Supporting Information

Douglas et al. 10.1073/pnas.1412277112

SI Text

Plant Growth and Flow Cytometry. Seeds were sterilized, plated on half-strength Murashige–Skoog nutrient medium, vernalized at 4 °C for 3 wk, and germinated at room temperature. Seedlings were planted in a standard potting mix in 1-L round pots and grown at the University of Toronto glasshouse. Ploidy was confirmed using flow cytometry with an internal standard (Plant Cytometry Services).

Genome Assembly, Variant Calling, and Phasing. A fragment-level assembly of *C. bursa-pastoris* was undertaken with the aim of generating kilobase-sized range sequences that would be informative for a gene-level realignment against *C. rubella*, rather than for a comprehensive genome assembly per se. The low sequence divergence between homeologous sequences poses challenges for assembly and argued for a relatively conservative approach, as a consequence of which the Ray assembler (1) was selected over more aggressive assemblers (e.g., Velvet, European Bioinformatics Institute). Approximately 20 million 2 × 108-bp Illumina read pairs (4.3 gigabases) from two European accessions (12.4 and 16.9) were 3′-trimmed to remove sequences with a quality <30, and contigs were then generated by the Ray assembler version 1.4 (1) using an assembly sequence (Kmer) length of 31. A marginal degree of further scaffolding was then undertaken with the SOAPdenovo assembler (BGI; version 1.05) using the same paired-end data and a Kmer length of 51. However, this additional scaffolding only slightly extended the Ray contigs. The final scaffold N50 was 2.5 kb (0.23% uncalled bases), and the contig N50 was 2.4 kb, with a maximum scaffold/contig length of ∼25 kb. The total genome size was 210.5 megabases with 83% of the genome present in sequences longer than 0.5 kb.

For polymorphism analyses, Illumina reads for both *C. orientalis* and *C. bursa-pastoris* were mapped to the *C. rubella* reference genome (2), using Stampy version 13 (3), and Picard (broadinstitute.github.io/picard) was used for read sorting and file format conversion. Genotyping of SNPs and short indels was done using the GATK software package (4, 5), with realignment of sequences surrounding short indels (6). The resulting polymorphism data were combined with similarly processed whole-genome polymorphism data from 13 *C. grandiflora* individuals (7, 8).

To minimize the risk of spurious SNP calling, each site in the variant call format (VCF) files was rigorously filtered by a Phred quality score of 40 and depth. *C. bursa-pastoris* and *C. grandiflora* VCF file depth cutoffs were a minimum of 20 and a maximum of 100. Depth cutoffs for *C. orientalis* were a minimum of 15 and a maximum of 100. We also excluded 20-kb genomic windows where less than 30% of sites passed these quality and depth criteria for a given species. Many of these regions corresponded to pericentromeric regions of the genomes.

For the transcriptome analyses, we extracted RNA from leaf tissue with a Qiagen Plant RNEasy Plant Mini Kit from one *C. bursa-pastoris* individual (SE14), six *C. grandiflora* individuals, and three biological replicates each of two *C. orientalis* individuals. RNA-sequencing libraries were generated using the TruSeq RNA version 2 protocol (Illumina). RNA sequencing was conducted at the Genome Quebec Innovation Centre, the Centre for Applied Genomics, and the Uppsala SNP&SEQ Technology Platform, Uppsala University, on Illumina HiSeq 2000 instruments.

All RNA-sequencing data were mapped to an exon-only *Capsella* reference, and variants were called using the same procedure as the genomic DNA above. We focused on sites that were determined to have fixed differences between the two *C. bursa-pastoris* subgenomes and that also showed no evidence of allelic mapping bias from genomic sequence (<40% or >60% of genomic reads mapping to a particular homeolog). Homeolog-specific expression was measured in the *C. bursa-pastoris* sample by calculating the proportion of all reads mapping to the *C. grandiflora*-descended (*C. bp* A) subgenome using the "DepthPerAlleleBySample" values found in the VCF file. We only measured homeolog-specific expression in genes with at least two informative "heterozygous" sites (i.e., two sites with fixed differences between the subgenomes), leaving us with 5,587 genes. We defined homeolog-specific expression of a gene and corresponding homeolog-specific "silencing" as cases where at least 95% of reads mapped to one subgenome. We used the "HTSeq.scripts.count" feature of HTSeq (9) (mode = intersection_nonempty) to count the number of pairs of reads that mapped to each gene in each of our *C. grandiflora* and *C. orientalis* samples. We normalized library size and estimated fold change in expression between the two species using DESeq (10). The number of differentially expressed genes between *C. grandiflora* and *C. orientalis* was determined using the DESeq significance test based on the negative binomial distribution (10). Note that the exact number of differentially expressed genes is probably underestimated due to the relatively low number of replicates. We identified 241 of 8,164 genes as differentially expressed between the two species based on a false discovery rate (FDR)-corrected P value <0.01.

To identify transposable element insertions in *C. bursa-pastoris*, we used the Popoolation TE approach developed by Kofler et al. (11). Using default settings, we ran the pipeline on paired-end Illumina (108-bp) samples from eight of our *C. bursa-pastoris* individuals, using the *C. rubella* genome (2) as the reference genome. Because Popoolation TE was originally developed for pooled population samples designed to infer population-wide frequencies, we modified it to apply to individual samples using frequency cutoffs to define heterozygous and homozygous insertions (11). Principal component analysis of insertion presence/absence in *C. bursa-pastoris* was analyzed, along with insertions previously identified in our samples of *C. orientalis* and *C. grandiflora* (12), using SPSS version 22 (SPSS, Inc.) under default settings.

Phasing of the *C. bursa-pastoris* homeologs was conducted with HapCUT version 0.6 (13), which has been used successfully to phase tetraploid wheat (14). HapCUT was run on SNPs passing our stringent heterozygosity and depth filters, and the resulting phased blocks were thus composed of high-confidence polymorphisms. Each *C. bursa-pastoris* sample was phased individually, and phased blocks were then identified as of unknown origin, descended from *C. grandiflora* or *C. orientalis*, or discordant between the two based on SNPs shared with the progenitors. SNPs identified within these blocks were compared across all samples, and any block containing an SNP inconsistently assigned to different subgenomes was removed from downstream analyses. This procedure could lead us to underestimate polymorphism levels in *C. bursa-pastoris* globally, but it should not bias our inference of selection or affect inference of biased fractionation, because no particular category of SNPs is preferentially removed and both subgenomes are expected to be equally affected by this procedure. In total, 523,425 SNPs were used to infer the parental origin of the phased blocks, with 16% of these SNPs resulting in discordant assignments across samples. The procedure was validated by comparison with Sanger sequence data for six independently amplified and sequenced genomic fragments from the two subgenomes. Principal component analysis on phased

SNPs was run using the resulting dataset, with analysis restricted to common SNPs at a frequency of six or more across the entire dataset. Principal component analysis of the phased SNPs was run using SPSS version 22 under default settings.

Comparative Genomics Analyses. A multiple whole-genome alignment between the *C. rubella* reference genome (2) and Illumina fragment assemblies of *C. bursa-pastoris*, *C. grandiflora* (2), *C. orientalis* (12), and *N. paniculata* (2) was conducted essentially as described by Haudry et al. (7). Briefly, each fragment assembly was initially aligned against the soft-masked (RepeatMasker, www.repeatmasker.org) *C. rubella* reference sequence using the alignment program LASTZ (15) with parameters --gapped --nochain --gfextend --strand=both. Alignments were then chained using axtChain (Kent tools; University of California, Santa Cruz) with a minimum chain score of 10,000 and a slightly customized linear gap table. Chains were then selected for the subset of most likely orthologous chains having the maximum alignment score against the *C. rubella* reference, retaining only a single chain for each *C. rubella* sequence for all assemblies except *C. bursa-pastoris*, in which up to two chains could be selected providing there was sufficient evidence for two good orthologous alignments. Finally, the individual alignments against *C. rubella* were iteratively merged by phylogenetic distance using the program MULTIZ (16) with default parameters to create a multiple alignment.

We constructed ML phylogenies for alignments of each assembled fragment that included two distinct *C. bursa-pastoris* sequences, as well as the orthologous *C. grandiflora*, *C. rubella*, *C. grandiflora*, and *N. paniculata* sequences. We constructed phylogenies using RAxML's (17) rapid bootstrap algorithm to find the best-scoring ML tree. Each phylogeny had 100 bootstrap replicates and used *N. paniculata* as the outgroup. We excluded trees with less than 80% bootstrap support at any branch from further analysis. We then used a custom Perl script to count the number of resulting phylogenies corresponding to each topology.

We used two approaches to validate our phylogenetic inference. First, we assessed whether similar patterns were observed with Sanger data and larger sample sizes for *C. bursa-pastoris*. Second, we assessed whether patterns of fixed heterozygosity and fixed differences between the diploid putative ancestors were in agreement with expectations under our phylogenetic hypothesis.

For the first validation, we assessed phylogenetic patterns at nine independent nuclear genes, where both subgenomes have previously been amplified and sequenced in *C. bursa-pastoris* with homeolog-specific primers (18–20). We amplified and sequenced the same loci in *C. orientalis* using these previous primers and used MUSCLE (21) to align our sequences to publicly available data for the same loci from *C. grandiflora* [sequences phased using PHASE2.1 (22, 23); *C. rubella* and *C. bursa-pastoris* (19, 20, 24–28)]. As outgroups, we used *Arabidopsis thaliana* and/or *N. paniculata.* All positions with gaps or missing data were removed, and data for each locus were collapsed into unique haplotypes using FaBox 1.40 (http://users-birc.au.dk/biopv/php/fabox/). Subsequently neighbor-joining trees were reconstructed in MEGA5 (29), with distances estimated based on the composite ML method (30) and support evaluated using 1,000 bootstrap replicates. In all cases, one of the *C. bursa-pastoris* subgenomes showed very high similarity to *C. orientalis* and the other clustered with *C. grandiflora* or *C. rubella* sequences, validating our inference of an allopolyploid origin of *C. bursa-pastoris* based on massively parallel sequencing data (Fig. S1).

Population Genetic Analyses. To infer demographic parameters associated with allopolyploid speciation in *C. bursa-pastoris*, we analyzed site frequency spectra for 60,225 SNPs at intergenic nonconserved regions and fourfold synonymous sites. Specifically, we used fastsimcoal2.1 (31) to infer demographic parameters based on the multidimensional site frequency spectrum for *C. grandiflora*, *C. orientalis*, and the two *C. bursa-pastoris* homeologous genomes. Estimates were obtained using the composite ML approach under four models that differed in the type of population size change (stepwise or exponential) allowed and in the presence or absence of postpolyploidization asymmetrical migration (Fig. 2). All parameter estimates were global ML estimates from 50 independent fastsimcoal2.1 runs, with a minimum of 50,000 and a maximum of 250,000 coalescent simulations, as well as 10–40 cycles of the likelihood maximization algorithm. Multidimensional site frequency spectrum (SFS) entries with less than five SNPs were pooled to avoid negative effects on the estimation procedure, as suggested by Excoffier et al. (31). We assumed a mutation rate of 7×10^{-9} per base pair and generation (32) and a generation time of 1 y when converting estimates to units of years and individuals. Confidence intervals of parameter estimates were obtained by parametric bootstrapping, with 100 bootstrap replicates per model. Model fit was assessed using Akaike's information criterion (AIC) and Akaike's weight of evidence, as in the study by Excoffier et al. (31). Note that because of possible linkage disequilibrium (LD) among our SNPs, particularly in selfing populations, confidence intervals and the strength of AIC model support should be treated with some caution. However, the parameter estimates themselves under the composite likelihood approach are expected to be robust, and we obtained comparable timing estimates across models (Fig. 2) and when using different subsets of sites with likely differences in their LD structure (nonconserved noncoding sequence and fourfold degenerate sites), arguing that our main conclusions are not affected by LD between SNPs. Furthermore, to investigate further the possible influence of LD on the demographic inference, we reran the models excluding sites less than 10 kb apart, and the major conclusions, including the best-fitting model, were found to be unaffected.

Molecular Evolutionary Analyses.

Inference of gene loss. We took two distinct approaches to identify putative deletion events. First, we used HTSeq (9) to count the number of reads mapping to each annotated gene in the *Capsella* reference genome, using the "intersection_nonempty" option. After normalizing each sample by the total number of reads, we identified genes that showed significant reductions in coverage in the *C. bursa-pastoris* samples in paired tests of both *C. grandiflora* and *C. orientalis*. By requiring significant depth reductions relative to both parental species, we should minimize problems associated with biased read mapping and/or ancestral copy number variation. Significance was assessed using two-tailed *t* tests assuming unequal variance. Only tests with *P* values less than 0.01 against both species (corresponding to an FDR of ∼5%) were treated as significant. Furthermore, to restrict our analyses to putative single-copy deletion events, rather than variance in read mapping success and/or high copy number genes, we only considered cases where the fold change in coverage in *C. bursa-pastoris* ranged from 0.25 to 0.65. Many of the identified deletions appear to span a single gene, but a number are also found in multigene clusters. Using our whole-genome de novo assembly, one-quarter of these events had detectable breakpoints within genomic scaffolds, with the remainder spanning repetitive sequences where breakpoints could not be resolved.

As a second approach, we used Pindel (33) to identify large deletions spanning whole genes and small indels affecting coding regions. Pindel was run for each sample independently and compared with the gene annotation. Gene deletions were called as deletions covering 80% of a locus. Overlapping variants between individuals and species were identified with the BEDTools (34) "intersect" command. Gene deletions and inversions required 80% overlap to be called as orthologous, and shorter indels required complete overlap.

Effect prediction of polymorphisms. The software package SnpEff version 3.5 was used to predict the genomic effects of SNPs and structural variants (35), given the *C. rubella* genome annotation (2). Mutations more likely to cause loss-of-function effects were identified by using the "-lof" option and parsing for mutations flagged for "HIGH" effect. This set included polymorphisms knocking out splice sites and start or stop codons and causing the gain of stop codons, as well as frameshift deletions and short insertions. To eliminate major-effect mutations that may have occurred in our *C. rubella* reference genome, we used *N. paniculata* to polarize these changes and retained only derived SNPs in our focal species. Possible compensatory mutations for all of these putatively deleterious mutations were accounted for within 50 bp of each mutation. For instance, frameshift mutations within the same gene and within 50 bp of each other were excluded from the analysis. Similarly, the gain of a stop codon within 50 bp of a lost stop codon would be called a compensatory mutation. Polymorphisms fixed between *C. bursa-pastoris* subgenomes were included as a separate category to test for the fixation of deleterious mutations within subgenomes. Because all mutations were called relative to the *C. rubella* assembly, all mutations were polarized by using one *N. paniculata* individual as an outgroup. Any putatively deleterious mutation also found in *N. paniculata* was excluded from the analysis.

Functional category enrichment. GOs, the classification of genes into classes of molecular function, cellular components, and biological process, were inferred from *Arabidopsis thaliana* using the Virtual Plant online server version 1.3 (36). Because *Capsella* is closely related to *Arabidopsis*, the majority of orthologous genes in *A. thaliana* are likely to have the same function in the *Capsella* species. A total of 19,520 genes in *Arabidopsis* that have known orthologs in *C. rubella* were included in this analysis.

Genes containing putatively deleterious SNPs, according to the SnpEff algorithm, in *C. bursa-pastoris* were analyzed with Virtual Plant's BioMaps tools (36). "Fixed heterozygotes" were considered separately. Genes associated with GO categories that were found to be enriched among singletons retained following the ancient WGD event in *Arabidopsis* (37) were acquired from the Virtual Plant database. Only GO categories associated with less than 1,000 genes were included for downstream analyses. Categories including "organelle" (6,778 genes), "cell part" (14,508 genes), and "cellular metabolic process" (5,998 genes) are examples of excluded groups. These GO terms are parents of the smaller GO classes included, so their removal likely reduced noise without compromising statistical power.

Scanning for selection. Polymorphisms at fourfold synonymous sites were counted in 1-kb, 10-kb, and 50-kb windows in coding regions upstream and downstream of selected putatively deleterious mutations in *C. bursa-pastoris*, and windows were averaged over mutations of the same category. The numbers of these neutral mutations were normalized by divergence to *N. paniculata*. Normalizing by divergence controls for differences in mutation rate across the genome. Again, fixed heterozygotes were considered separately. Singleton genes whose orthologs were retained following the ancient genome duplication in *Arabidopsis* were scanned (37), as were related genes from the same GO categories. Finally, nuclear genes associated with the chloroplast were considered separately, because they were significantly enriched for loss-of-function mutations in past analyses. VCFtools version 0.1.12a (38) was also used to assess nucleotide diversity over all site types in the same-sized windows surrounding fixed, putatively deleterious mutations using the "–window-pi" option. Neutral expectations of diversity surrounding a given fixation were generated by analyzing diversity in 50-kb windows surrounding fourfold synonymous fixations. For each window, 95% confidence intervals were computed through bootstrapping by substitution ($n = 1{,}000$). The result, shown in Fig. S5, is robust to all window sizes, although it is noisier at smaller window sizes.

DFE inference. The DFE for deleterious mutations was inferred using the ML approach of Keightley and Eyre-Walker (39), which compares the allele frequency spectra (AFS) for deleterious and neutral sites to infer the strength of purifying selection. Only sites that passed filtering criteria within all three *Capsella* species were analyzed. Polymorphisms segregating in other species were removed from the analysis. Frequency data were taken from *C. bursa-pastoris*, *C. grandiflora*, and *C. orientalis* for several site classes: zerofold nonsynonymous, fourfold synonymous, and conserved noncoding sites that were identified in a comparison involving nine Brassicaceae species using the *Capsella* genome as the reference (8). Fixed heterozygotes were excluded in these analyses, because they are not segregating within either subgenome. To ensure the number of chromosomes being compared among species was the same, the *C. grandiflora* dataset was down-sampled to 10 individuals. Heterozygous sites in *C. grandiflora* were converted to homozygotes by randomly selecting one of the two bases so as to mitigate any bias caused by higher heterozygosity in this species. To determine 95% confidence intervals, 200 bootstrap replicates of each site class were used to recalculate the folded AFS and numbers of invariant sites. The significance level for pairwise comparisons was determined as described by Keightley and Eyre-Walker (39). To account for uncertainty due to LD when estimating the DFE, significance tests and bootstrap confidence intervals were generated by resampling 10-kb blocks of SNPs.

Forward simulations. Forward population genetic simulations were conducted to investigate how the demographic changes accompanying the evolution of polyploidization shaped the observed patterns of relaxed selection in the allopolyploid *C. bursa-pastoris*. Simulations were performed using SLiM software (40), which implements a Wright–Fisher model with selection and nonoverlapping generations. Selective and demographics parameters estimated for *C. grandiflora* were used to simulate an ancestral outcrossing population. The ancestral population was completely outcrossing ($t = 1.00$) and comprised 529 individuals (census size, N), following a rescaling of the N_e estimates under a stepwise model (Fig. 2*A*). One-kilobase genes were simulated to investigate patterns of selection occurring within coding regions. The mutation rate ($\mu = 3 \times 10^{-5}$ per site per generation) and recombination rate ($r = 3 \times 10^{-2}$ per site per generation) were chosen to match the observed diversity for *C. grandiflora* (2). Note that because forward simulations have a rescaled N_e size downward due to computation time, the mutation rate and recombination rate estimates are similarly rescaled upward to match compound parameters inferred for *C. grandiflora*. We modeled 33% of sites within each gene as neutral and 66% as deleterious. Selective coefficients for deleterious sites were drawn from a gamma distribution with a mean N of 431.22 and shape parameter of 0.279, matching the inferred DFE for *C. grandiflora* (Fig. 4). Neutral and deleterious sites were interspersed randomly across each gene. All mutations were additive (dominance coefficient: $h = 0.5$), and there were no beneficial ones. Simulations were run for $10N$ generations to reach stationary equilibrium, where N was measured in terms of the ancestral population size. At that time, we created a population split resulting in a second population with $t = 0.02$ and $n = 53$, effectively simulating the combined effects of a shift to predominant selfing and a 10-fold size reduction experienced by *C. bursa-pastoris*. Although our inferred best demographic model was one with an exponential population size change, our forward simulations cannot accommodate exponential growth, and we therefore focused on the stepwise demographic change model (Fig. 2*A*). However, the bottleneck under a stepwise model should adequately capture the demographic influence on selection efficacy because the current N_e of *C. grandiflora* under a stepwise model was similar to the ancestral N_e at the time of speciation under the exponential model. Furthermore, much of the change in the exponential model was driven by the expansion along

the *C. grandiflora* lineage, with little inferred exponential population size change inferred for *C. bursa-pastoris*. We tracked mutations in the ancestral and newly split simulated populations over $0.35N$ generations, approximately the time since the split of the *C. grandiflora*-descended *C. bursa-pastoris* subgenome (Fig. 2). There were 5,000 runs of simulations, effectively simulating 5,000 independent genes. For each run, 10 individuals from each population, equivalent to the sample size of the empirical data, were randomly sampled to generate nonsynonymous and synonymous AFS. These AFS were compared for each population to infer the DFE, and bootstrap confidence intervals for each $N_e s$ category of the DFE were generated by resampling across genes as applied for the empirical investigations.

Sample Information. *C. bursa-pastoris* sample designations (region; collector) are as follows:

70.5 (Artemida, Greece; Kate St. Onge)

5.16 (Valladolid, Spain; Santiago Gonzalez-Martinez)

13.16 (Krakow, Poland; Sandra Sherwood)

39-12-28 (Bacia, Italy; Kate St. Onge)

12.4 (Halle, Germany; Walter Durka)

16.9 (Nijmegen, The Netherlands; Koen Verhoeven)

VLA (Vladivostok, Russia; Martin Lascoux)

PL (Puli, Taiwan, China; Y.-W. Yang)

SE14 (Harnosand, Sweden; Svante Holm)

RK32 (Reykjavik, Iceland; John Paul Foxe)

Sequencing Accessions. Raw reads were uploaded to GenBank's Short Read Archive (SRA) under the following BioProject accession numbers: *C. bursa-pastoris* genomic data (PRJNA268827), *C. bursa-pastoris* RNA-sequencing (PRJNA268847), and *C. grandiflora* RNAsequencing (PRJNA268848). *C. orientalis* RNA-sequencing data were uploaded to the European Bioinformatics Institute under study accession no. PRJEB7879.

Publicly available sequencing from the SRA used in this study included genomic sequencing data for 13 *C. grandiflora* individuals under BioProject PRJNA254516 (7, 8) and 10 *C. orientalis* individuals under BioProject PRJNA245911.

Also, to validate our conclusion regarding the hybrid origins of *C. bursa-pastoris*, we used available coding sequences for nine loci from *C. grandiflora* (*Cg*), *C. rubella* (*Cr*), *N. paniculata* (*Np*), and both of the *C. bursa-pastoris* subgenomes (*Cbp* A and *Cbp* B). Popset numbers and GenBank accession numbers for these sequences are shown below, organized by *Arabidopsis* ortholog:

At1g77120: 160334389 (*Cbp* A); 160334601 (*Cbp* B); 160334571 (*Cr*); 341865754 (*Cg*)

At5g10140: 160335879 (*Cbp* A); 160336063 (*Cbp* B); 160335193 (*Cr*); 341866744 (*Cg*)

At4g02560: 160336247 (*Cbp* A); 160336429 (*Cbp* B); 160335645 (*Cr*); 341865968 (*Cg*); DQ343348.1 (*Np*)

At4g00650: 160335277 (*Cbp* A); 160335461 (*Cbp* B); 160335235 (*Cr*); FJ650267.1, FJ650266.1, FJ650265.1, FJ650264.1, FJ650263.1, FJ650262.1 (*Cg*)

At1g03560: 260780309 (*Cbp* A); 260765968 (*Cbp* B); 341606816 (*Cr*); 341605684 (*Cg*)

At1g15240: JQ418695.1–JQ418745.1 (*Cbp*); JQ418746.1–JQ418751.1 (*Cr*); JX065232.1–JX065247.1 (*Cg*)

At2g26730: JQ418752.1–JQ418801.1 (*Cbp*); FJ182814.1–FJ182826.1 (*Cr*); FJ182827.1–FJ182845.1 (*Cg*)

At4g08920: 160334825, 260765934 (*Cbp* A); 160335009, 261047491 (*Cbp* B); 160334783, 341607428 (*Cr*); 341606524 (*Cg*)

At5g51670: JQ418859.1–JQ418909.1 (*Cbp*); FJ183262.1–FJ183275.1, JQ418910.1–JQ418914.1 (*Cr*); FJ183276.1–FJ183293.1 (*Cg*)

1. Boisvert S, Laviolette F, Corbeil J (2010) Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol* 17(11): 1519–1533.
2. Slotte T, et al. (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet* 45(7):831–835.
3. Lunter G, Goodson M (2011) Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21(6):936–939.
4. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.
5. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.
6. Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 11(1110):11.10.1–11.10.33.
7. Haudry A, et al. (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* 45(8):891–898.
8. Williamson RJ, et al. (2014) Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genet* 10(9): e1004622.
9. Anders S, Pyl PT, Huber W (2015) HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166–169.
10. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
11. Kofler R, Betancourt AJ, Schlötterer C (2012) Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet* 8(1):e1002487.
12. Ågren JA, et al. (2014) Mating system shifts and transposable element evolution in the plant genus *Capsella*. *BMC Genomics* 15:602.
13. Bansal V, Bafna V (2008) HapCUT: An efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* 24(16):i153–i159.
14. Krasileva KV, et al.; IWGS Consortium (2013) Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol* 14(6):R66.
15. Harris RS (2007) Improved pairwise alignment of genomic DNA. PhD thesis (The Pennsylvania State University, University Park, PA).
16. Blanchette M, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14(4):708–715.
17. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
18. Slotte T, Ceplitis A, Neuffer B, Hurka H, Lascoux M (2006) Intrageneric phylogeny of *Capsella* (Brassicaceae) and the origin of the tetraploid *C. bursa-pastoris* based on chloroplast and nuclear DNA sequences. *Am J Bot* 93(11):1714–1724.
19. Slotte T, Huang H, Lascoux M, Ceplitis A (2008) Polyploid speciation did not confer instant reproductive isolation in *Capsella* (Brassicaceae). *Mol Biol Evol* 25(7):1472–1481.
20. Slotte T, et al. (2009) Splicing variation at a FLOWERING LOCUS C homeolog is associated with flowering time variation in the tetraploid *Capsella bursa-pastoris*. *Genetics* 183(1):337–345.
21. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
22. Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73(5):1162–1169.
23. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68(4):978–989.
24. Guo Y-L, et al. (2009) Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. *Proc Natl Acad Sci USA* 106(13):5246–5251.
25. Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Mol Biol Evol* 27(8):1813–1821.
26. Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D (2011) Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. *Genome Biol Evol* 3:868–880.
27. St Onge KR, et al. (2012) Coalescent-based analysis distinguishes between allo- and autopolyploid origin in Shepherd's Purse (*Capsella bursa-pastoris*). *Mol Biol Evol* 29(7): 1721–1733.
28. St Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE (2011) Contrasting demographic history and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species with different mating systems. *Mol Ecol* 20(16):3306–3320.
29. Tamura K, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731–2739.
30. Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci USA* 101(30):11030–11035.

31. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genet* 9(10):e1003905.
32. Ossowski S, et al. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327(5961):92–94.
33. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25(21):2865–2871.
34. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
35. Cingolani P, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80–92.
36. Katari MS, et al. (2010) VirtualPlant: A software platform to support systems biology research. *Plant Physiol* 152(2):500–515.
37. De Smet R, et al. (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci USA* 110(8): 2898–2903.
38. Danecek P, et al.; 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.
39. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177(4):2251–2261.
40. Messer PW (2013) SLiM: Simulating evolution with selection and linkage. *Genetics* 194(4):1037–1039.

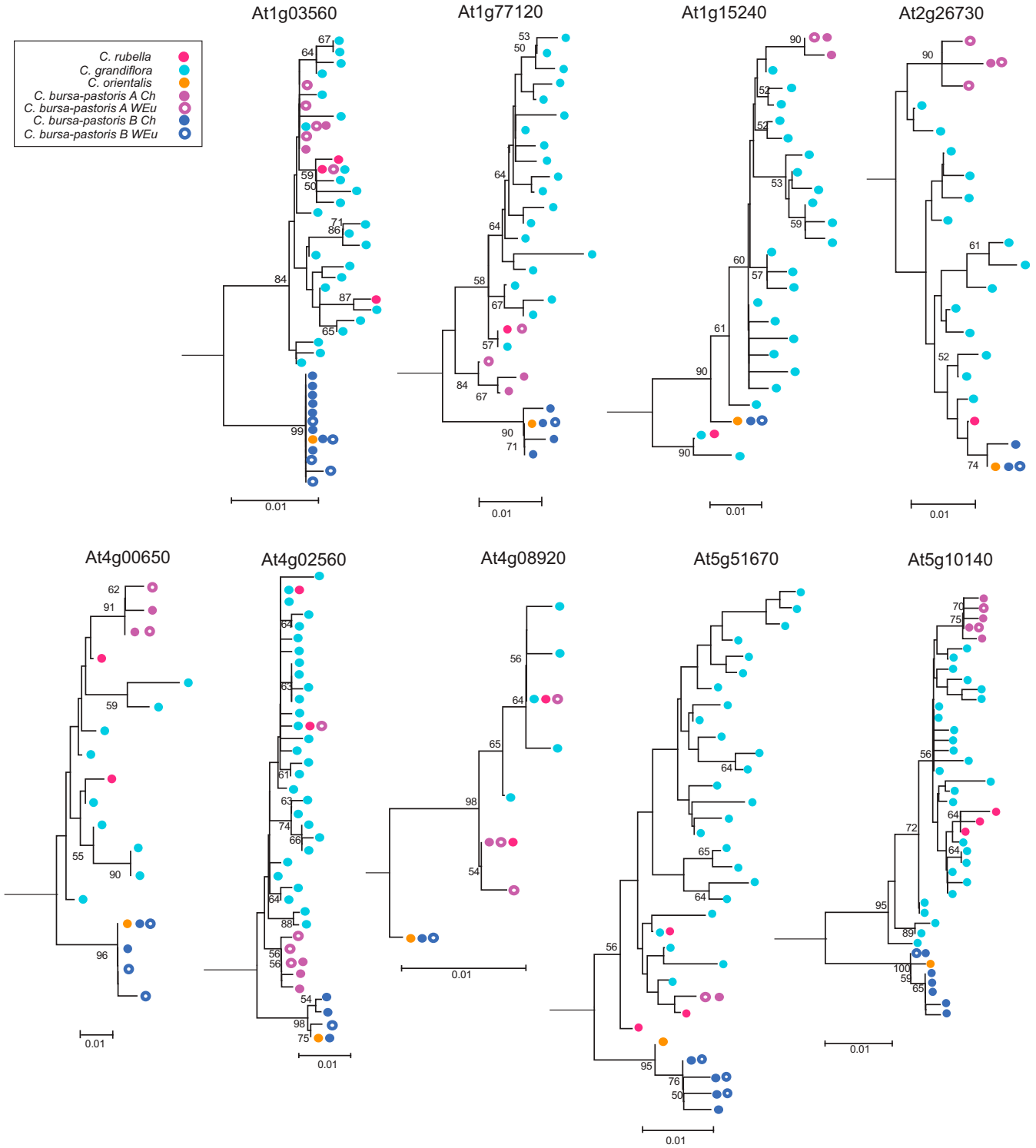

Fig. S1. Neighbor-joining trees for nine loci amplified using homeolog-specific PCR and sequenced at an average of 76 (range: 24–105) *C. bursa-pastoris* accessions across the native range of the species in Eurasia. Terminal nodes correspond to all unique haplotypes found for each locus. Labels indicate species, geographical origin [China (Ch) vs. Western Eurasia (WEu)], and subgenome designation (A or B), where applicable. As outgroups, we used *N. paniculata* or *A. thaliana* sequences. Only bootstrap values over 50% are shown.

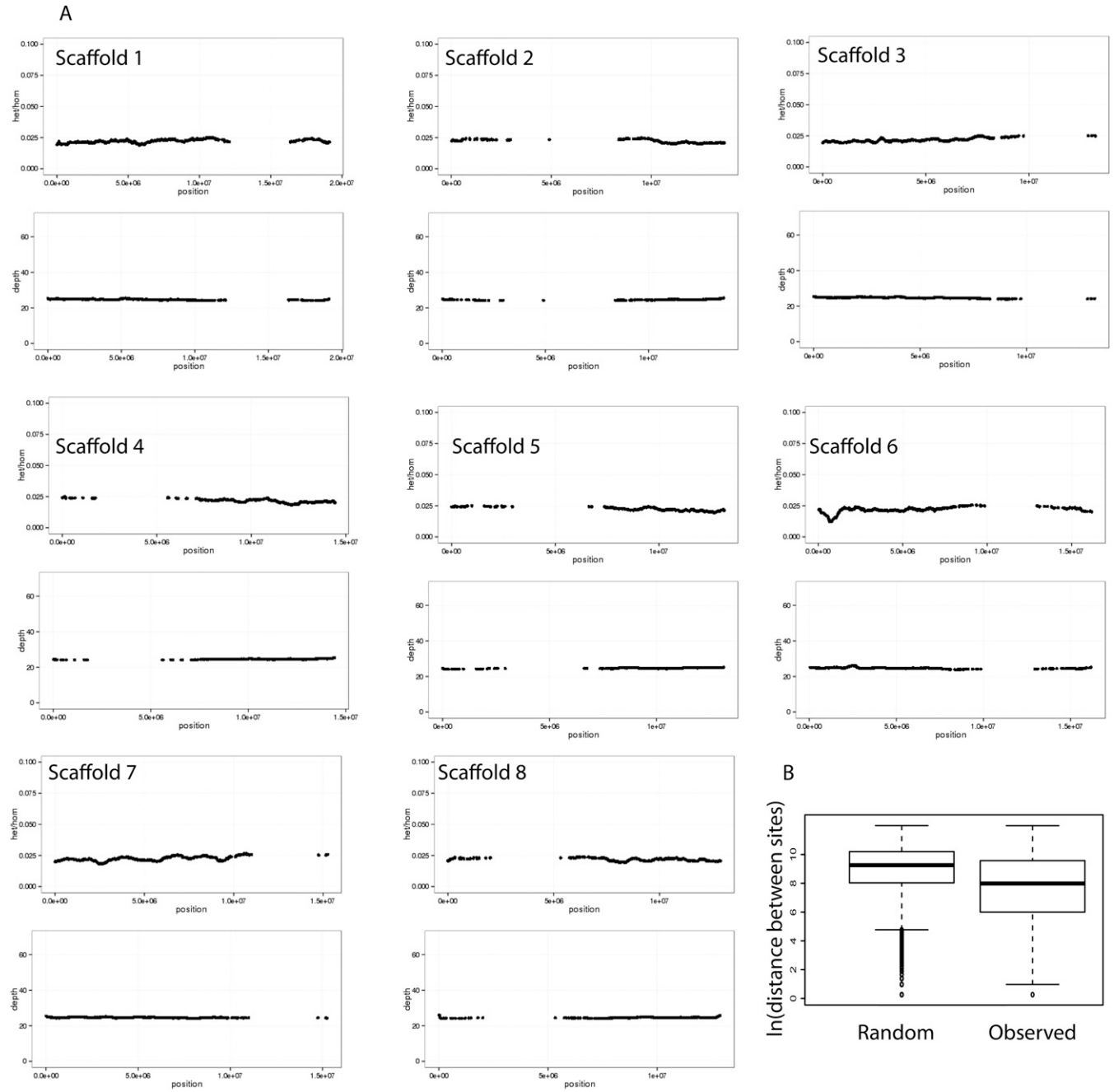

Fig. S2. (*A*) Sliding windows of heterozygosity and coverage in *C. bursa-pastoris*. Values shown are the number of heterozygous relative to homozygous genotype calls in 5,000 SNP windows and the average number of reads covering a genomic position. Five thousand SNP windows correspond to ~170,000 bp on average. (*B*) Distance to the nearest site missing a fixed difference between subgenomes that is present between the diploids. Distributions from random sampling and empirical data are shown.

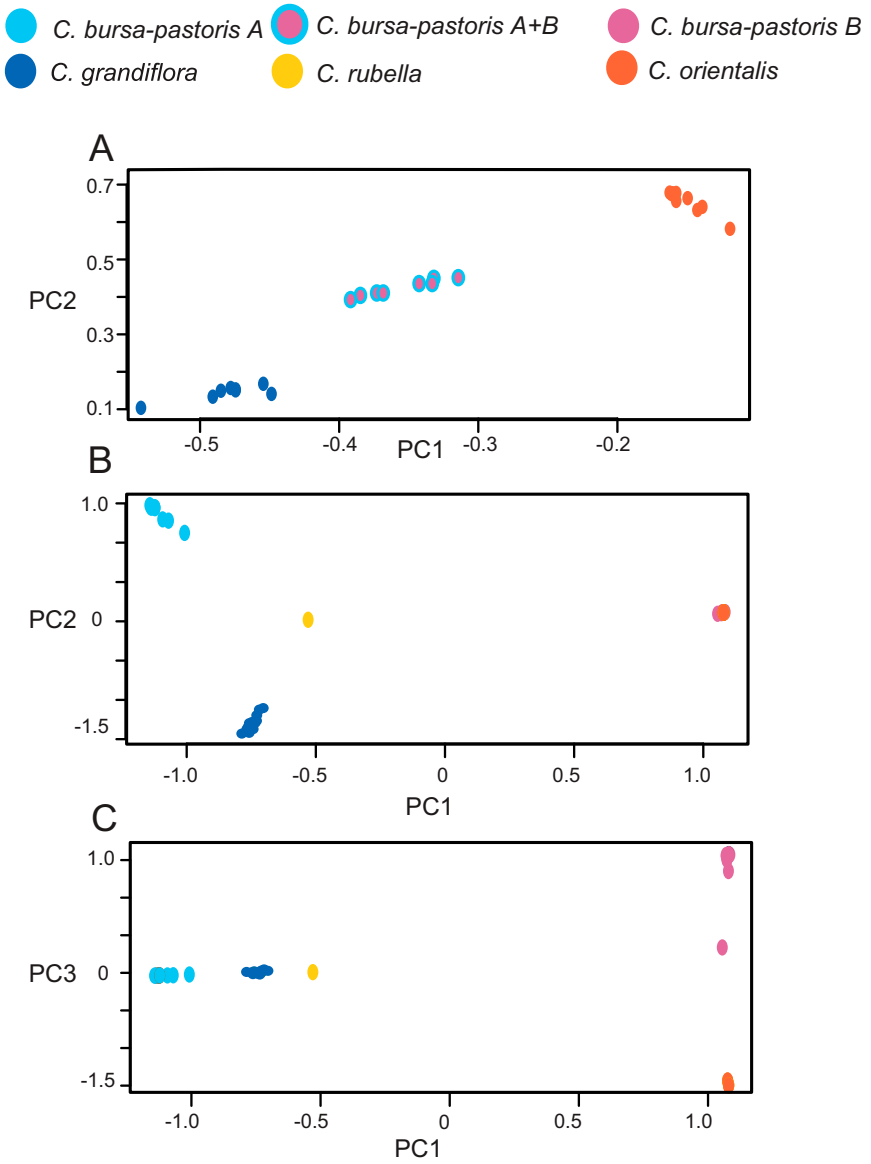

Fig. S3. Principal component (PC) analysis of *Capsella* species. (*A*) Analysis based on transposable element presence and absence across the three *Capsella* species, with *C. bursa-pastoris* insertions not phased by the subgenome of origin. The first two axes are shown. Analysis of SNP data from phased subgenomes of *C. bursa-pastoris*, with the first two axes (*B*) and the first and third axes (*C*) shown.

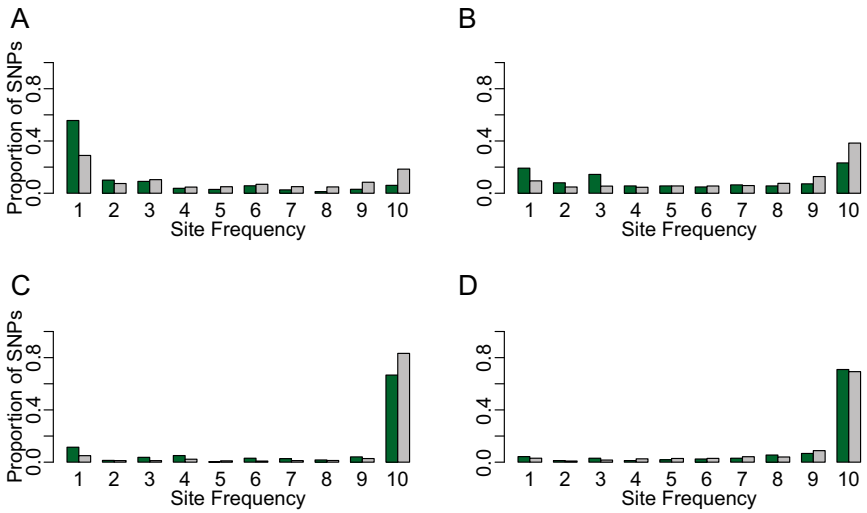

Fig. S4. *C. bursa-pastoris* site frequency spectra of deleterious SNPs (green) and fourfold synonymous SNPs (gray) segregating uniquely in *C. bursa-pastoris* (*A*), in both *C. bursa-pastoris* and *C. grandiflora* (*B*), in both *C. bursa-pastoris* and *C. orientalis* (*C*), and segregating in all three species (*D*). The deleterious SNP categories included are stop codon gained, stop codon lost, start codon lost, and splice site lost.

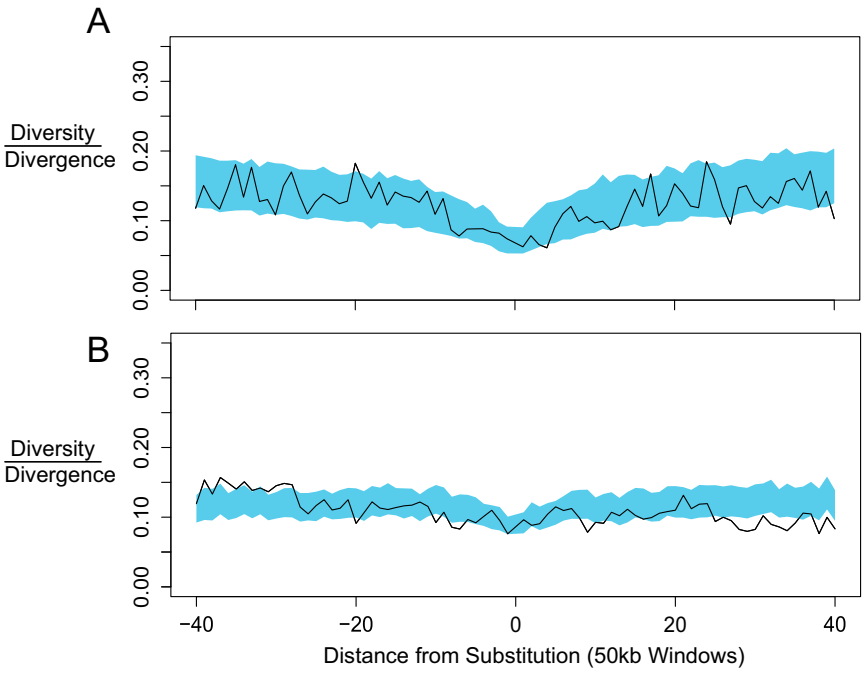

Fig. S5. Neutral fourfold synonymous diversity surrounding fixed putatively deleterious mutations in the two subgenomes of *C. bursa-pastoris*: *C. bursa-pastoris* subgenome A (*A*) and *C. bursa-pastoris* subgenome B (*B*). Diversity is normalized by divergence with *N. paniculata*. Shaded regions correspond to the bootstrapped (n = 1,000) 95% confidence interval of neutral diversity surrounding fixed fourfold synonymous mutations.

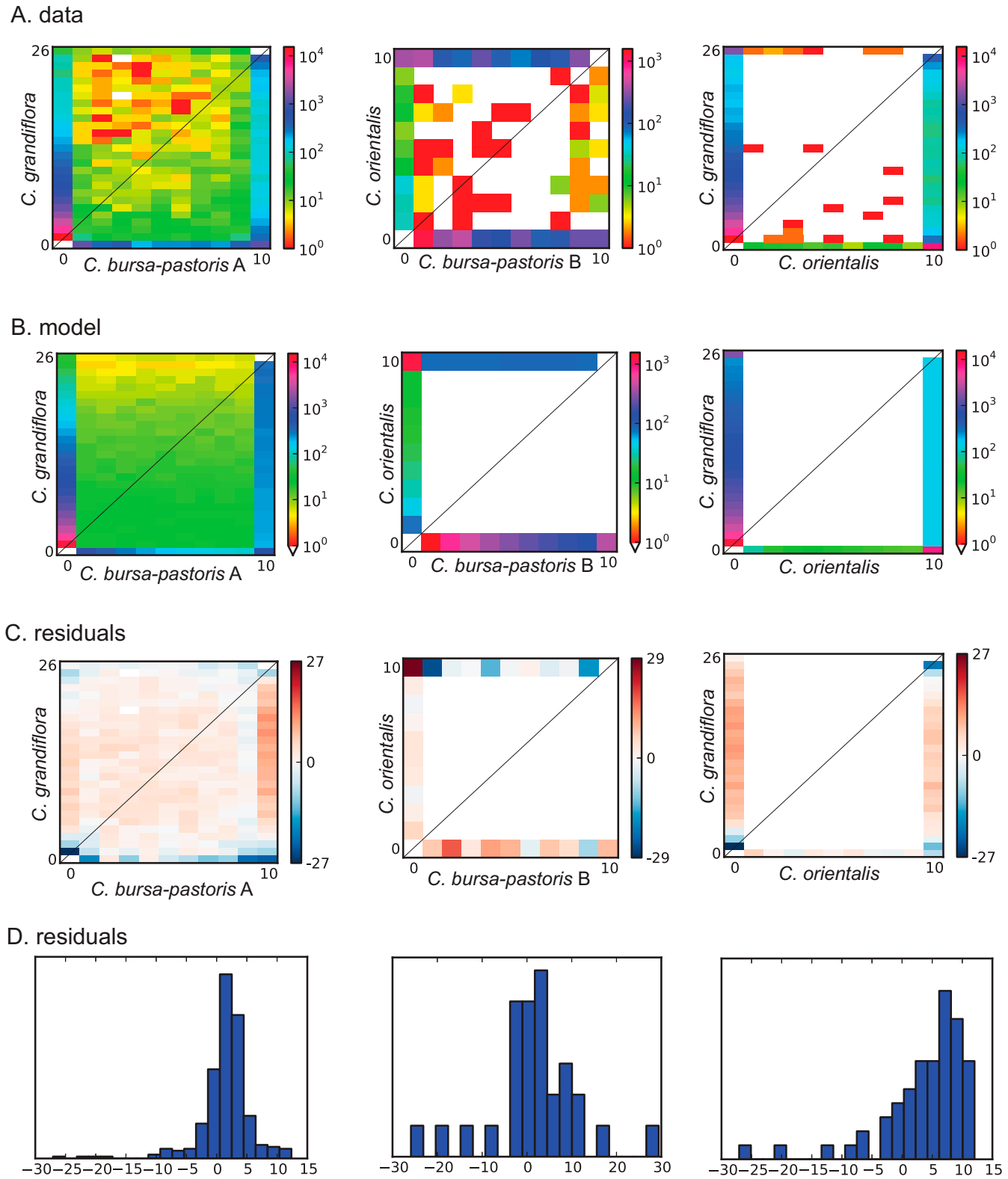

Fig. S6. Assessment of model fit for the best-fit exponential population size change model without migration. (*A*) Observed joint SFS for *C. grandiflora* vs. *C.* bursa-pastoris A, *C. orientalis* vs. *C. bursa-pastoris* B, and *C. grandiflora* vs. *C. orientalis*. (*B*) Expected joint SFS under the global ML parameters inferred by fastsimcoal2.1 for this model. (*C*) Model residuals across joint SFS. (*D*) Histograms of model residuals.

Table S1. Enrichment tests of putative inactivating mutations

Species	Mutation set	GO class	Identification no.	Term	Enrichment	Background	FDR
C. bursa-pastoris	Stop codon gained	BP	GO:0006259	DNA metabolic process	66/1,856	284/17,573	9E-05
C. bursa-pastoris	Stop codon gained	BP	GO:0006310	DNA recombination	18/1,856	52/17,573	0.01
C. bursa-pastoris	Stop codon gained	BP	GO:0006974	Response to DNA damage stimulus	36/1,856	163/17,573	0.02
C. bursa-pastoris	Stop codon gained	BP	GO:0022402	Cell cycle process	26/1,856	103/17,573	0.03
C. bursa-pastoris	Stop codon gained	BP	GO:0006508	Proteolysis	80/1,856	489/17,573	0.03
C. bursa-pastoris	Stop codon gained	MF	GO:0003824	Catalytic activity	817/1,885	6,278/17,843	2E-09
C. bursa-pastoris	Splice site lost	BP	GO:0006259	DNA metabolic process	44/1,402	284/17,573	0.03
C. bursa-pastoris	Splice site lost	MF	GO:0003824	Catalytic activity	585/1,429	6,278/17,843	0.001
C. grandiflora	Stop codon gained	BP	GO:0006259	DNA metabolic process	33/837	284/17,573	0.001
C. grandiflora	Stop codon gained	BP	GO:0006281	DNA repair	19/837	152/17,573	0.03
C. grandiflora	Stop codon gained	BP	GO:0006974	Response to DNA damage stimulus	20/837	163/17,573	0.03
C. grandiflora	Stop codon gained	MF	GO:0003824	Catalytic activity	391/860	6,278/17,843	5E-07
C. grandiflora	Splice site lost	MF	GO:0003824	Catalytic activity	323/749	6,278/17,843	0.003

BP, biological process MF, molecular function.

Table S2. Demographic parameter estimates with 95% confidence intervals for four models of allopolyploid speciation in *Capsella*

Model	N_e*Cg*	N_e*Co*	N_e*Cbp* A	N_e*Cbp* B	*T1*(*Cbp*)	*T2*(*Cg-Co*)	*2Nm*(*Cg-Cbp* A)	*rCg*	*rCo*	*rCbp*
Stepwise change	529 (184–692)	54 (17–67)	52 (18–71)	104 (42–131)	184 (63–251)	736 (237–981)	—	—	—	—
Stepwise change with migration	531 (230–552)	57 (24–60)	6 (2–6)	102 (45–109)	197 (80–211)	784 (335–847)	0.14 (0.04–0.52)	—	—	—
Exponential change	840 (148–868)	4 (1–6)	37 (6–55)	75 (12–101)	128 (22–177)	931 (161–1,159)	—	-4.1×10^{-6} (-3.5×10^{-5} to 1.3×10^{-6})	2.6×10^{-5} (1.6×10^{-5} to 1.4×10^{-4})	4.8×10^{-7} (-4.0×10^{-5} to 9.4×10^{-6})
Exponential change with migration	946 (383–1,077)	7 (4–14)	5 (3–6)	56 (53–130)	164 (100–269)	914 (448–1,188)	0.12 (0.06–0.34)	-4.4×10^{-6} (-6.2×10^{-6} to 1.3×10^{-6})	1.6×10^{-5} (4.3×10^{-6} to 2.3×10^{-5})	1.0×10^{-5} (-7.3×10^{-6} to 6.8×10^{-6})

Estimates of N_es for *Cg*, *Co*, and *Cbp* A and *Cbp* B are given in thousands of individuals, and estimates of the timing of the origin of *Cbp* [*T1*(*Cbp*)] and the split between *Cg* and *Co* [*T2*(*Cg-Co*)] are given in thousand years before present (kya). Note that for models with an exponential population size change, N_es correspond to current N_es.

Other Supporting Information Files

Dataset S1 (XLSX)