MBE Advance Access published January 27, 2012

Coalescent-based analysis distinguishes between allo- and autopolyploid origin in shepherd's purse *(Capsella bursa-pastoris)*.

Kate St.Onge*[1], John Paul Foxe*[4], Li Junrui*[5,7], Li Haipeng[5], Karl Holm[1], Pádraic Corcoran[2], Tanja Slotte[2], Martin Lascoux[1,5,†], Stephen Wright[3,6†]

*These authors contributed equally to this work

[1]Department of Evolutionary Functional Genomics, Uppsala University, 752 36 Uppsala, Sweden

[2]Department of Evolutionary Biology, Uppsala University 752 36 Uppsala, Sweden

[3]Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, On M5S 3B2, Canada

[4]Department of Biology, York University, 4700 Keele St. Toronto, Ontario M3J 1P3, Canada

[5]Laboratory of Evolutionary Genomics, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai, China

[6]Centre for Analysis of Genome Evolution and Function, University of Toronto

[7]Graduate School of the Chinese Academy of Sciences, Beijing 100039, China

[†]Corresponding authors: Stephen Wright and Martin Lascoux

E-mail: Stephen.Wright@utoronto.ca, Martin.Lascoux@ebc.uu.se

Phone: +1 (416) 946-8508, +46 (18) 4716416

Fax: +1 (416) 978-5878, +46 (18) 4716457

Keywords: autopoplyploidy, coalescent, Brassicaceae, Approximate bayesian computation, IM model

Running title: Autopolyploid speciation in *Capsella bursa-pastoris*

ABSTRACT

Polyploidization plays an important role in plant speciation. The most recent estimates report that up to 15% of angiosperm speciation events and 31% in ferns are accompanied by changes in ploidy level. Polyploids can arise either through autopolyploidy, when the sets of chromosomes originate from a single species, or through allopolyploidy, when they originate from different species. In this study we used two different coalescent-based methods to determine the date and mode of the polyploidization event that led to the tetraploid cosmopolitan weed, *Capsella bursa-pastoris*. We sampled 78 *C. bursa-pastoris* accessions, and 53 and 43 accessions from the only two other members of this genus, *C. grandiflora* and *C. rubella,* respectively, and sequenced these accessions at 14 unlinked nuclear loci with locus-specific primers in order to be able to distinguish the two homeologues in the tetraploid. A large fraction of fixed differences between homeologous genes in *C. bursa-pastoris* are segregating as polymorphisms in *C. grandiflora*, consistent with an autopolyploid origin followed by disomic inheritance. To test this, we first estimated the demographic parameters of an isolation-with-migration model in a pairwise fashion between *C. grandiflora* and both genomes of *C. bursa-pastoris* and used these parameters in coalescent simulations to test the mode of origin of *C. bursa-pastoris*. Secondly we used Approximate Bayesian Computation to compare an allopolyploid and an autopolyploid model. Both analyses led to the conclusion that *C. bursa-pastoris* originated less than one million years ago by doubling of the *C. grandiflora* genome.

INTRODUCTION

Polyploidy, or whole genome duplication, is widespread in plants. Polyploidy occurs in virtually all groups of vascular plants including ferns, mosses and algae (Otto and Whitton 2000). The most recent estimate of the prevalence of polyploids using phylogenetic data reports that 15% of speciation events in angiosperms and 31% in ferns are accompanied by changes in ploidy level (Wood et al. 2009), over 4 times higher than previous estimates of 2-4% in angiosperms and 7% in ferns (Otto and Whitton 2000). Analysis of fossil and genomic data estimate that 47%-100% of angiosperms have a polyploidy event at some point in their histories (Masterson 1994; Cui et al. 2006) and genomic studies have revealed that chromosomally diploid plant species, such as *Arabidopsis*, *Populus*, *Vitis*, and *Oryza* went through one or many rounds of polyploidisation during their evolution (e.g. Fawcett et al. 2009).

Typically, polyploids are divided into two categories based on their mode of origin, allopolyploids and autopolyploids. Allopolyploids have two full genome complements originating from two different species. These polyploids are expected to display disomic inheritance and form bivalents at meiosis, although disomic inheritance is not a strict indicator of allopolyploidy. All polyploid Brassica species studied so far are allopolyploids: *Brassica carinata*, *B. juncea*, and *B. napus* are tetraploids created from hybridization of the species *B. nigra*, *B. rapa* and *B. oleracea* in different combinations (U 1935). On the other hand, autopolyploids result from genome doubling within a species. Genome doubling can occur spontaneously or following the fusion of unreduced diploid gametes. Examples of autopolyploid plants include alfalfa and potato, and it was recently shown that the domesticated apple had an ancient autopolyploid origin (Velasco et al. 2010). Autopolyploids are typically expected to display polysomic inheritance and form multivalents at meiosis, although the generality of this rule has started to be questioned. Some autopolyploids are known to display disomic inheritance and this is probably more frequent than previously assumed (Soltis et al. 2010). However, how this occurs or how quickly disomic inheritance can evolve from polysomic inheritance is still poorly known (Cifuentes et al. 2010), although indirect evidence suggests that it can take place rapidly (Parisod et al. 2010). If disomic inheritance follows a period of polysomic inheritance, divergence times estimated from duplicated genes will reflect the time of onset of disomic inheritance, rather than the time of polyploidization (Gaut and Doebley 1997).

Determining the origin of polyploid species is an important aspect of speciation genetics and is central to our understanding of the mechanisms of formation of polyploids. While issues such as multiple origins of polyploid species, extinction of parental lineages and sampling of standing variation from progenitor species complicate this task (Doyle and Egan 2009; Soltis et al. 2010), recent advances in coalescent modeling have meanwhile facilitated it (Noor and Feder 2006; Becquet and Przeworski 2007; Hey and Nielsen 2007; Hey 2010). In particular, models of isolation-with-migration (IM) allow the differentiation of ancestral polymorphism from introgression and provide statistically sound estimates of divergence events (Wakeley and Hey 1997; Nielsen and Wakeley 2001). Using these models, diploid speciation processes have been studied in many organisms including *Drosophila* (Wang et al. 1997; Hey and Nielsen 2007), *Arabidopsis* (Ramos-Onsins et al. 2004), *Oryza* (Zhang and Ge 2007) and *Capsella* (Foxe et al. 2009). However, the use of coalescent-based models to study polyploidy and speciation has so far been limited with the notable exception of the studies of Jakobsson et al. (2006) in *A. suecica*, where an allopolyploid origin from *A. thaliana* and *A. arenosa* was known, and of *Capsella bursa-pastoris* in Slotte et al. (2008) and of *Arabidopsis lyrata* ssp. *kamchatica* of Taiwan in Wang et al. (2010).

The genus *Capsella* belongs to the mustard family (*Brassicaceae*) and is an attractive model genus because it is a young genus that contains few species with different mating systems and ploidy levels. The genus includes three species: *C. bursa-pastoris* (L.) Medik., a selfing tetraploid that displays a disomic inheritance and two diploid species, the outcrosser *C. grandiflora* (Fauché & Chaub.) Boiss., and the selfer *C. rubella* Reuter (Shull 1929; Hurka and Neuffer 1997). Previous studies suggested that *C. grandiflora* is ancestral to *C. bursa-pastoris* and *C. rubella* (Hurka and Neuffer 1997) and more recent findings confirmed that *C. rubella* diverged from *C. grandiflora* as recently or more recently than the Last Glacial Maximum (LGM, 18,000 years ago, St.Onge et al. 2011; 13,500 years ago, Foxe et al. 2009). *C. bursa-pastoris* has a worldwide distribution that can partly be explained anthropogenically. In contrast to *C. grandiflora* and *C. rubella*, *C. bursa-pastoris* can be found on each continent and thrives in a wide range of climates (Hurka and Neuffer 1997).

It is still unknown if *C. bursa-pastoris* is of autopolyploid or allopolyploid origin, and both possibilities have been suggested in previous work. Early isozyme electrophoresis indicated that *C. bursa-pastoris* shared alleles with both *C. grandiflora* and *C. rubella* and was hence thought to be an allopolyploid between these two species (Hurka et al. 1989). Later, evidence from restriction site variation in the chloroplast genome indicating that *C. rubella* was a more recently derived species led to the suggestion that *C. bursa-pastoris* was an ancient autopolyploid of *C. grandiflora* (Hurka and Neuffer 1997), despite the fact that *C. bursa-pastoris* displays disomic inheritance. Most recently, phylogenetic analysis suggested again that *C. bursa-pastoris* may be an allopolyploid, although not between *C. grandiflora* and *C. rubella* (Slotte et al. 2006).

A major limitation of these past studies is that they lack comprehensive data from all three *Capsella* species. In particular, the lack of large population data from *C. grandiflora*, the species of the genus known to harbor the most genetic variation, makes it difficult to conclusively determine the polyploid origin of *C. bursa-pastoris*. Here, we use DNA sequence data from 14 unlinked nuclear loci from large samples of all three *Capsella* species. In the absence of linkage data, assigning homeologues to particular genome copies in *C. bursa-pastoris* is not possible. To address this, we took the extreme possibility that more divergent copies from *C. grandiflora* all come from the same lineage. Since this would be most likely under an allopolyploid model, this allows us to explicitly test the plausibility of this model compared to autopolyploidy. To compare the fit of the data to allopolyploid vs. autopolyploid models of speciation we used a novel coalescent-based approach. First, we estimate the parameters of an Isolation-with-Migration model for pairs of species and then use these parameters in coalescent simulations to test the fit of the data to different models. Second, we use Approximate

Bayesian Computation (ABC, Beaumont 2010) to implement a two-split model and test our two competing hypotheses, the allopolyploid and the autopolyploid models. As Figure 1 shows, if *C. bursa-pastoris* has an autopolyploid origin we would expect the divergence time between the two homeologues to be as recent as, or (if there was an initial period of polysomic inheritance) more recent than the time at which *C. bursa-pastoris* derived from *C. grandiflora*, suggesting a simple way to test whether *C. bursa-pastoris* is of auto- or allopolyploid origin.

MATERIALS AND METHODS

Sample collection

Genetic data was collected from 78 accessions of *C. bursa-pastoris* from China, Taiwan, Israel and Europe, 43 accessions of *C. rubella* from Africa, South America, Europe and Israel and 53 accessions of *C. grandiflora* from Greece, covering a large portion of the narrow distribution of this species. Because this study focuses on the origin of *C. bursa-pastoris*, we have excluded samples from the Americas as Capsella species are a recent introduction there (Hurka and Neuffer, 1997). All our accessions come from natural populations from which we have collected seeds. In this study, we used a single accession per sampled population in most cases (see Table S1). Genetic data was also collected from one accession of *Neslia paniculata*, which was used as an outgroup in some analyses. *Neslia* is more recently diverged from *Capsella* than *Arabidopsis* (Bailey et al. 2006), providing a closer outgroup for inferences about *Capsella* divergence. Plants were grown in standard long-day conditions and DNA was extracted from fresh tissue of each individual using the QIAgen DNeasy Plant Mini Kit (QIAGEN, Valencia, California, USA). Accessions and their geographic origins are given in Table S1.

PCR and Sequencing

Fourteen gene fragments were selected for sequencing in this panel of individuals. These genes were found to be single copy in both diploids and duplicated in *C. bursa-pastoris,* as expected in a tetraploid. For four of the loci (At1g77120 (ADH), At5g10140 (FLC), At4g00650 (FRI) and At4g02560 (LD)), PCR primers for the diploid species and homeologue-specific primers for *C. bursa-pastoris* were designed as described by Slotte et al. (2006) and Slotte et al. (2008). For eight genes (At1g01040; At1g03560, At1g15240, At1G65450, At2g26730, At4g14190, At5g51670, At5g53020) primers for the diploid species were designed as described in Ross-Ibarra et al. (2008) and Foxe et al. (2009). For two additional loci (At2g18790 (PHYB) and At5g42800 (DFR)), primers were designed following a similar strategy. For all loci, initial primers were designed using Primer3 version 0.4.0 (Rozen and Skaletsky 2000) or PrimerQuest (Integrated DNA Technologies, Inc.) to amplify between 400-1000 bps using the *A. thaliana* genome sequence. The *A. thaliana* sequences were aligned to other Brassicaceae sequences when available to identify conserved regions. Both forward and reverse strands of the amplicons were sequenced directly at Lark Technologies (Houston, Texas), the Genome Quebec Innovation Centre (McGill University, Canada) or the Macrogen sequencing facility in Korea (Macrogen, Korea). Sequences were aligned and checked manually for heterozygous sites using either Sequencher version 4.7 (Gene Codes, Ann Arbor, MI) and Genedoc (Nicholas *et al.* 1997) or Codoncode Aligner version 2.0.6 (CodonCode, Dedham, MA). To differentiate the two homeologues of *C. bursa-pastoris*, the resulting sequences were used to design new homeologue-specific primers as in Slotte et al. (2006). In particular, we designed primers specific to SNPs showing fixed 'heterozygosity' amongst all of our samples, representing fixed SNP differences between homeologues. Each homeologue-specific amplicon was then sequenced directly and aligned as above. Based on direct sequencing of these samples only a single haplotype per homeologue was found for all of our primer pairs, implying homozygosity of our inbred samples. Details of the new primers for this study are shown in supplementary file S1. Sites with indels were removed before proceeding with analysis. The program PHASE 2.1 (Stephens et al. 2001), implemented in DnaSP 5.0 (Librado and Rozas 2009) was used to infer haplotypes in *C. grandiflora*. Additionally, each gene fragment was aligned with the homologous *A. thaliana* gene to infer the ancestral state of polymorphic sites. Loci and accessions where only one homeologue amplified were removed. New nucleotide sequences generated in this study that are greater than 200bp in length have been deposited in GenBank (accession numbers JQ418636-JQ419488). Complete sequence alignments, and sequence data from regions less than 200bp in length, are available upon request to the corresponding authors.

Summary statistics and estimation of species trees

A central challenge for our study is the difficulty in assigning homeologous genes to separate genomes of origin, designated as the *C. bursa-pastoris* A and B genomes. Homeologues were assigned to A and B genomes based upon the minimum number of synonymous substitutions between *C. grandiflora* and each homeologue as estimated using DnaSP version 5.0 (Librado and Rozas 2009). The most distant homeologue was assigned to the B genome while the other was assigned to A (Table S2; similar to Slotte *et al.* 2006 and Slotte *et al.* 2008; however, in these papers classification was based on all sites and *C. rubella* was used instead of *C. grandiflora*). These putative genomes were analysed separately for all subsequent analyses. Importantly, this classification effectively biases our analysis toward rejecting the hypothesis of the autopolyploid origin of *C. bursa-pastoris*. In particular, if the allopolyploid model is correct, the A and B homeologues likely represent distinct genomes with different parental origins, while under the autopolyploid model their difference is simply due to stochastic noise in the coalescent process, and the sorting does not reflect genome structure.

Classic genetic diversity summary statistics π (Tajima 1983) and Tajima's D (Tajima 1989) were calculated for synonymous sites in each species using a modified version of the Polymorphorama perl code (http://ib.berkeley.edu/labs/bachtrog/data/polyMORPHOrama/polyMORPHOrama.html) written by D. Bachtrog (UC Berkeley) and P. Andolfatto (Princeton University). The joint frequency spectra of derived polymorphic variants and the number of shared derived polymorphisms, unique polymorphisms, and fixed differences between each of the four genomes (Wakeley and Hey 1997) were calculated separately in a pairwise fashion using a Perl script written by S. Wright and a C program written by J. Li.

The molecular phylogenetic program BEST v. 1 (Bayesian estimation of species trees) (Liu 2008), which implements a Bayesian hierarchical model while accounting for the presence of deep coalescent events, was used to estimate the *Capsella* genus species tree using our multi-locus dataset (Liu 2008). Models within the BEST program assume (i) No population substructure within each population, (ii) No gene flow after species divergence and, (iii) No recombination within loci. Some of these assumptions, in particular the last one, will likely be violated. For example, recombination will be present in *C. grandiflora* and will make the length of terminal branches and the total branch length larger, and the time to the most recent common ancestor smaller (Schierup and Hein 2000). The program reportedly works best using concatenated alignments with little missing data. Consequently, we ran BEST using the 7 loci in this dataset that had the most consistent sampling of individuals across loci (At1g03560, At1g15240, At1g65450, At2g26730, At4g14190, At5g51670 and At5g53020). Alignments were concatenated using MacClade version 4.08 (available from http://macclade.org/). BEST was run in two ways, once using *A. thaliana* as an outgroup and again including both *A. thaliana* and *N. paniculata* (where available). In each case BEST was run twice, with 4 chains for a maximum of 2 million generations, with a burnin of 200,000 generations, sampling every 100 generations.

MIMAR and coalescent simulations

A first test of the null hypothesis that *C. bursa-pastoris* is an autopolyploid of *C. grandiflora* was done by first estimating the parameters of an isolation-with-migration model using the program MIMAR (Becquet and Przeworski 2007), and then performing coalescent simulations based on these parameters to test the null hypothesis (Hudson 2002). Because previous studies showed that *C. rubella* diverged very recently from *C. grandiflora* (Foxe et al. 2009; St.Onge et al. 2011), *C. rubella* was initially not included in this analysis. Furthermore, sites with >2 segregating bases were also excluded. MIMAR simulations were run in a pairwise fashion using *C. bursa-pastoris* A, *C. bursa-pastoris* B and *C. grandiflora* and allowing for three different models of migration between genomes: 1) absence of migration 2) symmetrical migration and 3) asymmetrical migration. Additionally, all analyses were run both with the ancestral effective population size unconstrained or assumed to be identical to the effective size of *C. grandiflora*. Prior limits for all parameters can be found in Table S3; these priors were set based on short initial runs with very wide priors. The program was run as described in Foxe et al. (2009), with the exception that each simulation was run for a total of 10,080 min. (1 week). We note that the model implemented by MIMAR does not allow a temporary reduction in Ne at the polyploid origin and so effectively allows multiple polyploid origins. Thus, inferences of effective population sizes should be considered a weighted average since divergence, rather than a direct estimate of the number of founders.

Because MIMAR simulations only model two taxa at a time, it does not on its own provide an explicit test of the mode of polyploid speciation. We therefore conducted coalescent simulations using MIMAR parameter estimates under models of both autopolyploidy and allopolyploidy. These models are depicted in Figure 1. Importantly, the differences between these models are the split times between *C. grandiflora* and the two genomes of *C. bursa-pastoris*. Under the autopolyploid model all three divergence times are the same, or the divergence time of the A and B homeologues is shorter than the time of either to *C. grandiflora*, if there was a period of polysomic inheritance. Under allopolyploidy the divergence time between *C. grandiflora* and *C. bursa-pastoris* B and between *C. bursa-pastoris* A and B are much longer than that between *C. grandiflora* and *C. bursa-pastoris* A. To model autopolyploid speciation, we used the inferred divergence time from MIMAR runs considering the two homeologues of *C. bursa-pastoris*, since this should provide an estimate of the lower bound for the time of autopolyploid origin. Under the allopolyploid model, the A and B copies in *C. bursa-pastoris* truly represent distinct genomes with different parental origins, and we used the two inferred divergence times from the MIMAR runs of *C. grandiflora* with the two distinct homeologue sets.

To compare our simulated data to our empirical data we used summary statistics introduced by Wakeley and Hey (1997) for each locus: the number of polymorphisms specific to the samples from populations 1 and 2 (called s1 and s2, respectively), where the population pairs correspond to *C. grandiflora/C. bursa-pastoris* A, *C. grandiflora/ C. bursa-pastoris* B and *C. bursa-pastoris* A /*C. bursa-pastoris* B, the number of shared polymorphisms between two samples (sp), and the number of sites fixed in either sample (f1 and f2, depending on which of the two species carries the ancestral state). We conducted simulations under both auto- and allopolyploidy models using the program *ms* (Hudson 2002) and the demographic parameters inferred with MIMAR. Namely, in the autopolyploid model we used T_1 as the divergence time between the two genomes and *C. grandiflora* (Figure 1) and in the allopolyploid model we used T_2' for the divergence between *C. bursa-pastoris* A and *C. grandiflora* and T_2" for the divergence between *C. bursa-pastoris* B and *C. grandiflora* (Figure 1). For each of the 14 genes we assumed that 10 chromosomes were sampled in each species and ran 10,000 simulations. We then calculated shared and fixed sites for each run and the mean over runs for each locus (additional information is available in Supplementary file S2 where the same analysis was carried out but considering both *C. grandiflora* and *C. rubella*).

Using these simulations we determined which of the summary statistics described above were informative in differentiating the two models. We found that unique polymorphisms (s1 and s2) did not differ between the two models. This may seem intuitive given that unique polymorphisms mostly reflect genealogies within that species, and therefore give limited information about speciation and divergence between the species in a genus. We therefore did not use these sites further. In contrast fixed sites (f1 and f2, depending on which of the two species carries the ancestral state) and shared polymorphisms (sp) did differ between ploidy models. Again this is intuitive as fixed differences correspond to mutations that happened in the early stages of speciation and are closely associated to divergence time, whereas shared polymorphisms, assuming they represent shared ancestral polymorphism and not recent introgression, represent polymorphism that were segregating in the ancestor and therefore give information about ancestral effective population sizes and divergence times. To make use of these two informative statistics we calculated their difference in the following way. If f1(C.bp B, C.g) and f1(C.bp A, C.g) are the number of fixed sites between *C. grandiflora* and *C. bursa-pastoris* B and *C. bursa-pastoris* A, respectively then their difference, fix_diff = f1(C.bp A, C.g) - f1(C.bp B, C.g). For convenience we used f1 to define fix_diff but the conclusions were the same when we used f2 (data not shown). If sp(C.bp A, C.g) and sp(C.bp B, C.g) represent the number of shared sites between *C. grandiflora* and *C. bursa-pastoris* A and *C. bursa-*

pastoris B, respectively, then the difference between them, shared_diff = sp(C.bp A, C.g) - sp(C.bp B, C.g).

We calculated the two differences defined above in our observed data and used them to test the null hypothesis that *C. bursa-pastoris* is an autopolyploid. Essentially we used a goodness-of-fit test to test the fit of our null hypothesis to the empirical data. We compared the observed values of the mean of fix_diff and shared_diff over the fourteen loci with the distribution of the same mean for the 10,000 simulation runs obtained under the autopolyploid model. When calculating the *p*-values for the autopolyploid model we used a two-tailed test for both test statistics. The *p*-value is therefore the fraction of simulations in which the absolute value of the mean is higher than the observed mean, with p-values of <0.05 indicating that our empirical values lie in the tails of the simulated distributions. We also assessed the fit of our alternative hypothesis, allopolyploidy, in a similar way except that in this case the tests were one-tailed because of the bias we created in our dataset. Therefore the *p*-value of the fixed_diff statistics is the fraction of simulations in which the mean is higher than the observed mean while in the case of shared_diff it is the fraction of simulations in which the mean is lower than the observed mean.

Approximate Bayesian Computation

An Approximate Bayesian Computation (ABC) analysis was used to evaluate two-split models resulting in *C. grandiflora* and the A and B genomes of *C. bursa-pastoris*. This analysis was performed using the program Seqlib-1.6 (De Mita et al., 2007) (http://sourceforge.net/projects/seqlib/), on the silent sites of the dataset. Because *C. rubella* has recently evolved from *C. grandiflora* and the variation in the species is more or less a subset of the variation found in *C. grandiflora* (Foxe et al. 2009; Guo et al. 2009, St Onge et al. 2011), we chose not to include it in the present analysis (see results and discussion). We evaluated two possible arrangements of coalescent events involving the three lineages 1) the A and B genomes of *C. bursa-pastoris* coalesce first, followed by this lineage coalescing with *C. grandiflora* 2) *C. grandiflora* coalesces first with the A genome of *C. bursa-pastoris*, followed by coalescence with the B genome. Model 1 represents an autopolyploidy event, where the divergence times of both *C. bursa-pastoris* genomes from to *C. grandiflora* are the same (Figure 1). Under model 2, the *C. bursa-pastoris* B genome is more diverged from *C. grandiflora* than *C. bursa-pastoris* A, representing an allopolyploidy event (Figure 1). These models have 9 parameters: the population mutation rates of each lineage, θ1, θ2 and θ3 where θ2 and θ3 are relative to θ1, the population recombination rate, ρ, a combined migration rate between all lineages, the dates of each divergence event, where the second event is additive to the first and the population sizes after each coalescent event, relative to θ1. It should be noted again that, for each gene, the *C. bursa-pastoris* allele most divergent from *C. grandiflora* was assigned to the B genome, effectively biasing our analysis towards model 2.

The ABC analysis of our two two-split models was performed using a set of 13 summary statistics; the number of shared, fixed and unique polymorphisms in all possible configurations with the three populations. We first performed initial runs with 1,000,000 samples using wide priors (Table S3). Using the local linear regression method described by Beaumont *et al.* (2002), 0.1% of the samples best fitting our empirical data were selected and used to create a prior for the ABC run. This allows us to explore the region of high probability identified in the initial run. 500,000 samples are taken in the ABC run, and 0.2% of the best fitting samples were used to estimate model parameters. A goodness-of-fit (GoF) test was used to validate the results of the ABC analysis. This test consisted of two sets of simulations, one using the point estimates for each parameter estimated in the ABC and one using the posterior distributions of each parameter. Further details on the goodness-of-fit test are in the Supplementary File S3..

A second analysis was performed using the same method but with only *C. bursa-pastoris* accessions from China. This was done to assess the influence of putatively introgressed alleles from *C. rubella* that only occurred in Europe (Slotte et al. 2008).

Testing for interlocus gene conversion in C. bursa-pastoris

We followed the approach of Slotte and colleagues (2008) to test for gene conversion between homeologues. In particular, we calculated the minimum number of recombination events, Rm, between homeologues (Hudson and Kaplan 1985) using DNAsp 5.0 (Librado and Rozas 2009), and tested for gene conversion using the geneconv software (Sawyer, 1989).

RESULTS

Patterns of polymorphism and phylogeny

Synonymous site diversity, measured as π, was higher in *C. grandiflora* than in *C. rubella* and *C. bursa-pastoris* A and B; median values were 0.028 for *C. grandiflora,* while they were zero for the latter two species (Figure S1). This is in agreement with expectations based on the respective mating systems of these species and previous studies (Slotte et al. 2008; Foxe et al. 2009; St.Onge et al. 2011). In particular the low level of nucleotide diversity observed in *C. rubella* is consistent with the presence of a severe population bottleneck associated to the shift to selfing (Foxe et al. 2009; Guo et al. 2009, St.Onge et al. 2011). The reduction in diversity seen in both *C. bursa-pastoris* A and B may also be the result of a recent bottleneck at speciation and transition to selfing. However, all species showed a high variance in diversity, with *C. rubella* showing the most extreme variance, with synonymous π values varying from 0 to the extremely high value of 0.15 for the DFR (At5g42800) locus. Resequencing of the full *C. rubella* genome and mRNA resequencing indicate that the high variance in diversity is a genomewide characteristic of the species (Wright et al, unpublished; D. Weigel, pers comm). This locus also showed high, but less elevated, polymorphism in *C. grandiflora* (0.08). Excluding DFR, the average synonymous diversity was 0.027 in *C. grandiflora*, 0.004 in *C. rubella*, 0.003 in *C. bursa-pastoris A*, and 0.003 in *C. bursa-pastoris B*. The average Tajima's D values at synonymous sites were negative for *C. bursa-pastoris* A (-0.19) and B genomes (-0.9), possibly reflective of recent population expansion. In *C. grandiflora*, synonymous Tajima's D was close to zero (-0.08), consistent with previous conclusions suggesting that this species is close to demographic equilibrium (Foxe et al. 2009; St.Onge et al. 2011). In *C. rubella*, synonymous Tajma's D was slightly negative (-0.2.).

The minimum number of synonymous substitutions was calculated in a pairwise fashion between *C. grandiflora* and *C. bursa-pastoris* A and B (Table S2). Under an allopolyploidy model we would expect a higher number of synonymous substitutions between *C. grandiflora* and *C. bursa-pastoris* B than between *C. grandiflora* and *C. bursa-pastoris* A. We do of course observe this since we have used the minimum number of synonymous substitutions to *C. grandiflora* to assign alleles to the A and B genomes, assigning the more distant allele to the B genome. However, for most loci, there is only a slight difference in this quantity between homeologues, suggesting that the two homeologues are nearly equal in their distance from standing *C. grandiflora* haplotype variation. Furthermore, the minimum number of synonymous substitutions between the two *C. bursa-pastoris* genomes is higher than either comparison with *C. grandiflora* as previously observed (Slotte et al. 2006). Likewise, we observe 29 fixed synonymous differences between *C. bursa-pastoris* A and B compared with 2 between *C. grandiflora* and *C. bursa-pastoris* A and 19 between *C. grandiflora* and *C. bursa-pastoris* B (Figure 2). The cause of this large difference in fixed sites observed between the *C. bursa-pastoris* genomes is likely their small effective population size causing alleles to drift to fixation quickly. On the other hand, the large effective population size of *C. grandiflora* would allow the maintenance of many shared alleles with both *C. bursa-pastoris* genomes.

Looking at the pattern of fixed differences between homeologues in *C. bursa-pastoris* reveals a striking pattern; 43% of fixed differences between homeologues are segregating with our *C. grandiflora* sample. Furthermore, if we restrict this to the 7 genes with large *C. grandiflora* samples (>20 chromosomes), this fraction increases to 52%. This retention of *C. grandiflora* polymorphism as fixed differences between homeologues in *C. bursa-pastoris* is consistent with an autopolyploid model, where distinct haplotypes sampled from the ancestral *C. grandiflora* population were ‘frozen’ as gene duplicates during polyploidization. Under this scenario, the remaining fixed differences would reflect rare SNPs not sampled in *C. grandiflora* and/or new mutations and fixation events following speciation. Considerably fewer fixed differences between homeologues are still segregating in *C. rubella* (20%).

In terms of identical haplotypes, we identified identical haplotypes between *C. bursa-pastoris* A and the other two species for all but three of our loci. Of the genes showing haplotype sharing 4 loci showed sharing with both species, 4 showed sharing only with *Capsella rubella*, and 2 showed sharing with *C. grandiflora* alone. Although the excess haplotype sharing in *C. rubella* is consistent with the inference of introgression (Slotte et al. 2008), it is important to note that the requirement of inferring phase in *C. grandiflora* and extensive recombination may erode some of the signal of haplotype sharing. Indeed, for the seven loci where we have relatively large *C. grandiflora* sample sizes for better inferences of phased haplotypes, only one locus shows *C. rubella* only haplotype sharing, and for this one it is only a single *C. rubella* individual that shows the shared haplotype.

We estimated the species tree of the *Capsella* genus using the program BEST, which implements a Bayesian hierarchical model while accounting for the presence of deep coalescences (Liu 2008). The analysis was performed twice, first by including *A. thaliana* as an outgroup (Figure S2-A) and second by including both *A. thaliana* and *Neslia paniculata*, where available, as outgroups (Figure S2-B). *C. grandiflora* was not shown to be more closely related to either *C. bursa-pastoris* A or B in either of the resulting trees. In fact, the tree resulting from the first analysis is the expected tree under an autopolyploidy model where the branch lengths between the two *C. bursa-pastoris* genomes and *C. grandiflora* are equal. Despite biasing our analysis toward the allopolyploidy model our results thus lend support to the autopolyploidy hypothesis.

Demographic model fitting: MIMAR and ms simulations

We used the program MIMAR (Becquet and Przeworski 2007) to fit models of isolation with migration in a pairwise fashion to *C. grandiflora* and *C. bursa-pastoris* A and B. The model assumes that a single ancestral population of size N_a splits into two descendant populations at time t, and the two descendant populations have distinct population sizes. Models including symmetric migration, asymmetric migration and no migration between the two derived populations were analysed for all three species pairs.

The results, however, show no evidence for migration between *C. grandiflora* and *C. bursa-pastoris*, so we only report the results from analyses assuming no migration between descendant populations.

Mimar runs that included gene flow, both between the *C. bursa-pastoris* homeologues and from *C. bursa-pastoris* to *C. grandiflora*, showed modes that approached zero (Table S4), providing little evidence for extensive gene conversion between homeologues and/or introgression from *C. grandiflora* following divergence. We therefore focus the presentation of the results on the no-migration model, although all results are reported in Table S4. C. *bursa-pastoris* A and B show a 5- and 7- fold decrease in effective population size, respectively, compared with *C. grandiflora* (Figure 3A; Table S4), with effective population sizes around 50,000-80,000 for *C. bursa-pastoris* A and B and values around 410,000 for *C grandiflora*, if we assume a mutation rate of 1.5×10^{-8} /site/year (Koch et al. 2000). The estimated time of divergence between each pair of genomes were 278,000 years between *C. grandiflora* and *C. bursa-pastoris* A, 1.1 million years between *C. grandiflora* and *C. bursa-pastoris* B and 563,000 years between the two *C. bursa-pastoris* genomes (Figure 3). It is not unexpected that the divergence time is much older between *C. grandiflora* and *C. bursa-pastoris* B compared with the *C. grandiflora* and *C. bursa-pastoris* A divergence time, since we have biased our analysis toward finding this result. What is striking is that the divergence time estimate between the two *C. bursa-pastoris* genomes is intermediate between the other two estimates, and the 90% highest posterior density (HPD) overlaps the HPD intervals between *C. grandiflora* and both *C. bursa-pastoris* homeologues. Under an allopolyploidy model the divergence time between the two *C. bursa-pastoris* genomes should be the same as the divergence between *C. grandiflora* and *C. bursa-pastoris* B, and significantly different from divergence between *C. grandiflora* and *C. bursa-pastoris* A. This suggests that the true divergence between C. *grandiflora* and both *C. bursa-pastoris* copies reflects an autopolyploid event about 563,000 years ago.

To further test whether the data fit an autopolyploid model we used test statistics based on shared and fixed sites. We calculated these summary statistics for both the observed data and the data simulated under both models. Most of the differences between the two models are confined to the fixed and shared sites (Table S5). We calculated two further statistics, the differences in both the number of fixed and the number of shared polymorphic sites between *C. grandiflora* and *C. bursa-pastoris* B, on the one hand and *C. grandiflora* and *C. bursa-pastoris* A, on the other hand. We used our two statistics to test for significant departures from the autopolyploid and allopolyploid models. Neither statistics in our observed data depart significantly from the simulated values under the autopolyploid model (P = 0.4339 for fixed differences and P = 0.3673 for shared differences) while both depart significantly under the allopolyploid model (P = 0.0032 for fixed differences and P = 0.0008 for shared differences) (Figure 4). We therefore cannot reject the autopolyploid hypothesis, while we can reject the allopolyploid model.

Demographic model fitting: Approximate Bayesian computation

Model 2 (allopolyploidy) of our two-split analysis failed to converge in the initial run, making it impossible to continue on to the ABC run. Model 1 (autopolyploidy), however, did produce usable samples indicating that this model fits better our data than model 2. Furthermore, the posterior distributions of most parameters have clear modes, showing that the data is informative for this model (Figure 5). The point estimates of the current population size of the A and B genomes of *C. bursa-pastoris* are 15,000 and 22,000 respectively (90% CR: 12,000-23,000 for *C. bursa-pastoris* A and 1,500-43,400 for *C. bursa-pastoris* B), while the estimate for *C. grandiflora* is 91,000 (90% CR: 32,600-162,000). The date of the first divergence event, between *C. bursa-pastoris* A and *C. bursa-pastoris* B, is 649,000 years (90% CR: 314,000-1,187,000 years), when assuming a generation time of 1 year and a mutation rate of 1.5×10^{-8}. The date of the second divergence event (739,000 years, 90%CR 361,000-1,443,000) is close to the time of the first divergence event suggesting that the A and B genomes of *C. bursa-pastoris* diverged from each other at a relatively similar time to when they diverged from *C. grandiflora*, thereby strongly supporting an autopolyploid origin of *C. bursa-pastoris*. It was not possible to estimate the population sizes after each coalescent event as the posteriors of these parameters were not informative. Goodness-of-fit tests indicate that the resulting model fit our data reasonably well. We calculated Tajima's D, θw and $\theta\pi$ for each genome from our goodness-of-fit simulations and S_{nn}, G_{ST} and K_{ST} among the genomes using Seqlib's build-in goodness-of-fit test and found that all summary statistics fit our data (two-tailed P-value > 0.05) except for Tajima's D (supplementary file S3). The reduced fit to Tajima's D may be reflective of population expansion following divergence.

To explore the possible impact of introgression events between *C. bursa-pastoris* and *C. rubella* on our inferences, the same analysis, using model 1, was performed using only Chinese *C.bursa-pastoris* samples, which were previously inferred to not be subject to introgression (Slotte et al. 2008). Introgressed alleles would be expected to decrease the divergence time between *C. grandiflora* and *C. bursa-pastoris*. Although the point estimates of the two divergence times were older for this analysis than for the total dataset, the 90% CR was extremely wide and overlapping with time estimates from the full dataset. However, this analysis was not very informative because the divergence time parameters and several other parameter estimates from this analysis had very wide 90% CRs, or/and had no clear mode. Importantly the posterior of the date of the first coalescent event encompasses the prior for this parameter (Figure S3). This may be due to lack of data in the Chinese samples, which have much less diversity than the European ones. This is probably due to the recent origin of the Chinese *C. bursa-pastoris* populations (Slotte et al. 2008). In fact, this reduction in diversity is supported by our Chinese-only ABC analysis, as θ is one of the few well-inferred parameters of the model (effective population size of Chinese *C. bursa-pastoris* 4550, 90% CR: 3,383-12,033)

Gene conversion and interlocus recombination

The results indicating a lack of gene flow between homeologues suggest that there has not been extensive gene conversion and/or historical recombination events, but we also conducted explicit tests for this. None of our loci showed evidence for gene conversion between *C. bursa-pastoris* homeologues using the geneconv software. However, two highly polymorphic loci, DFR (R_m=4) and At4g14190 (R_m=2), showed non-zero minimum number of recombination events between the two homeologues, suggesting the possibility of some level of interlocus gene conversion. Given that these loci, particularly DFR, are highly polymorphic in the diploid species, it is possible that the recombination events may have originated in the ancestral population rather than be due to homeologous gene conversion. Indeed, one of the recombination events in At4G14190 is also present in *C. grandiflora* (data not shown).

DISCUSSION

So far, it has proven difficult to establish whether *C. bursa-pastoris* is an allopolyploid or an autopolyploid. Various studies have resulted in often-conflicting theories as to the evolutionary origins of *C. bursa-pastoris,* some lending support to an allopolyploid origin (Hurka et al. 1989; Slotte et al. 2006) and others to an autopolyploid one (Hurka and Neuffer 1997). Because divergence is recent and extensive shared polymorphisms persist, a coalescent-based approach incorporating population samples and multilocus nuclear data becomes crucial to accurately distinguish models of polyploid speciation. In the present study we used sequence polymorphism and divergence at 14 nuclear loci and two different coalescent-based approaches to test whether *C. bursa-pastoris* had an autopolyploid or allopolyploid origin.

We conducted three types of analysis to investigate the two possible origins of *C. bursa-pastoris*. First we examined the diversity among the three *Capsella* species and inferred their phylogeny using the program BEST. Second we estimated the parameters of an isolation-with-migration model with the program MIMAR and used these estimates to conduct coalescent simulations under both models. Finally, we used Approximate Bayesian Computation to estimate parameters of two-split models representing our null and alternative hypotheses. We could not reject an autopolyploid origin of *C. grandiflora* in any of these analyses, whereas our results were inconsistent with an allopolyploid model. Based on our analyses, the lower bound of the time of origin of *C. bursa-pastoris* is between 270,000 and 700,000 years ago. *C. bursa-pastoris* would thus still be much older than *C. rubella* which most likely diverged from *C. grandiflora* less than 50,000 years ago (Foxe et al. 2009; StOnge et al. 2011) allowing us to rule out the suggestion that *C. bursa-pastoris* could be an allopolyploid of *C. rubella* and *C. grandiflora* (Hurka et al. 1989) in agreement with the conclusion of Slotte et al. (2006). Even if these time estimates should be taken with a grain of salt given the uncertainty around mutation rates (Beilstein et al. 2010; Ossowski et al. 2010) a rather recent autopolyploid origin would be consistent with the low level of diversity in *C. bursa-pastoris*, and it would also mean that disomic inheritance has evolved quite rapidly in this species. The ABC analysis indicates that the divergence time of the two homeologueous chromosomes of *C. bursa-pastoris* is very close to the divergence between *C. grandiflora* and *C. bursa-pastoris*, suggesting that if there was a period of tetrasomic inheritance it was short relative to the age of the tetraploid species. It has been shown in other species that polyploids with tetrasomic segregation (pairing of four homologous chromosomes during meiosis) tend to rediploidize over time as mutations accumulate and chromosomes diverge (Ramsey and Schemske 1998; Soltis et al. 2010). This process can indeed occur rather quickly and diploidization can proceed through structural rearrangements within 30 generations in *A. thaliana* (Parisod et al. 2010). Furthermore, autopolyploids with small chromosomes or low chiasma frequencies may exhibit disomic inheritance immediately after their formation (Stebbins 1971). It is also possible that autopolyploid formation from a highly diverse ancestral population such as *C. grandiflora*, may enhance the speed at which disomic inheritance can occur.

Many polyploid species have multiple origins (Soltis et al. 2003). In a previous study Slotte et al. (2006) argued that the low nucleotide diversity observed for cpDNA sequences and at seven chloroplast microsatellite loci supports a single origin of *C. bursa-pastoris*. The chloroplast sequences resulted in a strongly supported phylogeny in which *C. bursa-pastoris* is sister to both diploid species. This topology is consistent with an ancient origin of *C. bursa-pastoris* from *C. grandiflora* given the fact that *C. rubella* derived from *C. grandiflora* much more recently. The level of variation in *C. bursa-pastoris* across the 14 loci is similarly low, and is a consequence of a 5-7 fold decrease of the effective population size compared to *C. grandiflora*. This reduction is not as severe as the reduction in population size observed in *C. rubella* (100-1,500 fold reduction, Foxe et al. 2009; 18 fold reduction, St.Onge et al. 2011). This may be the result of a combination of factors. Recurrent polyploid formation would increase genetic variation but would not leave such a strong bottleneck signature; while this might seem to contradict the lack of variation observed in cpDNA, this could reflect subsequent coalescent events in the chloroplast following species formation (Ceplitis et al. 2005; Slotte et al. 2006). Alternatively, the severity of the bottleneck could have been lessened by early gene flow from *C. grandiflora* via pollen, which would not affect diversity in cpDNA. A third alternative is that the evidence for a severe population bottleneck might simply have eroded with time as the divergence of *C. bursa-pastoris* from *C. grandiflora* is much older than the divergence of *C. rubella* from *C. grandiflora*; a more detailed model of small founding population size followed by a recovery in population size is likely equally consistent with the data, and might explain our observed negative values of Tajima's D.

Gene conversion can have a strong impact on the histories of duplicated genes and genomes (e.g. Osada and Innan 2008) and, in principle, extensive gene conversion in *C. bursa-pastoris* could also have affected our results. Extensive gene conversion could theoretically cause an allopolyploid genome to appear as an autopolyploid under our analysis. However, for this to have happened in *C. bursa-pastoris* the amount of gene conversion would have had to be very extensive, which seems highly unlikely. We identified only two of our loci with evidence of interlocus recombination using the minimum number of recombination events, and no evidence for gene conversion using geneconv. Furthermore, the loci showing gene conversion are highly polymorphic in the diploid species, raising the possibility that the identified recombination events could be due to their retention from ancestral polymorphism and/or due to introgression events. Even though gene conversion is unlikely to have been potent enough to alter our conclusion it might still have contributed to the pattern of divergence among the different genomes. Assuming autopolyploidization and speciation occurred simultaneously we would expect the A and B genomes of *C. bursa-pastoris* to split from *C. grandiflora* at the same time. However, we observe a slight gap in the mean values of these dates. This could be caused by early gene conversion between the A and B genomes, making them appear to be slightly more recently diverged from each other than either is to *C. grandiflora* although a period of initial tetrasomic inheritance, as discussed previously, might be a more parsimonious explanation. Overall, the similar divergence times between homeologues and *C. grandiflora* make long periods of disomy and/or gene conversion unlikely.

Another factor that might have influenced our results is introgression. Previous work has identified evidence of introgression from *C. rubella* to *C. bursa-pastoris* (Slotte et al. 2008). Evidence for introgression was detected in European populations of *C. bursa-pastoris* but was absent in China where *C. rubella* is absent. Since these introgressed alleles would generally be grouped with the A genome, they are expected to increase the divergence between the A and B genomes of *C. bursa-pastoris* and thereby favor our allopolyploid hypothesis. Introgression is therefore not expected to alter our conclusion that *C. bursa-pastoris* has an autopolyploid origin. It would, however, be expected to cause the inferred divergence date between *C. bursa-pastoris* and *C. grandiflora* to be younger. To examine the possible role of introgression from *C. rubella*, and to confirm our general conclusions using *C. rubella* instead of *C. grandiflora*, we conducted *mimar* analysis with asymmetrical gene flow for *C. rubella* and both *C. bursa-pastoris* A and B. Parameter estimates for these runs had particularly wide confidence intervals, likely due, at least in part, to the loss of information on ancestral polymorphism caused by the severe bottleneck in *C. rubella*. Nevertheless, the results are consistent with our previous conclusions: divergence estimates between *C. bursa-pastoris* A and B fall in between the divergence times estimated between *C. bursa-pastoris* A and *C. rubella* (mode: 66,066, 95% HPD: 22022-3.9 million years) and *C. bursa-pastoris* B and *C. rubella* (mode: 3.1 million years, 95% HPD: 2.2 million-4.0 million years). Furthermore, simulations of autopolyploid models of the observed data conform well to our observed comparisons of *C. rubella* to *C. bursa-pastoris*, while we get higher rejection rates for the allopolyploid model (Supplementary file S2). To further test if the inferred divergence times were being affected by putatively introgressed alleles we conducted an ABC analysis using only the Chinese samples. Although this analysis was not very informative, the 90% CR of the first inferred divergence time using only China's *C. bursa-pastoris* samples was overlapping with the estimate from the total dataset, suggesting that introgression from *C. rubella* into *C. bursa-pastoris* did not have a strong impact on our conclusion. Finally, the patterns of haplotype sharing do not indicate that extensive introgression from *C. rubella* is likely to greatly influence our analysis; haplotype sharing was generally comparable for both diploid species. With genome-wide data from large samples of all three species, it will be interesting to re-examine the extent to which haplotype sharing reflects ancestral polymorphism vs. gene flow following speciation.

It is important to note that all of our modelling approaches focus on a simplified model of speciation and divergence, and it is possible that additional model mis-specifications, particularly in the allopolyploid model, could be leading to a higher rejection rate. For example, Mimar assumes a single population size change following divergence and a constant migration rate, and subsequent population size changes and/or changes in gene conversion rates between homeologues over time could be complicating our inferences. However, our simulations lead us to conclude that the autopolyploid model can explain our data quite well, and it is not obvious why model mis-specification would be a problem specific only to the allopolyploidy model. Nevertheless, it will be important to confirm out conclusions with large-scale genomic data, where the patterns of haplotype structure and divergence across chromosomes can also be incorporated into these analyses.

CONCLUSIONS

Our study confirms the usefulness of coalescent-based approaches when studying the mode of origin of polyploids, although as pointed by Doyle and Egan (2009) precise time estimates remain elusive and are highly dependent on demographic details and on assumptions on mutation rates. While these results shed much light on the evolutionary origin of *C. bursa-pastoris*, little is still known about the extensive phenotypic changes that have occurred in both *C. bursa-pastoris* and *C. rubella*. Understanding the genomic context and underlying evolutionary forces that have promoted these changes will be of considerable interest in future studies.

Acknowledgements: We thank Anna E. Palmé for helpful comments on earlier versions of this manuscript. ML thanks the Swedish Research Council and the Philip Sörensen Foundation for support and the Chinese Academy of Sciences for funding a visiting professorship. TS thanks the Liljewalch and Sernander foundations at Uppsala University for funding to conduct sampling of *C. grandiflora*. KS would like to thank the Nilsson-Ehle-fonden, the Royal Swedish Academy of Sciences and the Sernander foundation at Uppsala University. SW thanks the Natural Sciences and Engineering Research Council (NSERC) for funding.

References

Bailey CD, Koch MA,_Mayer M,_Mummenhoff K, O'Kane Jr SL,Warwick SI, Windham MD, Al-Shehbaz IA. 2006. Toward a global phylogeny of the Brassicaceae. Mol. Biol. Evol. 23:2142–2160.

Beaumont MA. 2010. Approximate Bayesian Computation in evolution and ecology. Annu. Rev. Ecol. Evol. Syst. 41: 379–406.

Beaumont M, Zhang W, Balding D. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025-2035.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proc. Natl. Acad. Sci., USA 107: 18724-18728.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Res 17:1505-1519.

Ceplitis A, Su Y, Lascoux M. 2005. Bayesian inference of evolutionary history from chloroplast microsatellites in the cosmopolitan weed *Capsella bursa-pastoris* (Brassicaceae). Molecular Ecology 14:4221-4233

Cifuentes M., Grandont L, Moore G, Chèvre AM, Jenczewski E. 2010. Genetic regulation of meiosis in polyploid species: new insights into an old question. New Phytol. 186: 29-36.

Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738 –749.

De Mita S, Ronfort J, McKhann HI, Poncet C, El Malki $_{R,}$ Bataillon T. 2007. Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in *Medicago truncatula*. Genetics 177: 2123–2133.

Doyle JJ, Egan AN. 2009. Dating the origins of polyploidy events. New Phytol. 186:73-85

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. Proc Natl Acad Sci, USA. 106: 5737–5742.

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Rapid morphological evolution and speciation associated with the evolution of selfing in *Capsella*. Proc Natl Acad Sci, USA. 106:5241-5245.

Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci, USA. 94: 6809-6814.

Guo YL, Bechsgaard J, Slotte T, Neuffer B, Lascoux M, Weigel D, Schierup MH. 2009. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proc. Natl. Acad. Sci., USA 106:5246-5251.

Hey J. 2010. Isolation with migration models for more than two populations. Mol Biol Evol. 27: 905-920.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci, USA. 104: 2785-2790.

Hudson RR 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Hudson, RR, Kaplan N. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA-sequences. Genetics, 111: 147–164.

Hurka H, Freundner S, Brown AH, Plantholt U. 1989. Aspartate aminotransferase isozymes in the genus *Capsella* (Brassicaceae): subcellular location, gene duplication, and polymorphism. Biochemical genetics 27:77-90.

Hurka H, Neuffer B. 1997. Evolutionary processes in the genus *Capsella* (Brassicaceae). Plant Syst Evol. 206:295-316.

Jakobsson M, Hagenblad J, Tavaré S, Säll T, Halldén C, Lind-Halldén S, Nordborg M. 2006. A unique recent origin of the allotetraploid species *Arabidopsis suecica*: Evidence from nuclear DNA markers. Mol Biol Evol. 23:1217-1231.

Koch M, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol Biol Evol 17:1483-1498.

Librado, P, Rozas, J 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24: 2542-2543.

Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in the majority of angiosperms. Science 264: 421–423.

Nicholas KB, Nicholas HB Jr., Deerfield DW II. 1997 GeneDoc: Analysis and Visualization of Genetic Variation, EMBNEW.NEWS 4:14

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885-896.

Noor MA, Feder JL. 2006. Speciation genetics: evolving approaches. Nature Reviews Genetics 7:851-861.

Osada, N., Innan I. 2008. Duplication and gene conversion in the *Drosophila melanogaster* genome. PLoS Genet 4(12): e1000305

Ossowski S, Schneeberger K, Lucas-Lledó JI*,* Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327: 92-94.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Ann. Rev. Genet. 34:401-437.

Parisod C., Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. New Phytol. 186:5-17.

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aguadé M. 2004. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics 166:373-388.

Ramsey J, Schemske D. 1998. Pathways, mechanisms and rates of polyploid formation in flowering plants. Ann. Rev Ecol Syst. 29:467-501.

Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth

D, Gaut BS. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS One. 6:e2411.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 132:365–386.

Sawyer S. 1989. Statistical test for detecting gene conversion. Mol. Biol. Evol. 6: 526-538.

Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156: 879-891.

Shull GH. 1929. Species hybridization among old and new species of Shepherd's Purse. Int Congr Plant Sci. 1:837–888.

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 93:1714-1724.

Slotte T, Huang H, Lascoux M, Ceplitis A. 2008. Polyploid speciation did not confer instant reproductive isolation in *Capsella* (Brassicaceae). Mol Biol Evol. 25:1472-1481.

Soltis DE, Soltis PS, Tate JA. 2003. Advances in the study of polyploidy since Plant Speciation. New Phytol 161:173-191.

Soltis DE, Buggs RJA, Doyle JJ, Soltis PS. 2010. What we still don't know about polyploidy. Taxon. 59:1387-1403.

Stebbins GL. 1971. Chromosomal Evolution in Higher Plants. Edward Arnold: London.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978-989.

StOnge K, Källman T, Slotte T, Lascoux M, Palmé AE. 2011. Divergent population history and structure in two closely related species (*Capsella rubella* and *C. grandiflora*) with different mating systems. Molecular Ecology 20: 3306–3320.

Tajima F. 1983. Evolutionary relationships of DNA-sequences in finite populations. Genetics 105: 437-460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

U, N. 1935. Genome analysis in Brassica with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Jpn J Bot 7: 389–452.

Velasco R, and 86 co-authors 2010. The genome of the domesticated apple (*Malus* • *domestica* Borkh.). Nature Genetics 42:833-841.

Wakeley J., Hey J. 1997. Estimating ancestral population parameters. Genetics 145:847-855.

Wang RL, Wakeley J, Hey J. 1997. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. Genetics 147:1091-1106.

Wang W-K, Ho C-W, Hung K-H, Wang, K-H, Huang, C-C, Araki, H, Hwang, C-C, Hsu, T-W, Osada N, Chiang, T-Y (2010) Multilocus analysis of genetic divergence between outcrossing Arabidopsis species: evidence of genome-wide admixture. New Phytol 188: 488-500

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Pop Biol 7: 256-276.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci USA 106:13875-13879

Zhang L-B, Ge S. 2007. Multilocus analysis of nucleotide variation and speciation in *Oryza officinalis* and its close relatives. Mol Biol Evol 24:769-783.

Figure text

Figure 1. Model diagrams of the null hypothesis, autopolyploidy, and alternative hypothesis, allopolyploidy.

Figure 2. Number of synonymous fixed differences between all pairs of *C. bursa-pastoris* A, *C. bursa-pastoris* B, *C. grandiflora* and *C. rubella.*

Figure 3. Marginal posterior distributions of speciation parameters estimated by MIMAR, with posterior modes showing good fit to data summaries. $\theta = 4Ne\mu$ where *Ne* is the effective population size and μ is the mutation rate (1.5 × 10^{-8}/site/year)

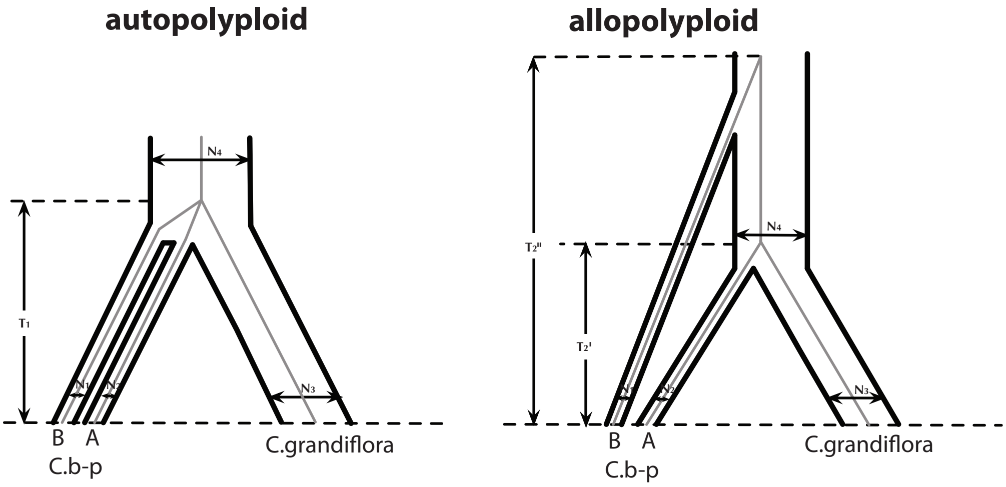
A) Constrained model: the model assumes equal effective population sizes in the ancestor as in present-day *C.grandiflora*: Model 1; Species 1 = *C. grandiflora*, Species 2 = *C. bursa-pastoris A*. The model is represented by continuous lines. Model 2; Species 1 = *C. grandiflora*, Species 2 = *C. bursa-pastoris B.* The model is shown by a dotted line. Tgen Divergence time (years) between *C. grandiflora* and *C. bursa-pastoris* A and between *C. grandiflora* and *C. bursa-pastoris* B

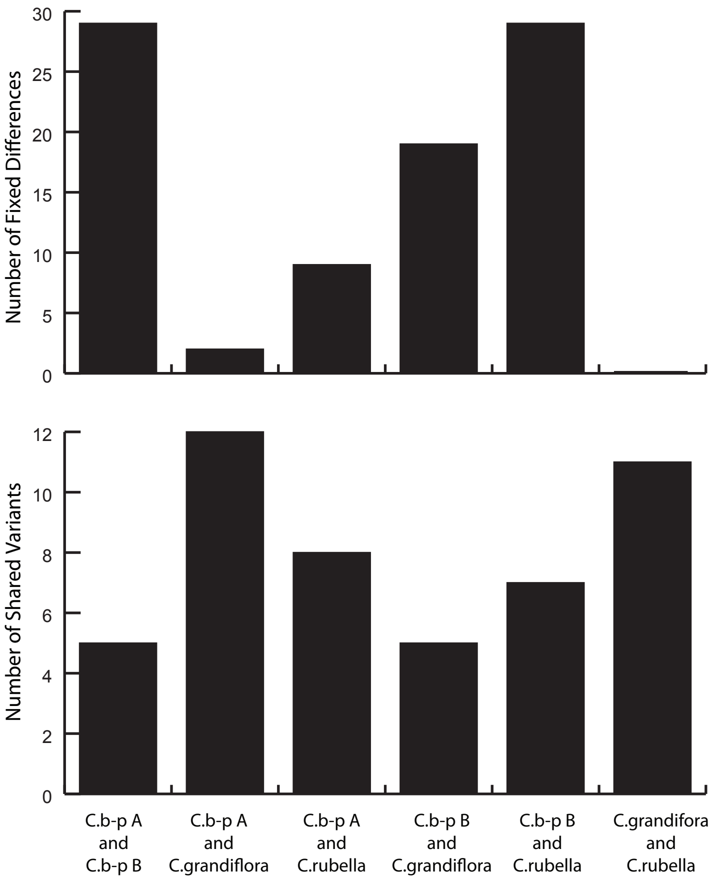
B) Unconstrained model: θA ancestral *C. grandiflora*, θ1 *C. bursa-pastoris* A (continuous line), θ2 *C. bursa-pastoris* B (dotted line). Tgen Divergence time (years) between *C. bursa-pastoris* A and *C. bursa-pastoris* B.

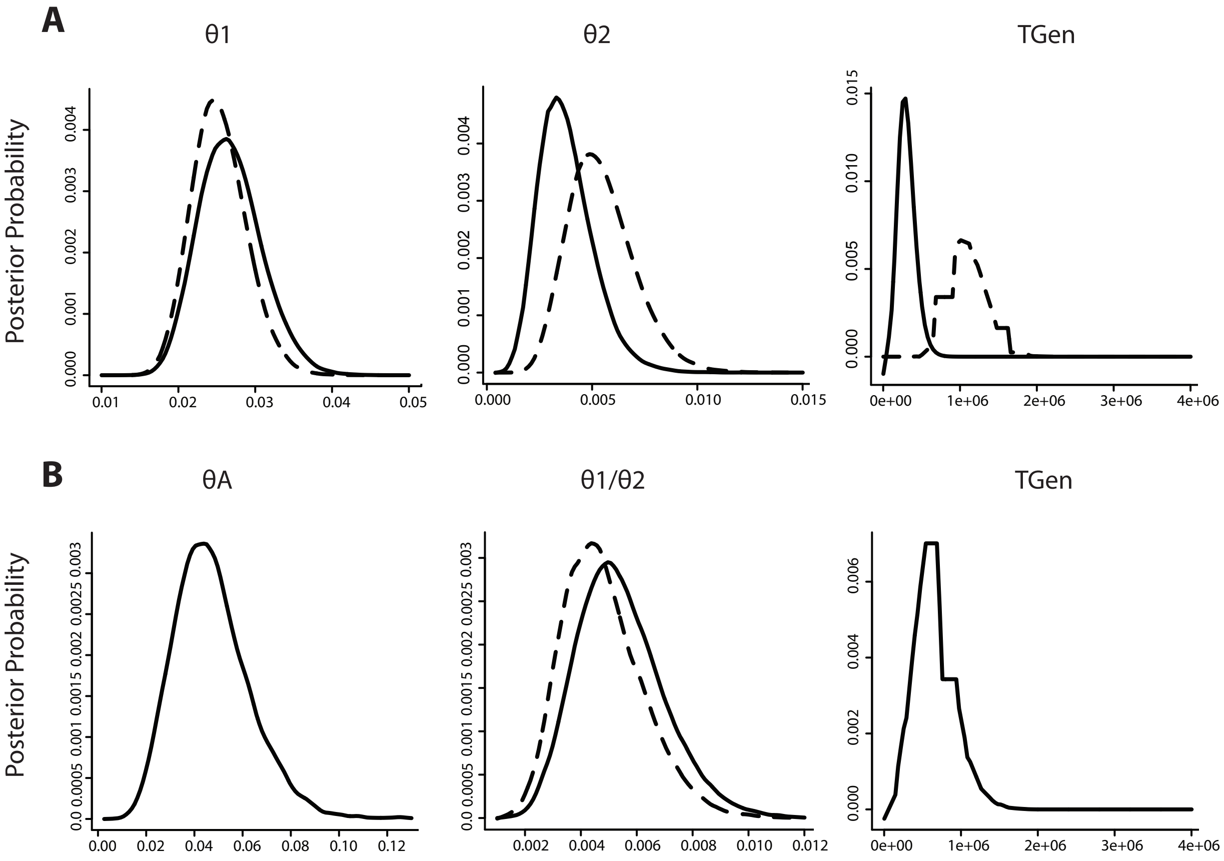
Figure 4: Density distribution of the simulated values of the summary statistics under (A) autopolyploidy and (B) allopolyploidy. The left column gives the distribution of the mean of *fix_diff* over the fourteen genes, where *fix_diff* is the difference between the number of fixed sites of each of the homoelogues when it is compared to *C. grandiflora*. The right column gives the same for *shared_diff,* the difference between the number of shared polymorphic sites of each of the homoelogues to when it is compared to *C. grandiflora*. The blue vertical line is the observed value. P values are given in the upper right corner of each plot. See text for details.

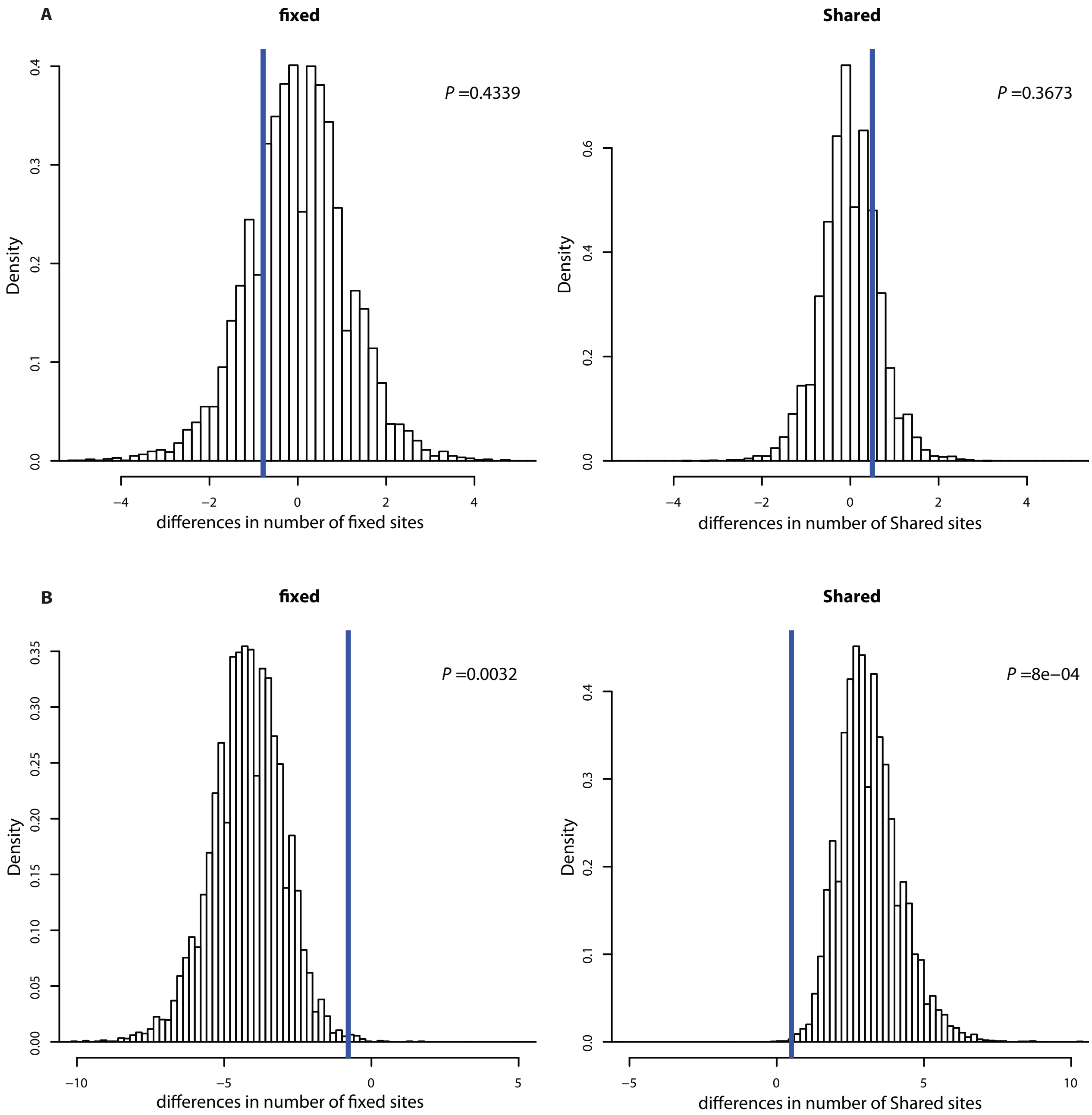
Figure 5. Posterior distributions of informative parameters in the two-split model where

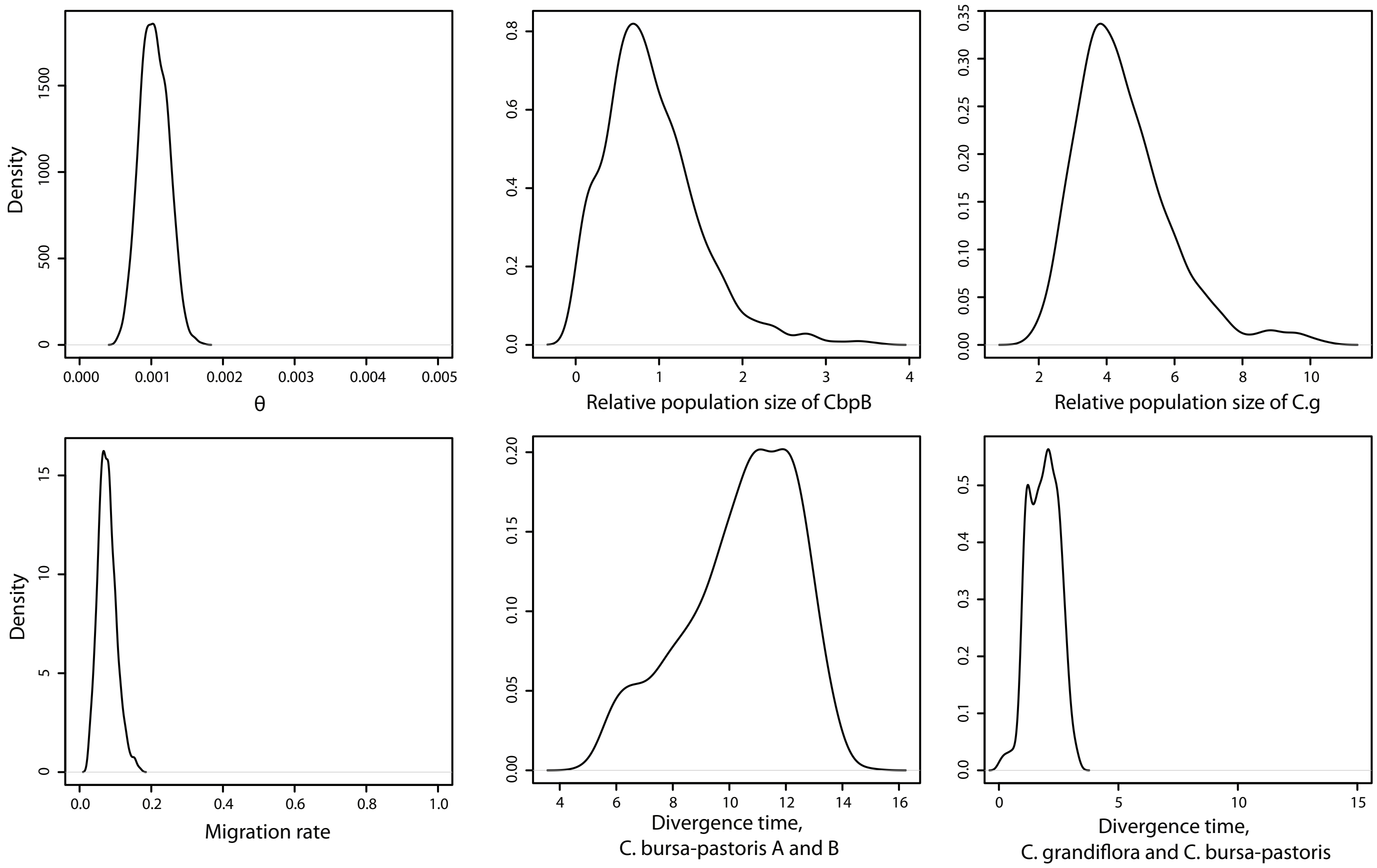
the two *C. bursa-pastoris* genomes coalesce first, followed by coalescence of their common ancestor with *C. grandiflora*. $\theta=4N_e\mu$ where N_e is the effective population size of *C. bursa-pastoris* A and μ is the mutation rate (1.5×10^{-8}/site/year). Other effective population sizes and the divergence times are relative to this first estimate. Divergence times are on a scale of $4N_e\times$generations and the second date parameter is additive to the first.


autopolyploid
N₄
T₁
N₁
N₂
N₃
B
A
C.b-p
C.grandiflora
allopolyploid
T₂''
T₂'
N₄
N₁
N₂
N₃
B
A
C.b-p
C.grandiflora


Number of Fixed Differences
30
25
20
15
10
5
0
Number of Shared Variants
12
10
8
6
4
2
0
C.b-p A and C.b-p B
C.b-p A and C.grandiflora
C.b-p A and C.rubella
C.b-p B and C.grandiflora
C.b-p B and C.rubella
C.grandifora and C.rubella


A
θ1
θ2
TGen
Posterior Probability
0.000
0.001
0.002
0.003
0.004
0.01
0.02
0.03
0.04
0.05
0.000
0.001
0.002
0.003
0.004
0.000
0.005
0.010
0.015
0.000
0.005
0.010
0.015
0e+00
1e+06
2e+06
3e+06
4e+06
B
θA
θ1/θ2
TGen
Posterior Probability
0.0
0.0005
0.001
0.0015
0.002
0.0025
0.003
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.0
0.0005
0.001
0.0015
0.002
0.0025
0.003
0.002
0.004
0.006
0.008
0.010
0.012
0.000
0.002
0.004
0.006
0e+00
1e+06
2e+06
3e+06
4e+06


A
fixed
P =0.4339
Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
differences in number of fixed sites
Shared
P =0.3673
Density
0.0
0.2
0.4
0.6
−4
−2
0
2
4
differences in number of Shared sites
B
fixed
P =0.0032
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
−10
−5
0
5
differences in number of fixed sites
Shared
P =8e−04
Density
0.0
0.1
0.2
0.3
0.4
−5
0
5
10
differences in number of Shared sites


Density
0
500
1000
1500
0.000
0.001
0.002
0.003
0.004
0.005
θ
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
Relative population size of CbpB
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
4
6
8
10
Relative population size of C.g
Density
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
Migration rate
0.00
0.05
0.10
0.15
0.20
4
6
8
10
12
14
16
Divergence time,
C. bursa-pastoris A and B
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
Divergence time,
C. grandiflora and C. bursa-pastoris