration of the contact map shows that it has typical diagonal patterns with no regions that may indicate on misassembly (Fig. 2c). After obtaining corrected assembly and checking its quality, we annotated it using a combination of homology based and de novo prediction approaches. The annotation included 65,207 protein-coding genes, 63,776 (97.8 %) out of which are located in the chromosome scaffolds and are evenly distributed across subgenomes (31,875 in the subgenome O and 31,901 in the subgenome R). In order to facilitate further studies employing this annotation, the genes were put into correspondence with orthologous *Arabidopsis thaliana* genes. The gene naming reflects this correspondence following the format: (number of chromosome).(number of gene in *Capsella*).(number of gene in *A. thaliana*)_(type of search of a homologous gene – Orthofinder or blast), for example, «Cbp.O1.g00001000.AT1G02140_O» for a gene that is located in the chromosome O1, is a first gene annotated on this chromosome and is orthologous to *A. thaliana* gene AT1G02140. The genes are numbered with space of 1000 in order to enable the addition of genes that might be found in other accessions of *C. bursa-pastoris* in further studies. In order to estimate the quality of annotation, we performed several tests, in particular, calculated BUSCO metrics for the set of predicted proteins (for all proteins together and in a subgenome-wise manner and performed comparison with *A. thaliana* in terms of such parameters as CDS length, the number of exons, and the percent of single-exon genes. This approach follows recently published recommendations for the improvement and quality control of plant genome annotations [21]. The BUSCO completeness score for all proteins together was in the range of 97.7–94.1%, for subgenomes—from 90.4 to 96.8% (Additional file 2: Table S1). The fraction of single-exon genes and the CDS

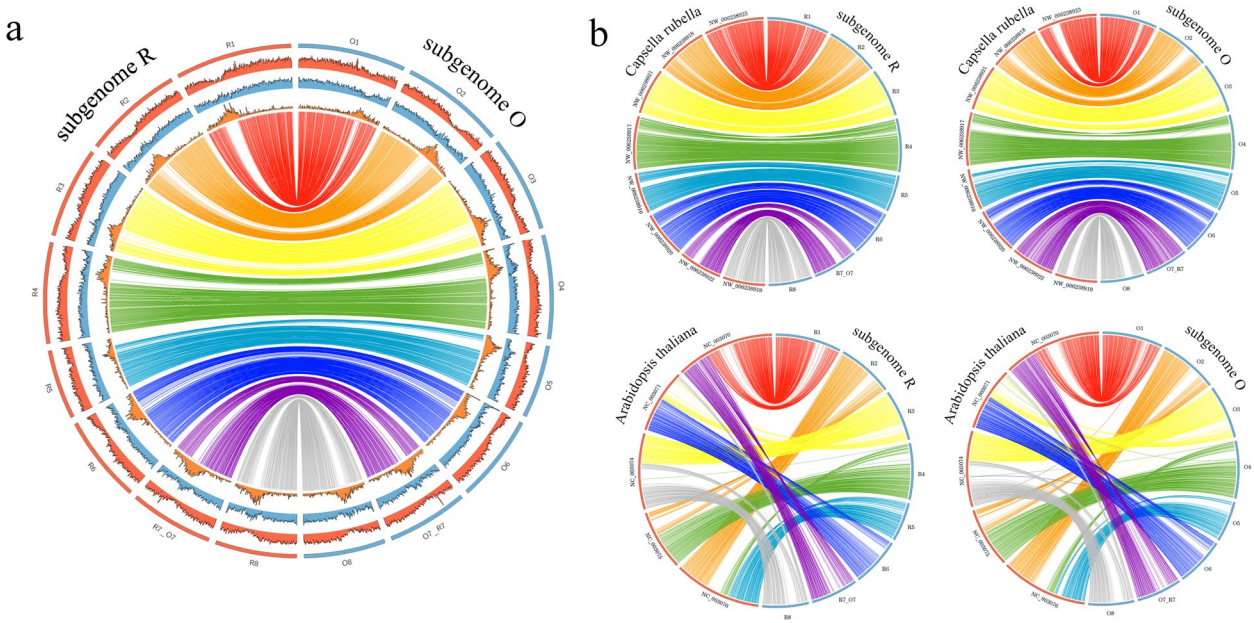

Fig. 3. Circos plots showing the comparison of *C. bursa-pastoris* subgenomes **a** between themselves and **b** with *C. rubella* and *A. thaliana*. On the panel **a**, the density of repeats is shown in orange, the density of genes in blue and GC content in red

length were close to ones in *A. thaliana* (Additional file 2: Table S2).

The R and O subgenomes are mostly colinear one with another and with the *C. rubella* parent (the only one for which chromosome-scale assembly is available). As whole-genome alignments show, there are no major rearrangements or indels between homeologous chromosomes (Fig. 3a). Many polyploids of hybrid origin have undergone biased fractionation of the subgenomes, where one of the subgenomes preferentially loses genetic material and other preferentially retains. This is clearly not the case for *C. bursa-pastoris*. This finding is congruent with the results stemming from the annotation: both subgenomes carry similar number of genes. The distribution of genes and repeats is also similar; the repeats have a sharp peak at a specific gene-poor region that presumably corresponds to a centromere. The only notable difference is the small fragment at the beginning of chromosome R5 for which there is no homologous fragment in O subgenome (neither in O5 nor in any other place of the subgenome). In order to ensure that this is not an assembly artifact, we checked its coverage in R5 and it does not deviate from the average across the genome; we also looked for the presence of highly similar sequences in the contigs of the initial assembly that were not included in the final one and did not find any. This region is found in *C. rubella* genome; this suggests that the absence in O subgenome is the result of deletion which occurred after the divergence *C. grandiflora/rubella* and *C. orientalis*. The *C. rubella* genome has a small inversion in the beginning of chromosome 4 relative to both R and O subgenomes (Fig. 3b). The fact that R and O subgenomes share the same orientation shows that it was ancestral while the inverted orientation found in *C. rubella* is lineage-specific and occurred during speciation of *C. rubella* after the formation of *C. bursa-pastoris*.

Both *C. bursa-pastoris* subgenomes retain strong similarity and colinearity with *A. thaliana* genome (Fig. 3b), with *A. thaliana* chromosomes aligning to several chromosomes of each of the subgenomes. This is in congruence with previous data on *C. rubella* [22, 23] and highlights two contrasting tendencies in the genome evolution in Brassicacae—whole-genome duplication and fusions of chromosomes with further reduction of non-coding DNA.

The availability of a subgenome-resolved genome sequence and a wide set of WGS data on *C. bursa-pastoris* and parental species derived from earlier studies [24, 25] allows to analyze the evolution of each of the subgenomes separately. To do this, we used the samples from species that is donor of the other subgenome as outgroup. For example, when considering O subgenomes, *C. rubella* and *C. grandiflora* were used as outgroup and vice versa. The reads of diploid species were mapped on only one subgenome—the one that was under consideration. For *C. bursa-pastoris* samples, we mapped reads on both subgenomes and then considered only the regions corresponding to one subgenome. Using this approach, we got 6,651,155 variable

positions (SNP, indels were not used) for subgenome R and 4,967,014 for subgenome O on a set of 56 samples of *C. bursa-pastoris*, 7 samples of *C. rubella*, and 21 for *C. grandiflora*, and 15 for *C. orientalis* (Additional file 2: Table S3, Fig. 4a). These genome-wide data on genetic variation were used to infer relationships between samples using phylogenetic trees and neighbor-net approach. All these analyses were carried out separately for R and O subgenomes. In terms of the relationships within *C. bursa-pastoris*, phylogenetic trees resulting from the analysis of O and R subgenomes are mostly congruent. Both show the separation of Asian (ASI) and Middle East (ME) populations recognized in earlier studies [24, 25]; the third group, European (EU) is recognized as monophyletic only in trees inferred from O subgenomes while in R-trees it is present as a grade while ASI and ME are sister clades (Fig. 4a). Neighbor-net analysis offers higher resolution and allows to get more detailed and accurate view of the relationships between *C. bursa-pastoris* populations. The networks inferred from different subgenomes are highly congruent (Fig. 4b, c) and show that *C. bursa-pastoris* is subdivided into six subgroups. Three of them correspond to the EU + ME populations and are found in Europe, with group 1 confined mostly to northern Europe and groups 2 and 3—to Eastern (the group 3 corresponds to ME). Other three occur in Asia (group 4—northern Kazakhstan and groups 5 and 6—China) (Fig. 4d). A notable feature of both networks and trees is the asymmetry concerning the position of *C. orientalis* relative to the subgenome O and *C. rubella*/*C grandiflora* to the subgenome R. For the R-lineage (with *C. orientalis* as a root), *C. rubella*/*C grandiflora* is an outgroup to R subgenomes. In case of O-lineage (*C. orientalis* and the subgenome O, *C. rubella*/*C grandiflora* is a root), it is different—the clade uniting *C. orientalis* samples is nested together with the subgroups of *C. bursa-pastoris*. Similar observations were done in earlier studies [19] and two hypotheses were put up: the multiple origin of *C. bursa-pastoris*, independent origin of Asian *C. bursa-pastoris* and a present-day *C. orientalis* being closer to the ancestor of Asian group and the introgression from *C. orientalis* to the O subgenome. The availability of a genome sequence phased into subgenomes allows us to test those hypotheses and to put up an alternative one (see "Discussion"). As mentioned above, we found a homeologous exchange between chromosome 7 of *C. orientalis* origin and that of *C. rubella/grandiflora* origin. It led to the formation of two hybrid chromosomes, termed R7_O7 and O7_R7. The accession that we use as a source for reference (msu-wt) belongs to the EU population [26] or "group 1" according to the classification presented above. To test whether the hybrid chromosomes are found in Asian and Middle East groups or it is a feature unique for this accession, we performed the analysis based on chromatin contacts data acquired using HiC method [27]. We constructed and sequenced HiC libraries for representatives of Asian (Cbp_ASI) and Middle East (Cbp_ME) groups and mapped them on msu-wt reference. For both Cbp_ME and Cbp_ASI, contact maps for R7_O7 and O7_R7 demonstrate a clear diagonal pattern with no gaps and/or alternative joins. This indicates on the absence of structural changes between the reference genome and the genome from which HiC data are derived (Fig. 5a). In order to control whether this analysis is able to reveal such changes, we made a simulation of translocation (which will correspond to the absence of exchange between R7 and O7) by artificially modifying the reference and mapped HiC data to this modified reference. In this case, the contact map clearly shows the gap in contacts which is an indicator of the discordance between the reference and the HiC data (Fig. 5a, right upper panel). As an additional evidence of the colinearity of genomes from different populations of *C. bursa-pastoris* and the presence of hybrid chromosomes, we performed a crossing of accessions belonging to the populations 3 and 1 (the latter is the same accession msu-wt that was used for the genome assembly). We analyzed 30 F2 plants derived from two crosses and calculated the number of recombinations per chromosome (Fig. 5b). In case if the chromosome 7 had different structure, we would have observed decreased recombination in these chromosomes. Taken together, these analyses provide the evidence of that hybrid chromosomes exist in all three major clades of *C. bursa-pastoris*.

Another important question is the transfer of genetic material between populations and from parental species. In order to assess it, we analyzed population structure based on genome-wide search of SNPs. This was done separately for each of the subgenomes. Under k=4 (the optimal value as estimated by fastStructure), we did not find the evidences of widespread introgression from either parent to modern *C. bursa-pastoris* populations (Fig. 6). The most pronounced introgression observed was that of *C. orientalis* to O subgenome. Based on the results of fastStructure, it occurred in few samples from Asian populations (groups 4 and 5) which come from the areas overlapping with the areas of *C. orientalis* distribution [9]. Previous reports also detected introgression though the details differ (will be discussed further). Surprisingly, the analysis of population structure revealed the signs of introgression from *C. bursa-pastoris* to parental species. This is highly unexpected because it requires the hypothesis on differential elimination of one of the subgenomes in gametes; though such cases are known in some species [28], we think that this might also stem from technical issues (like contamination by *C. bursa-pastoris* material because these species are very hardly distinguishable) rather than from biological phenomena. Within

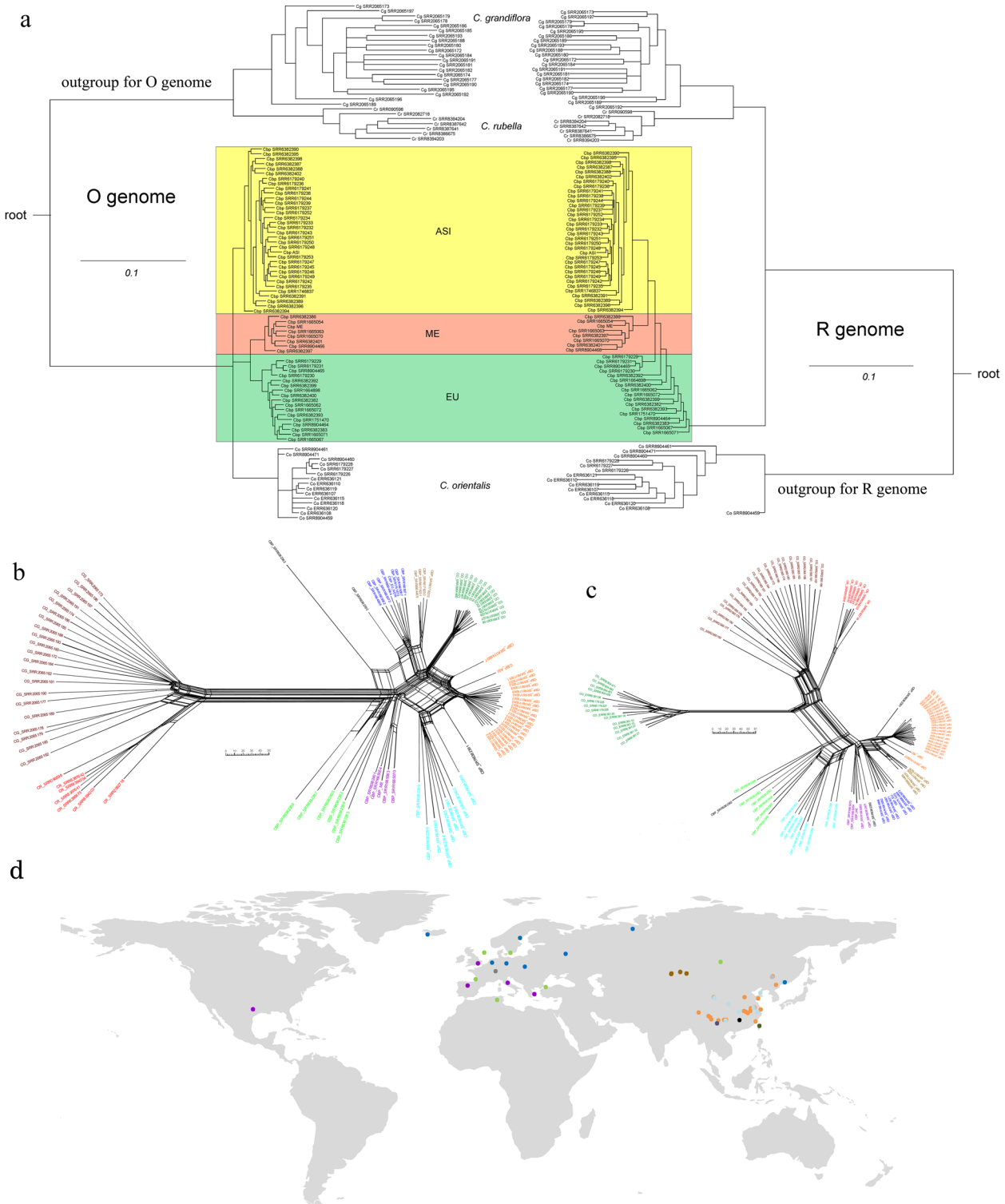

Fig. 4. Populations of *C. bursa-pastoris*. **a** Tanglegram of phylogenetic trees by subgenomes O and R. Colors indicate samples of groups ME, ASI, and EU. **b** Neighbor-net graph constructed for subgenome O. **c** Neighbor-net graph constructed of subgenome R. **d** Geographical map with the location of populations. The six subgroups are each marked with their own color and number. Accessions belonging to the same groups on both trees are marked with the same color. Samples with an unstable position are highlighted in black. The color is the same on a geographic map and on graphs

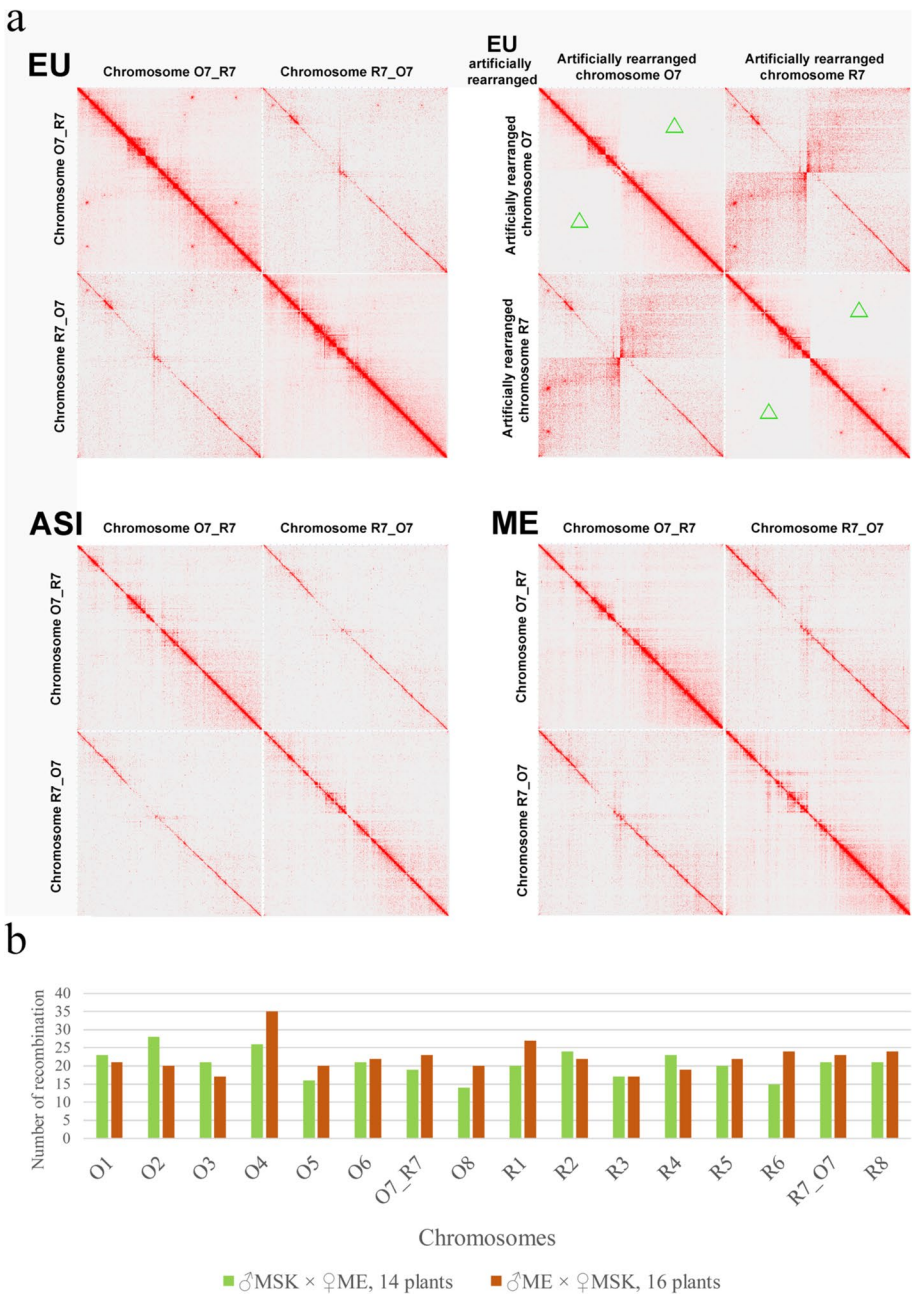

Fig. 5. Homeologous exchange between chromosomes 7 from O and R subgenome. **a** HiC maps for chromosome 7 for plants belonging to different populations (area without contacts in the analysis of the artificial chromosome is indicated by green triangles). Non-unique read mapping was allowed, in order to visualize the synteny of paralogous chromosomes. **b** The number of recombinations in chromosomes detected in F2 from crossing line msu-wt (EU population) and line Cbp_ME (ME population)

C. bursa-pastoris, the gene flow between populations was also moderate. Using admixture analysis under $K = 6$ (corresponds to the number of groups distinguished within *C. bursa-pastoris*), we found that R subgenomes (Additional file 1: Fig. S6) do not demonstrate any stratification while for O subgenomes, two groups are separated (corresponding to EU + ME and Asian populations). Within these groups, several samples show admixture with European populations. These samples correspond to group 5 (Fig. 4b) of the Asian clade. This might however be explained not only by introgression but also the retention of ancestral polymorphisms because this population represent the basal clade of the Asian group. In any case, there is

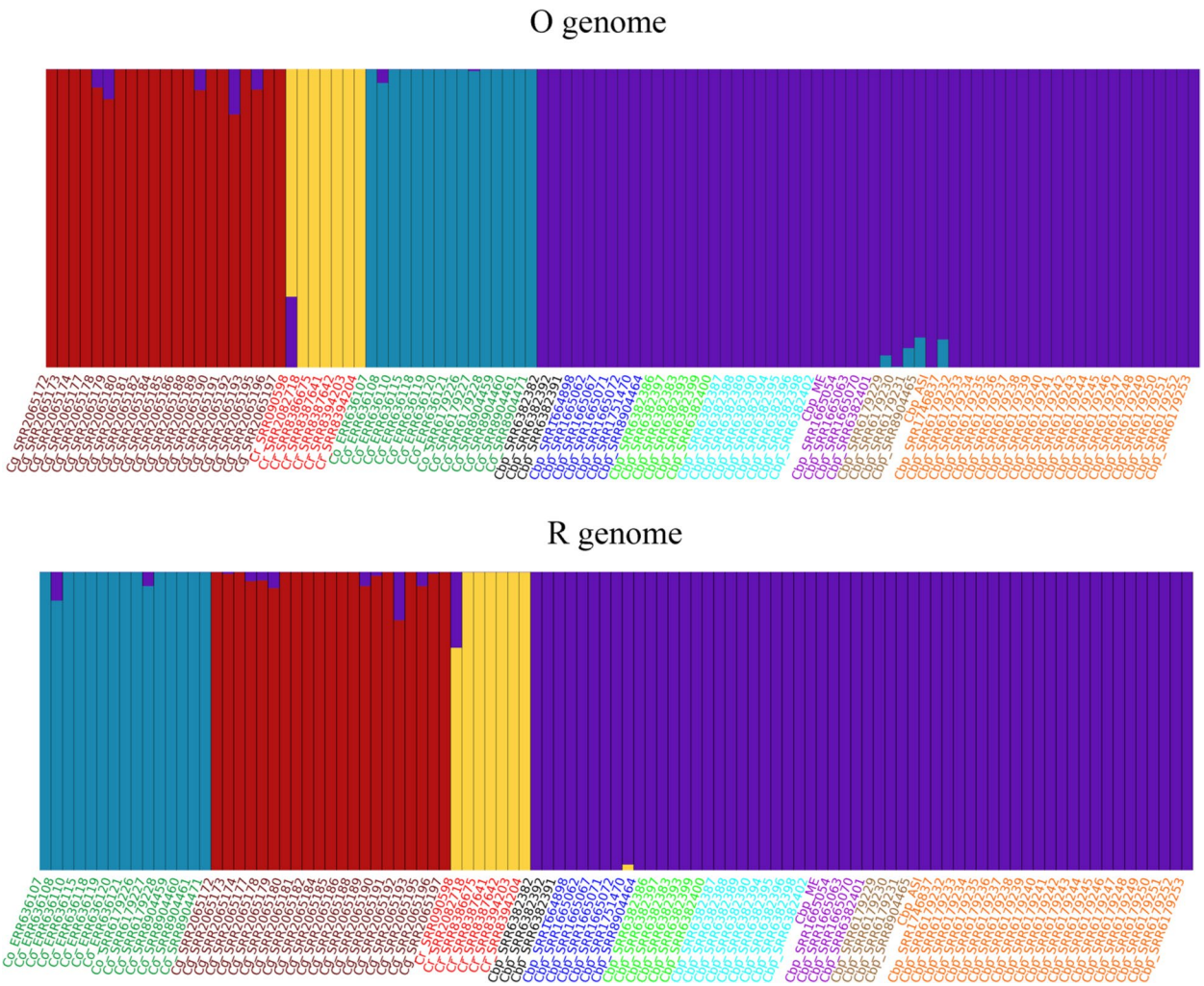

Fig. 6. Analysis of introgression by admixture analysis, for *K*=4. The colors of the line names correspond to the populations in Fig. 4b

no massive admixture between populations. The analysis by the alternative tool, D-suite [29], also revealed introgression from *C. orientalis* to *C. bursa-pastoris*. In the analysis of O subgenome, the introgression is the most pronounced for Asian populations (Additional file 1: Fig. S7a-c). Regarding the introgression within *C. bursa-pastoris* populations, D-suite provides results that are only partially congruent with that of fastStructure; these results suggest the introgression between ME and ASI populations (Additional file 1: Fig. S7d-f), while fastStructure shows the admixture between group 5 (Asian population) and European populations (see above). It should be noted however that these are not robust relative to the tree topology. The analysis of R subgenome also indicates on some introgression from *C. grandiflora/rubella* (Additional file 1: Fig. S8a-c), though much less pronounced than for *C. orientalis*. It also shows the introgression between populations of *C. bursa-pastoris* (Additional file 1: Fig. S8d-f), in contrast to the results of fastStructure.

Discussion

C. bursa-pastoris has long been a complex object. There was a long-standing controversy about its origin (either by allo- or by autopolyploidy [30–32]) which was solved only when genome-scale data came into practice [7, 14]. However, many important questions concerning the evolution of this species remained. Some of these questions were tackled using a genome of one parental species—*C. rubella*—as a reference for phasing-based analysis [25]. Our analysis of chromosome-scale assembly allowed to characterize the structural features of each of the subgenomes. We found that they are mostly colinear and carry similar number of genes. In contrast to what is seen in many other polyploids of hybrid origin [33], there is no biased fractionation leading to the subgenome

dominance. This is congruent with our earlier findings on the absence of asymmetry in gene expression profiles between subgenomes [14].

An important finding from the whole-genome analysis is the evidence of the homeologous exchanges (HE) between R and O subgenomes at early stages of *C. bursa-pastoris* evolution that led to a formation of two hybrid chromosomes. The homeologous exchanges are frequent in allopolyploids [34]; however, the fixation of each individual exchange event is rare and the fact that HE event is shared is a strong evidence of common ancestry [35]. The question on a single or multiple origin of *C. bursa-pastoris* is debatable; an independent origin was hypothesized, especially for Asian clade, though inconclusively [25]. We showed that the representatives of ME and Asian clade also have hybrid chromosomes, similar to msu-wt accession belonging to European clade. Though we cannot completely rule out the possibility that there are some accessions that lack HE or that the HE occurred independently in different clades, the current results strongly support the single origin of *C. bursa-pastoris*. Further study of *C. bursa-pastoris* accessions at a pangenome level would corroborate this.

Notably, the phylogenetic trees and networks inferred from R and O subgenomes are discordant relative to the position of *C. orientalis* and *C. rubella*/*C grandiflora* samples. These distinct evolutionary trajectories could be the evidence of multiple origin [25]. However, in view of the presence of HE in different populations of *C. bursa-pastoris*, this scenario is unlikely. Alternative explanation also put out in the earlier studies is the massive introgression from *C. orientalis* to *C. bursa-pastoris*. However, we showed that this is not the case as well since the introgression is limited. Taking into account our findings stemming from whole-genome analysis, we put out another hypothesis for these contrasting phylogenies. The origin of *C. bursa-pastoris* has been for a long time a controversial topic (see, e.g., [30–32, 36]). Current hypothesis based on the genomic data implies that the *C. bursa-pastoris* is a hybrid of *C. rubella*/*C grandiflora* (paternal parent) and *C. orientalis* (maternal parent). The support for this hypothesis was first provided in [7] and it was corroborated in [14, 31]. We suggest here a modification of this hypothesis that implies that *C. rubella*/*C grandiflora* is not a direct progenitor of *C. bursa-pastoris* but a sister species that diverged from a species that is a direct progenitor and is now extinct or not identified (Fig. 7). This is supported also by the higher divergence between Cr/Cg and R subgenome than between Co and O subgenome.

The genome of *C. rubella* was by now the only one assembled up to chromosome scale; the genome of *C.*

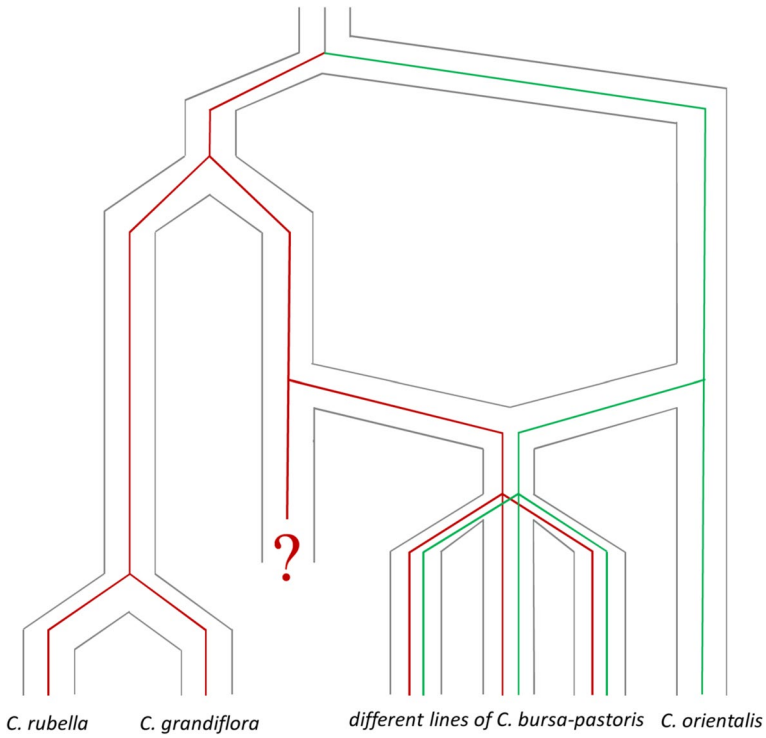

Fig. 7. Evolution of the two subgenomes of the tetraploid *Capsella bursa-pastoris* (Brassicaceae) and modified hypothesis on the origin of *Capsella bursa-pastoris*

orientalis is available only as a set of short contigs (N50 = 25 kb) [37] and is not used for any studies involving *C. bursa-pastoris*. Instead of this, the multistep pipeline involving mapping of *C. bursa-pastoris* reads on *C. rubella* genome and further phasing of SNP is used [25]. Given the higher divergence between *C. rubella* and R subgenome and the presumable non-parental relationships of *C. rubella/C grandiflora* and *C. bursa-pastoris*, this might obfuscate the results of the genomic and transcriptomic analysis based on this pipeline. The availability of the chromosome-scale assembly of *C. bursa-pastoris* genome thus opens the avenue to the unbiased view of the genome evolution of this plant and provides a reference for further omics-related studies in this emerging model species.

Materials and methods

Plant material, DNA extraction, library preparation, and sequencing

For PacBio long-read sequencing, fresh leaves of *C. bursa-pastoris* accession 'msu-wt' [14, 17] was used. This accession comes from Moscow, Russia, growing in vicinity of Lomonosov Moscow State University (the same accession was used in our previous studies under the names wt-msu, Msk or wt-msc-1). The leaves were quick-frozen in liquid nitrogen and transferred on dry ice to the DNA Link laboratory (South Korea, Seoul). DNA was isolated, and SMRTBell libraries were prepared and sequenced by the DNA Link laboratory. For WGS short-read sequencing which was performed for two accessions, termed here as Cbp_ASI and Cbp_ME, DNA was extracted from fresh leaves using CTAB method [38]. Shotgun libraries were prepared using the NEBNext Ultra II DNA library preparation kit (New England Biolabs, Ipswich, MA, USA) and sequenced on Illumina HiSeq2000 (Illumina, San Diego, CA, USA) platform using HiSeq SBS Kit (200 Cycles).

For chromatin capture-based Hi-C sequencing, the nuclei enrichment and isolation was performed on fresh leaf material using CelLytic PN Isolation/Extraction Kit (Sigma-Aldrich). DNA was extracted from plant cells and Hi-C libraries were prepared using EpiTect Hi-C Kit (Qiagen, Germany) according to the manufacturer's protocol. Hi-C libraries were sequenced on Illumina NextSeq500 (Illumina, San Diego, CA, USA) platform using High Output Kit v2.5 (150 Cycles).

The quantity of extracted DNA and prepared libraries were measured with Qubit (Thermo Fisher Scientific, Waltham, MA, USA) DNA assays. The size and quality of prepared libraries were validated by Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA) DNA fragment analysis and qPCR.

Genome assembly

Draft genome assembly of the *C. bursa-pastoris* genome was constructed from CCS PacBio reads (the average read length was 7948 bp) obtained from the DNA Link laboratory using the Canu software v. 2.0 [39] with the following parameters: genomeSize=300m—estimated genome size; correctedErrorRate=0.005—the allowed difference in the overlap between two reads is no more than 0.5%; minOverlapLength=1000—minimum read overlap length.

The reads of 50 samples of the F2 generation and one of the parent plants of the F0 generation used for the genetic map were mapped to the assembly, and SNP calling using the CLC Genomics workbench software v.20.0.3 (QIAGEN) was performed to verify and correct the assembly (see Additional file 1: Fig. S9). Based on the SNP data of the F0 generation parent, the database of homozygous SNP markers (245,060 markers in total) was constructed and used for the downstream analysis (see Additional file 1: Fig. S10). The allele status within each of the F2 plants was tested based on SNP calling results and marker coverage information for each marker. The results of this check were compiled in tables where rows correspond to the genetic markers of the database based on F0 sequencing results, columns correspond to the F2 plants, and each table cell contains information about the marker state. Data collection for the genetic map, SNP database construction, and data processing in detail is presented in Additional files 1,8,9. Marker states were as follows:

- "-1"—homozygous allele corresponding to the reference allele. The assembly from sequencing data of the first parent of the F0 plants (msu-wt) of the genetic map was used as a reference.
- "0" —heterozygous
- "1" —homozygous allele corresponding to the homozygote from the second parent of the F0 plants of the genetic map. The marker database has been created from the data on the homozygotes of this plant.
- "-"—unidentified marker.

Before searching for incorrectly assembled regions, the table with the states of the genetic map markers was filtered. Markers that were unidentified in more than half of the samples were removed. Markers that did not satisfy Pearson's criterion for the ratio of homozygotes to heterozygotes were also removed. The threshold for the criterion was chosen as 5.99, corresponding to a p-value < 0.05. Next, the filtered marker status table was manually processed for areas with a high recombination frequency. These areas have been removed from the

assembly. Markers belonging to deleted regions have also been removed from the marker state table.

At the next stage of data preparation for constructing a genetic map, the averages for markers within the contig were calculated for each contig and each sample. Based on the averages, a new table of marker states was compiled, in which contigs acted as markers. The following thresholds and a code corresponding to the input data of the MSTMAP software v. 1.0 [40] were used to transfer information about the average values of the states of markers in each specific position into the state of contig markers:

- "A"—homozygous allele corresponding to the reference. This state was assigned to a marker if the average for the contig corresponding to the marker in the sample was from −1 to −0.8, including the ends of the interval.
- "B"—homozygous allele corresponding to the homozygote from the second parent F0 plant of the genetic map. This state was assigned to a marker if the average for the contig corresponding to the marker in the sample was from 0.8 to 1, including the ends of the interval.
- "X"—heterozygote. This state was assigned to a marker if the average for the contig corresponding to the marker in the sample was from −0.2 to 0.2, including the ends of the interval.
- "-"—undefined marker. This state was assigned to a marker if the average for the contig corresponding to the marker in the sample was from −0.8 to −0.2 or from 0.2 to 0.8.

The resulting table of markers was processed in the MSTMAP software with the following parameters: "cut_off_p_value 0.0000001, no_map_dist 100.0, no_map_size 1, missing_threshold 0.5, population_type RIL2." As a result of this program, a set of 15 linkage groups was initially obtained. For 5 linkage groups, the analysis was restarted by the MSTMAP software with the changed parameter "cut_off_p_value 0.0000000001." As a result, 16 linkage groups were obtained; this corresponds to the haploid chromosome number in *C. bursa-pastoris*.

To scaffold contigs using HiC, HiC reads were mapped onto a set of filtered contigs. Mapping was carried out without taking into account paired readings using the CLC program. For the obtained alignments, the information about the pairing of reads and information about the links was restored using the Arima mapping Pipeline [41]. The obtained linkage information was further divided into linkage groups and separately processed using the AllHiC program [42], which was modified to use the information about the order of contigs in the linkage group. As a result, an assembly was obtained, consisting of 16 HiC scaffolds and contigs not included in them. Visualization of HiC maps was carried out using the JuiceBox program [43]. Based on the analysis of maps for each linkage group, scaffolding errors were identified and corrected.

Genome annotation

Annotation was the multistep procedure that started from two annotations—one was done using BRAKER v. 2.1.5 [44] with parameters "--esmode --softmasking" and the second—using liftoff v 1.6.1 [45]. Liftoff is a tool that can transfer annotations from closely related species (*Arabidopsis thaliana*, annotation TAIR10 in this case). The transfer was done separately for R and O subgenome, then the annotations for subgenomes were merged into one GFF file and all genes that did not contain valid ORF were removed using script RemoveNotValidORFsFromGffFile.pl. Then we added to this annotation the gene models generated with BRAKER using a script MergeGFFFiles.pl (only those gene models that did not overlap with filtered Liftoff annotation were added). For all genes annotated at this step, CDS and corresponding amino acid sequences were predicted. The amino acid sequences were used for the inference of orthology with *A. thaliana* (TAIR10 proteins) using Orthofinder v 2.4.0, separately for each of the subgenomes. Then we considered orthogroups that contained one or several *C. bursa-pastoris* genes from either R or O subgenome and exactly one *A. thaliana* gene. This allowed to make a correspondence between *Arabidopsis* gene and *C. bursa-pastoris* genes; this correspondence was included in the gene name in order to facilitate further analyses. The genes that were not the part of the orthogroups with single *A. thaliana* genes (the genes from *C. bursa-pastoris*-only orthogroups, or orthogroups with more than one *A. thaliana* gene or singletons) were the subject of blastp search against TAIR10 proteins. The best hit was taken as a corresponding and its identifier was included in the *C. bursa-pastoris* gene name. At the next step, we searched for the pairs of homeologs. In order to identify them, we first performed blastn search of all CDS from a chromosome (or part of the chromosome, in case of hybrid chromosomes R7_O7 and O7_R7) belonging to R subgenome against all CDS from homeologous chromosome of the O subgenome and vice versa. Then using a script GetListOfSequenceGenesForGFFFileWithSequenceOfHomeologous.pl, we compiled a list of pairs, using the information on the best blast hit and the position of a gene in the chromosome (i.e., if for example, a certain gene from R

subgenome has two hits in the homologous chromosome of the O subgenome, the gene with the position that better corresponds to the position in the R subgenome is considered as a homeolog). For the gene for which the homeologs were not found, we performed an additional round of gene prediction using liftoff. In this case, liftoff transfers the annotation from a gene that was found but left without pair to a subgenome where the homeolog of this gene should be but was not found. For example, let us consider the genes O1, O2, and O3 from the subgenome O; O1 and O3 have pairs of homeologs R1 and R3 correspondingly, and O2 does not have a homeolog. In this case, we use liftoff to predict a gene in a region that is located between R1 and R3, giving it a hint that this gene should be highly similar to O2. After this, the genes that did not contain valid ORF were filtered using script RemoveNotValidORFsFromGffFile.pl and merged with the main annotation using script MergeGFFFiles.pl. Then the inference of orthology, BLAST search and the search of homeologs were run again with the set that includes these additionally discovered genes. Based on the information on homeologous relationships between R- and O subgenome genes and on the similarity with *A. thaliana* genes, we named the genes in *C. bursa-pastoris* annotation using script ReformatGffFile.pl.

Phylogenetic trees and networks

For the inference of relationships between *C. bursa-pastoris*, *C. rubella*/*C. grandiflora* and *C. orientalis* and within *C. bursa-pastoris*, we mapped data from multiple individuals of these species on our assembly of *C. bursa-pastoris* genome. The data included sequences from earlier publications [24, 25] and generated in this study. In order to provide reliable results of SNP calling, we selected only those samples that had sequencing depth more than 10x and that represented individual plants, not pools of several neighboring plants. The data for *C. rubella*/*C. grandiflora* and *C. orientalis* were mapped separately on R subgenome and O subgenome, and for *C. bursa-pastoris*—on both subgenomes. For mapping and SNP calling, we used the tool CLC Genomics Workbench (Qiagen) (parameters identical described on Additional file 1: Fig. S9). Mapping and SNP calling setting were the same as used for the construction of genetic map (Additional file 1: Fig. S2). Then the lists of SNP were retrieved for each individual (for *C. bursa-pastoris* samples—for each subgenome separately) using the script CreateDBList.pl. All SNPs found in at least one sample were added to the list. Then the sequencing depth was calculated for each SNP from the list using samtools depth. The SNPs with sequencing depth less than 4 were replaced by N; heterozygous SNPs were replaced by corresponding letters denoting degenerate bases (Y for C/T, R for A/G, etc.). These data were used for the construction of pseudoalignment. For each sample, the sequence was constructed based on the reference sequence with replacement of reference positions carrying SNP by the nucleotide corresponding to the non-reference nucleotide found in this sample. For this, we used a custom script CreatePseudoAlignmentForRegion.pl. Separate pseudoalignments were constructed for each chromosome arm. Phylogenetic trees were constructed for each pseudoalignment using RAXML v. 8.0.26 with parameters "-m GTRCAT -T 100 -x 123456 -N 100 -p 092345."

Admixture analysis

VCF files with filtered SNPs obtained from the variant calling on the subgenome R and O of the reference genome were merged into two multisample VCF files for each subgenome separately using bcftools v.1.16 [46] (*merge* with option *--missing-to-ref*). VCF files contained samples of *C. rubella*, *C. grandiflora*, *C. orientalis*, and three major populations of *C. bursa-pastoris*, 102 samples in total in each file. Then VCFs were converted to plink binary format using plink software v1.90b6.21 [47, 48] to infer the population structure using fastStructure v.1.0 software [49], which accepts genotypes in plink binary format as input. Admixture proportions were calculated with or without parental species of *C. bursa-pastoris* with the *K* value (number of populations) from 1 to 10. The optimal model complexity was estimated using the chooseK Python script of the fastStructure software package. Admixture proportions inferred by fastStructure were visualized using distruct v.2.3 Python script [50].

We also performed analysis for the presence of introgression between the parental species and different *C. bursa-pastoris* lineages, as well as between the lineages themselves using the Dsuite software [29]. D-statistics and f4-ratio were calculated separately for subgenomes O and R. Parental species were used as outgroups: *C. grandiflora/C. rubella* for the O subgenome to detect introgression between *C. orientalis* and *C. bursa-pastoris* lineages, or using *C. orientalis* as an outgroup to assess introgression only between *C. bursa-pastoris* lineages; for the R subgenome, the same approach was used but the outgroup was *C. orientalis* or *C. grandiflora/C. rubella*, respectively. Since the topology of trees showing relationships between *C. bursa-pastoris* lineages is not completely resolved and D-suite does not allow polytomies, we performed analysis with different tree topologies corresponding to different ways of the resolution of polytomy.

Abbreviations

ASI Asian *C. bursa-pastoris* population

EU European *C. bursa-pastoris* population
HE Homeologous exchanges
ME Middle East *C. bursa-pastoris* population
O *C. bursa-pastoris* subgenome derived from *C. orientalis* parent
R *C. bursa-pastoris* subgenome derived from *C. rubella/grandiflora* parent
WGS Whole-genome sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-01832-1.

Additional file 1: Fig. S1. Scheme for data acquisition for the genetic map. **Fig. S2.** An example of a contig fragment with a colored state of the markers. Due to the low coverage some of the markers are "noisy". **Fig. S3.** An example of chimeric assembly. Two adjacent markers located at a distance of ~90 kbp have 33 "recombinations" per 100 chromosomes, which is impossible and indicates independent inheritance of markers.**Fig. S4.** An example of correction of a local chimeric assembly. a Insertion of a foreign fragment(s) and b view of the site after correction. The green dashed lines show the signals indicating the proximity of the sites. **Fig. S5.** Simulation of the subgenome separation procedure. Example of the coverage by reads of the parental species of some reference contigs created from the genomes of *C. orientalis* and *C. rubella* for subgenome separation. **Fig. S6.** Analysis of introgression by admixture analysis in *C. bursa-pastoris*, for K=6. The colors of the line names correspond to the populations in Fig. 4b.**Fig. S7.** Fbranch matrix plotted using Dsuite f4-statistics results for different tree topologies of parental species and lineages of *C. bursa-pastoris* for subgenome O. a for ((((ME,EU),ASI),Co),Cgr-Outgroup) tree; b for ((((ASI,EU),ME),Co),Cgr-Outgroup) tree; c for ((((ASI,ME),EU),Co),Cgr-Outgroup) tree; d for (((ME,EU),ASI),Co-Outgroup) tree; e for (((ASI,EU),ME),Co-Outgroup) tree; f for (((ASI,ME),EU),Co-Outgroup) tree. Co – *C. orientalis*, Cgr – *C. rubella/C. grandiflora*, and ASI, ME, EU – lineages of *C. bursa-pastoris*. **Fig. S8.** Fbranch matrix plotted using Dsuite f4-stastics results for different tree topologies of parental species and lineages of *C. bursa-pastoris* for subgenome R. a for ((((ME,EU),ASI),Cgr),Co-Outgroup) tree; b for ((((ASI,EU),ME),Cgr),Co-Outgroup) tree; c for ((((ASI,ME),EU),Cgr),Co-Outgroup) tree; d for (((ME,EU),ASI),Cgr-Outgroup) tree; e for (((ASI,EU),ME),Cgr-Outgroup) tree; f for (((ASI,ME),EU),Cgr-Outgroup) tree. Co – *C. orientalis*, Cgr – *C. rubella/C. grandiflora*, and ASI, ME, EU – lineages of *C. bursa-pastoris*. **Fig. S9.** F2 data processing. **Fig. S10.** Building the SNP Database.

Additional file 2: Table S1. BUSCO metrics for assembly and annotation. **Table S2.** Statistics of the protein-coding genes annotation in *C. bursa-pastoris*. **Table S3.** Set of samples used for genetic variation analysis.

Acknowledgements
Not applicable.

Authors' contributions
AAP conceived and coordinated the study and participated in the acquisition of experimental data, genome assembly, and writing of the manuscript, ASK performed genome assembly and participated in population genomics analysis, AVK participated in population genomics analysis, managed the data, and revised the manuscript, DOO performed admixture analysis, MSM participated in the acquisition of experimental data, MDL participated in the acquisition of experimental data, study design, and writing of the manuscript. All authors read and approved the final manuscript.

Funding
This study was supported Russian Science Foundation, project # 21-74-20145.

Availability of data and materials
Raw sequencing reads for are available in the NCBI under Bioproject PRJNA986448 (PacBio and HiC reads) [51], PRJNA986297 (individual plants F2 lel x msu-wt) [52], PRJNA986442 (individual plants F2 msu-wt x ME line) [53]. Genome assembly is available in NCBI under accession number GCA_001974645.2 (ASM197464v2) [54]. Custom scripts are available on GitHub via following link https://github.com/ArtemKasianov/CapsellaArticle2 [55].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 18 July 2023 Accepted: 23 January 2024
Published online: 05 March 2024

References

1. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet. 2017;18:411–24.
2. Madlung A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity. 2013;110:99–104.
3. Salman-Minkov A, Sabath N, Mayrose I. Whole-genome duplication as a key factor in crop domestication. Nat Plants. 2016;2:16115.
4. Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, Lysak MA, et al. Gradual evolution of allopolyploidy in *Arabidopsis suecica*. Nat Ecol Evol. 2021;5:1367–81.
5. Monnahan P, Kolář F, Baduel P, Sailer C, Koch J, Horvath R, et al. Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. Nat Ecol Evol. 2019;3:457–68.
6. Han T-S, Wu Q, Hou X-H, Li Z-W, Zou Y-P, Ge S, et al. Frequent introgressions from diploid species contribute to the adaptation of the tetraploid Shepherd's Purse (*Capsella bursa-pastoris*). Molecular Plant. 2015;8:427–38.
7. Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. Proc Natl Acad Sci. 2015;112:2806–11.
8. Plants of the World Online: *Capsella grandiflora* (Fauché & Chaub.) Boiss. 2023. https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:279976-1. Accessed 26 June 2023.
9. Plants of the World Online: *Capsella orientalis* Klokov. 2023. https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:280028-1. Accessed 26 June 2023.
10. Frenot Y, Chown SL, Whinam J, Selkirk PM, Convey P, Skotnicki M, et al. Biological invasions in the Antarctic: extent, impacts and implications. Biol Rev Camb Philos Soc. 2005;80:45–72.
11. Plants of the World Online: *Capsella bursa-pastoris* (L.) Medik. 2023. https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:30092589-2. Accessed 26 June 2023.
12. Choi B, Jeong H, Kim E. Phenotypic plasticity of *Capsella bursa-pastoris* (Brassicaceae) and its effect on fitness in response to temperature and soil moisture. Plant Spec Biol. 2019;34:5–10.
13. Wesse C, Welk E, Hurka H, Neuffer B. Geographical pattern of genetic diversity in *Capsella bursa-pastoris* (Brassicaceae) — A global perspective. Ecol Evol. 2020;11:199–213.
14. Kasianov AS, Klepikova AV, Kulakovskiy IV, Gerasimov ES, Fedotova AV, Besedina EG, et al. High-quality genome assembly of *Capsella bursa-pastoris* reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. Plant J. 2017;91:278–91.
15. Slotte T, Huang H-R, Holm K, Ceplitis A, Onge KSt, Chen J, et al. Splicing variation at a *FLOWERING LOCUS C* homeolog is associated with flowering time variation in the tetraploid *Capsella bursa-pastoris*. Genetics. 2009;183:337–45.
16. Ziermann J, Ritz MS, Hameister S, Abel C, Hoffmann MH, Neuffer B, et al. Floral visitation and reproductive traits of *Stamenoid petals*, a naturally occurring floral homeotic variant of *Capsella bursa-pastoris* (Brassicaceae). Planta. 2009;230:1239–49.
17. Klepikova AV, Shnayder ED, Kasianov AS, Remizowa MV, Sokoloff DD, Penin AA. *lepidium-like*, a naturally occurring mutant of *Capsella*

bursa-pastoris, and its implications on the evolution of petal loss in Cruciferae. Front Plant Sci. 2021;12.
18. Hintz M, Bartholmes C, Nutt P, Ziermann J, Hameister S, Neuffer B, et al. Catching a "hopeful monster": shepherd's purse (*Capsella bursa-pastoris*) as a model system to study the evolution of flower development. J Exp Bot. 2006;57:3531–42.
19. Kryvokhyzha D, Milesi P, Duan T, Orsucci M, Wright SI, Glémin S, et al. Towards the new normal: Transcriptomic convergence and genomic legacy of the two subgenomes of an allopolyploid weed (*Capsella bursa-pastoris*). PLOS Genet. 2019;15: e1008131.
20. Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. The dynamic ups and downs of genome size evolution in Brassicaceae. Mol Biol Evol. 2009;26:85–98.
21. Vuruputoor VS, Monyak D, Fetter KC, Webster C, Bhattarai A, Shrestha B, et al. Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. Appl Plant Sci. 2023;11: e11533.
22. Acarkan A, Roßberg M, Koch M, Schmidt R. Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. Plant J. 2000;23:55–62.
23. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, et al. Evolution of genome size in Brassicaceae. Ann Botany. 2005;95:229–35.
24. Huang H-R, Liu J-J, Xu Y, Lascoux M, Ge X-J, Wright SI. Homeologue-specific expression divergence in the recently formed tetraploid *Capsella bursa-pastoris* (Brassicaceae). New Phytologist. 2018;220:624–35.
25. Kryvokhyzha D, Salcedo A, Eriksson MC, Duan T, Tawari N, Chen J, et al. Parental legacy, demography, and admixture influenced the evolution of the two subgenomes of the tetraploid *Capsella bursa-pastoris* (Brassicaceae). PLOS Genet. 2019;15: e1007949.
26. Kryvokhyzha D, Holm K, Chen J, Cornille A, Glémin S, Wright SI, et al. The influence of population structure on gene expression and flowering time variation in the ubiquitous weed *Capsella bursa-pastoris* (Brassicaceae). Mole Ecol. 2016;25:1106–21.
27. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
28. Dedukh D, Litvinchuk S, Rosanov J, Mazepa G, Saifitdinova A, Shabanov D, et al. Optional endoreplication and selective elimination of parental genomes during oogenesis in diploid and triploid hybrid European water frogs. PLOS ONE. 2015;10: e0123304.
29. Malinsky M, Matschiner M, Svardal H. Dsuite - fast D-statistics and related admixture evidence from VCF files. Mol Ecol Res. 2021;21:584–95.
30. St Onge KR, Foxe JP, Li J, Li H, Holm K, Corcoran P, et al. Coalescent-based analysis distinguishes between allo- and autopolyploid origin in Shepherd's Purse (*Capsella bursa-pastoris*). Mol Biol Evol. 2012;29:1721–33.
31. Roux C, Pannell JR. Inferring the mode of origin of polyploid species from next-generation sequence data. Mol Ecol. 2015;24:1047–59.
32. Hurka H, Friesen N, German DA, Franzke A, Neuffer B. 'Missing link' species *Capsella orientalis* and *Capsella thracica* elucidate evolution of model plant genus *Capsella* (Brassicaceae). Mole Ecol. 2012;21:1223–38.
33. Bird KA, VanBuren R, Puzey JR, Edger PP. The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytologist. 2018;220:87–93.
34. Mason AS, Wendel JF. Homoeologous Exchanges, Segmental allopolyploidy, and polyploid genome evolution. Front Genet. 2020;11.
35. Lashermes P, Combes M-C, Hueber Y, Severac D, Dereeper A. Genome rearrangements derived from homoeologous recombination following allopolyploidy speciation in coffee. Plant J. 2014;78:674–85.
36. Slotte T, Ceplitis A, Neuffer B, Hurka H, Lascoux M. Intrageneric phylogeny of *Capsella* (Brassicaceae) and the origin of the tetraploid *C. bursa-pastoris* based on chloroplast and nuclear DNA sequences. Am J Bot. 2006;93(11):1714-24.
37. Ågren JA, Wang W, Koenig D, Neuffer B, Weigel D, Wright SI. Mating system shifts and transposable element evolution in the plant genus Capsella. BMC Genom. 2014;15:602.
38. Doyle JJ, Doyle JL. A Rapid DNA Isolation Procedure for Small Quantities of Fresh Leaf Tissue. Phytochemical Bull. 1987;19:11–5.
39. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
40. Wu Y, Bhat PR, Close TJ, Lonardi S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLOS Genet. 2008;4: e1000212.
41. ArimaGenomics. Mapping pipeline for data generated using Arima-HiC. https://github.com/ArimaGenomics/mapping_pipeline. Accessed 26 June 2023
42. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 2019;5:833–45.
43. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. Juicebox.js provides a cloud-based visualization system for Hi-C data. Cell Syst. 2018;6:256-258.e1.
44. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer; 2019. p. 65–95.
45. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021;37:1639–43.
46. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:giab008.
47. Chang CC. Data management and summary statistics with PLINK. In: Dutheil JY, editor. Statistical Population Genomics. New York, NY: Springer US; 2020:49–65.
48. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
49. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573–89.
50. Vikram E. Chhatre. Distruct. http://distruct2.popgen.org Accessed 26 June 2023.
51. Penin AA, Kasianov AS, Klepikova AV, Omelchenko DO, Makarenko MS, Logacheva MD. High-quality chromosome scale genome assembly of *Capsella bursa-pastoris*. Sequence Read Archive. 2023. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA986448/.
52. Penin AA, Kasianov AS, Klepikova AV, Omelchenko DO, Makarenko MS, Logacheva MD. Genetic map of *Capsella bursa-pastoris* F2 lel x Msk (DNA-seq). Sequence Read Archive. 2023. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA986297
53. Penin AA, Kasianov AS, Klepikova AV, Omelchenko DO, Makarenko MS, Logacheva MD. Genetic map of *Capsella bursa-pastoris* F2 Msk x Cbp_ME (DNA-seq). Sequence Read Archive. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA986442.
54. Penin AA, Kasianov AS, Klepikova AV, Omelchenko DO, Makarenko MS, Logacheva MD. Genome assembly ASM197464v2. GenBank. 2024. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_001974645.2.
55. Kasianov AS. CapsellaArticle2. GitHub. 2023. https://github.com/ArtemKasianov/CapsellaArticle2

Publisher's Note