Research Report
Biological Sciences: Evolution

Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*

Gavin Douglas[a,1], Gesseca Gos[a,1], Kim A. Steige[b,1], Adriana Salcedo[a], Karl Holm[b], J. Arvid Ågren[a], Khaled M. Hazzouri[a,c], Wei Wang[a], Adrian E. Platts[d], Emily B. Josephs[a], Robert J. Williamson[a], Barbara Neuffer[e], Martin Lascoux[b,2], Tanja Slotte[b,f,2], Stephen I. Wright[a,]

1 These authors contributed equally to the work

2 authors for correspondence, email:

stephen.wright@utoronto.ca

tanja.slotte@su.se

martin.lascoux@ebc.uu.ce

a Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Canada

b Department of Ecology and Genetics, Evolutionary Biology Centre, Science for life Laboratory, Uppsala University, Uppsala, Sweden

c Center for Genomics and Systems Biology, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates.

d McGill Centre for Bioinformatics, McGill University, Montreal, Canada

e Department of Botany, University of Osnabrück, Osnabrück, Germany

f Department of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

Keywords: polyploidy, population genomics, speciation, gene loss

Whole genome duplication events have occurred repeatedly during flowering plant evolution, and there is growing evidence for predictable patterns of gene retention and loss following polyploidization. Despite these important insights, the rate and processes governing the earliest stages of diploidization remain uncertain, and the relative importance of genetic drift vs. natural selection in the process of gene degeneration and loss is unclear. Here we conduct whole genome resequencing in *Capsella bursa-pastoris*, a recently formed tetraploid with one of the most widespread species distributions of any angiosperm. Whole genome data provide strong support for recent hybrid origins of the tetraploid species within the last 100-300,000 years from two diploid progenitors in the *Capsella* genus. Major-effect inactivating mutations are frequent, but many were inherited from the parental species and show no evidence of being fixed by positive selection. Despite a lack of large-scale gene loss, we observe a shift in the efficacy of natural selection genome-wide. Our results suggest that the earliest stages of diploidization are associated with quantitative genome-wide shifts in the strength and efficacy of selection rather than rapid gene loss, and that nonfunctionalization can receive a 'head start' through deleterious variants found in parental diploid populations.

SIGNIFICANCE

Plants have undergone repeated rounds of whole genome duplication, followed by gene degeneration and loss, with only a minority of genes eventually retained as duplicates. Using whole genome resequencing we examined the origins and earliest stages of genome evolution in the recent tetraploid *Capsella bursa-pastoris*. We conclude that the species has a hybrid origin from two distinct lineages in the *Capsella* genus within the last 100-300,000 years. Our analyses suggest the absence of rapid gene loss upon polyploidization, but provide evidence that the species has large numbers of inactivating mutations, many of which were inherited from the parental species. We detect a reduction in the strength of negative selection against deleterious mutations in the polyploid.

INTRODUCTION

There is growing evidence for substantial differences across populations and species in the strength of selection at the molecular level (1). Both estimates of purifying and positive selection differ strongly across taxa, highlighting the importance of understanding the factors influencing the strength, direction and sign of selection genome-wide (2, 3). In flowering plants, mating system shifts and polyploidy represent major repeated transitions that are thought to have a significant impact on the genome-wide strength and efficacy of selection (4-6). However, our understanding of the combined and unique effects of these transitions on genome diversity and selection remains limited.

While theory predicts a reduced efficacy of natural selection in self-fertilizing species (5, 6), the general adaptive significance of polyploidy is unclear. As polyploidy results in gene

duplication across the entire genome, it has long been thought to contribute raw material for the evolution of novel gene function (7). In particular, polyploids may experience more adaptive evolution than their diploid ancestors, because unconstrained gene duplicates provide raw material for natural selection to act on (8). Polyploids are also expected to have higher levels of heterozygosity and reduced inbreeding depression, which may be beneficial (9). On the other hand, theoretical work has shown that any reduction in inbreeding depression will be transient, and polyploids should harbor higher mutational load and can experience less efficient selection than diploid species (4). Finally, epigenetic and genetic changes can occur rapidly upon polyploid formation, but whether this contributes to polyploid establishment is not known (8, 9).

The vast majority of angiosperms have experienced whole genome duplication (WGD) due to polyploidization in their history. Following WGD, genomes undergo a process of rediploidization, where duplicate genes are lost, and only a minority of genes are retained as duplicates for extended periods of evolutionary time (10, 11). Genomic studies have documented ample variation in the time since the last WGD among plants (12-18). For instance, maize experienced a whole genome duplication event roughly 5-12 million years ago (19-21) and has retained 43% of its genes as duplicates (22, 23), while the *Arabidopsis* genome retained only about 20% of duplicate genes since its last WGD event over 20 million years ago (13, 15, 24-27). Given the important contribution of polyploidy to plant speciation (4, 28), this suggests that the time since the last polyploidization event may be a key parameter influencing variation in the strength and direction of selection across plant genomes.

Large-scale genome projects have provided a number of key insights into genome evolution in ancient polyploids (10, 12, 19). First, gene retention and loss is nonrandom with

respect to gene function, with signs of strong convergence in which genes are lost (10). This raises an open question about the extent to which gene loss proceeds solely via a process of relaxed purifying selection, or whether there is also positive selection for gene loss. Second, gene loss is often biased, such that genes from one chromosome duplicated by WGD (homeolog) are preferentially lost (12), or become expressed at lower levels than those on the other homeolog (19). This process is termed biased fractionation (22), and while the underlying mechanism remains unresolved, it could result from differences in epigenetic regulation or selective history of the parental species (24), and may occur rapidly upon polyploid formation or following a longer timescale of genome rearrangement and evolution.

While much of the work on genome evolution in polyploids has focused on ancient events, the early stages of gene degeneration and loss could provide important insights into the causes of genome evolution in polyploids (29, 30). Key open questions concern the role of drift vs. positive directional selection in early gene loss, the speed of early gene degeneration, and the role of past evolutionary history in the dynamics of genome evolution in polyploids. Comparative population genomics of polyploids and their progenitor species provides a key opportunity to better understand the dynamics associated with the early stages of polyploid evolution.

The genus *Capsella* includes the primarily self-fertilizing tetraploid *Capsella bursa-pastoris* (2n=4x=32) as well as three diploid species: the self-incompatible outcrosser *Capsella grandiflora*, the self-fertilizing *Capsella rubella*, and the self-compatible *Capsella orientalis* (2n=2x=16). These species differ greatly in their geographical distribution. The outcrosser *C. grandiflora* is limited to north-western Greece and Albania, whereas the selfer *C. rubella* has a

broader Mediterranean-central European distribution and the self-compatible *C. orientalis* is found from eastern Europe to central Asia (31). In contrast to its diploid congeners, the selfing tetraploid *C. bursa-pastoris* has a worldwide distribution that is at least in part anthropogenic (32, 33).

We have previously shown that the self-fertilizing diploid *C. rubella* is derived from the outcrossing, self-incompatible diploid *C. grandiflora*, and that these species split relatively recently, within the last 50,000-100,000 years ago (34-37). There is reason to believe that outcrossing is the ancestral state in this system, as *C. grandiflora* shows trans-specific shared polymorphism with *Arabidopsis lyrata* and *Arabidopsis halleri* at the self-incompatibility locus (35, 38, 39). The self-compatible *C. orientalis* is therefore also thought to be derived from an outcrossing, *C. grandiflora*-like ancestor, although further back in time than *C. rubella* (31).

The origin of the widespread tetraploid *C. bursa-pastoris* has however proven difficult to determine (33, 40). Various hypotheses on the origin of this species have been put forward (32, 40-44), and most recently, *C. bursa-pastoris* has been suggested to be an autopolyploid of *C. orientalis* (31), or *C. grandiflora* (32). Despite the hypotheses of autopolyploidy, *C. bursa-pastoris* appears to have strict disomic inheritance, with two separate homeologous genomes segregating independently (42).

In this study, we investigate the mode of formation and genomic consequences of polyploidization in *C. bursa-pastoris* using high-coverage massively parallel genomic sequence data. We first evaluate the origins of *C. bursa-pastoris* and conclude that the species has recent, hybrid, allopolyploid origins from the *C. orientalis* and the *C. grandiflora*/*C. rubella* lineages. By comparing genome-wide patterns of nucleotide polymorphism and gene inactivation between

C. bursa-pastoris and its two close relatives, *C. grandiflora* and *C. orientalis*, we then quantify the rate and sources of early gene inactivation following polyploid formation. Finally, we compare the genome-wide strength of selection in the diploids and the recently-derived allopolyploid. The results are important for an improved understanding of the consequences of polyploidization for the strength and direction of selection in plant genomes.

RESULTS and DISCUSSION

De novo assemblies and whole genome alignment across the Capsella *genus*

We generated *de novo* assemblies for *C. bursa-pastoris* and *C. orientalis* using Illumina genomic sequence (see Methods and SI Text). Scaffolds from these assemblies were then aligned to the *C. rubella* reference genome (37) using whole genome alignment, along with assemblies of *C. grandiflora* and the outgroup species *Neslia paniculata* (37). With the exception of *C. bursa-pastoris*, only a single orthologous chain was retained for each genomic region, while for *C. bursa-pastoris* we allowed the retention of two chains, based on the expectation of assembling sequences from the two homeologous chromosomes. In total, our *C. bursa-pastoris* assembly spans approximately 102MB of the 130MB *C. rubella* assembly, of which approximately 40% (42MB) is covered by two homeologous sequences, while the remainder is covered by a single sequence.

While it is possible that regions covered by a single *C. bursa-pastoris* sequence reflect large-scale gene deletions, it is likely that our assembly method often collapses the two homeologous sequences into a single assembled sequence, given the relatively low pairwise differences between the homeologs (average of 3.7% differences in assembled homeologs). For

investigating the genomic relationships across the *Capsella* genus, we therefore focus on the 42MB of the genome that contains two *C. bursa-pastoris* homeologs, and provide further support that these regions are representative of genome-wide patterns in follow-up polymorphism analyses below.

The evolutionary origin of C. bursa-pastoris

We constructed independent phylogenetic trees for each alignment from genomic fragments that included two homeologous regions of *C. bursa-pastoris*, and one each from *C. grandiflora*, *C. orientalis*, *C. rubella* and the outgroup sequence *N. paniculata*. We obtained 17,258 trees where all branches had at least 80% bootstrap support, 79% of which had one of the six topologies depicted in Figure 1. The total alignment length across these trees represents 13.6 MB of sequence across the genome. Of these trees, 70% showed clustering of one homeolog with the *C. grandiflora*/*C. rubella* lineage (Fig 1 A-C; hereafter the 'A' subgenome) and the second homeolog with *C. orientalis* (hereafter the 'B' subgenome), providing strong support for a hybrid origin of *C. bursa-pastoris*. Thus, we conclude that *C. bursa-pastoris* is an allopolyploid resulting from a hybridization event between these two lineages. As *C. bursa-pastoris* clusters with *C. orientalis* for maternally inherited chloroplast DNA (31), we conclude that the maternal parent of *C. bursa-pastoris* came from the *C. orientalis* lineage and the paternal contribution came from an ancestral population from the *C. grandiflora*/*C. rubella* lineage. Given the current disjunct distribution of *C. grandiflora* and *C. orientalis* (31), this implies that the ancestral range of either or both lineages must have been overlapping in the past, to allow for hybridization to occur.

To validate these findings, we analyzed a set of nine Sanger-sequenced loci where both *C. bursa-pastoris* homeologs were previously isolated using homeolog-specific PCR amplification techniques and Sanger-sequenced in a larger set of *C. bursa-pastoris* accessions covering the native range of *C. bursa-pastoris* in Eurasia (mean N=76, range 24-105). In all cases, one of the *C. bursa-pastoris* homeologs 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 (Figure S1).

Overall, our comparative 4-species genome analysis provides strong support for an allopolyploid origin for *C. bursa-pastoris*. To verify that these conclusions are not biased by a focus on genomic regions that have two homeologs assembled, and to investigate polyploid origins in more detail, we resequenced 9 additional genomes of both *C. bursa-pastoris* and *C. orientalis*. We compared single nucleotide polymorphism (SNP) sharing and divergence across these samples by mapping all reads to the reference *C. rubella* genome, and incorporated genome resequencing data from 13 *C. grandiflora* accessions from previous work (41, 43). Because we are mapping our tetraploid sequences to a diploid, we expect to see high levels of apparent 'heterozygosity'; as *C. bursa-pastoris* is highly selfing we expect the vast majority of these heterozygous sites to represent homozygous polymorphic differences between the two homeologous copies. Across the genome, both higher-order heterozygosity and depth appear to be quite uniform (Figure S2), suggesting the general absence of large-scale deletions and/or recombination events between homeologous chromosomes.

One category of particular interest is sites that exhibit 'fixed heterozygosity' across all of our *C. bursa-pastoris* samples, as these represent fixed differences between the homeologs that arose either because of initial sequence differences between the parental species or due to subsequent mutations that have spread through the tetraploid species. We find that 58% of the 105,082 fixed heterozygous positions are fixed SNP differences between *C. orientalis* and *C. grandiflora*. Of the remainder, 30% represent sites that are segregating within *C. grandiflora*, only 0.2% are segregating within the highly selfing *C. orientalis*, 0.06% in both species, while 12% of fixed heterozygous sites are not found segregating at all in our sample of *C. grandiflora* and *C. orientalis*. Conversely, of the 77,524 fixed differences identified between *C. orientalis* and *C. grandiflora*, 95.4% of these sites have both alleles in *C. bursa-pastoris* and 83% of these show fixed heterozygosity. Finally, we examined patterns of transposable element insertion polymorphism sharing among the three species. As expected, principal components analysis places *C. bursa-pastoris* intermediate between the two diploid species (Figure 2A), and 72% (2415) of insertions shared between two of the three species are shared between *C. bursa-pastoris* and *C. grandiflora*, 20% (675) are shared between *C. bursa-pastoris* and *C. orientalis*, while only 7% are shared uniquely between *C. grandiflora* and *C. orientalis*. These patterns are highly consistent with our earlier conclusions about hybrid origins, with the vast majority of fixed differences between homeologs having been sampled either from the differences between *C. orientalis* and *C. grandiflora* or segregating variation from within the highly polymorphic outcrossing species.

To estimate the timing of origin of *C. bursa-pastoris*, SNPs from each individual were phased and identified as segregating on the *C. grandiflora* or *C. orientalis* descended homeolog

(see methods and SI text). Principal components analysis of these phased SNPs shows strong clustering of the A subgenome with *C. grandiflora* and *C. rubella*, while the B subgenome clusters with *C. orientalis* (Figures 2B and 2C). We used coalescent simulations applying the composite likelihood method implemented in fastsimcoal2.1 (45) to fit models of speciation to the joint site frequency spectra data from the *C. bursa-pastoris* A subgenome, *C. grandiflora*, *C. bursa-pastoris* B subgenome and *C. orientalis*. We compared four different models using Akaike's Information Criterion (AIC), including models with either a stepwise or exponential change in population size, and with or without post-polyploidization gene flow. The models show general agreement in terms of timing of origin, but the best-fit model is one without migration but with exponential growth following speciation (Figure 3, see SI text). These analyses suggest that *C. bursa-pastoris* formed very recently, between 100,000 and 300,000 years ago, while the divergence time between the parental lineages of *C. orientalis* and *C. grandiflora* was considerably older, at approximately 900,000 years ago. We also estimate that the current effective population size of *C. bursa-pastoris* (approx. 37,000-75,000) is considerably larger than that of the highly selfing, nearly invariant diploid *C. orientalis* (4,000), but smaller than the effective population size of the highly outcrossing *C. grandiflora* (840,000).

Identification and characterization of putatively deleterious mutations

We took several approaches to determine whether duplicate genes are already deleted and/or degenerating within the tetraploid *Capsella bursa-pastoris*. First, to examine evidence for gene loss, we looked for genes with significant reductions in normalized coverage in our 10 *C. bursa-pastoris* genomes compared with the diploid parental species, *C. grandiflora* and *C.*

rubella. 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. Taking this approach, we find only 76 genes with statistical evidence of gene loss specific to *C. bursa-pastoris*, suggesting that whole gene loss events have not yet occurred on a large scale over approx. 200,000 years of divergence (Table S1). Of these putative deletions, many of them 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 spanned repetitive sequence where breakpoints could not be resolved. The identified deletions averaged 5015 bp and spanned 1067 to 19,643 bp in length, consistent with conclusions from ancient polyploids that most gene loss events are mediated by short deletion events (23). While some of these deletions are in annotated genes that often show presence/absence polymorphism even in diploids, such as F-box proteins and disease resistance genes (46), others appear to be in essential genes (Table S1).

Since the coverage analysis requires significant reductions in read number across our entire set of *C. bursa-pastoris* individuals from a broad geographic range including Europe and Asia, it is likely to be highly enriched for deletions that are at high frequency or fixed across the species. To examine the potential for sample-specific deletions, structural variants were called using Pindel (47), and we identified deletions spanning 80% or more of a gene. In *C. bursa-pastoris*, 150 gene deletions unique to the species were identified, while 155 and 26 were found in *C. grandiflora* and *C. orientalis* respectively. The relative number of gene deletions compared to all other polymorphic deletions was moderately higher in *C. bursa-pastoris* compared to both

C. grandiflora (Chi-squared p= 0.026) and *C. orientalis* (p= 0.039) and similar between the two diploids (p= 0.505).

Given the apparently low levels of whole gene loss in *C. bursa-pastoris*, we were interested in examining whether more subtle signs of gene degeneration were apparent. We therefore assessed the numbers and frequencies of putatively deleterious SNPs and insertion/deletion events (indels) causing frameshift mutations in coding regions. To eliminate major-effect mutations that may have occurred in our *C. rubella* reference genome, we used *Neslia paniculata* to polarize these changes and retained only those that represent derived SNPs in our focal species. Deleterious SNPs were defined as having likely loss-of-function effects on annotated genes. Putative deleterious SNPs and indels that showed evidence for compensatory mutations, including correcting frameshift mutations and SNPs restoring gene function (48) were also accounted for and removed from this analysis.

Even after these filters, we identified a large number of putatively deleterious SNPs in the *C. bursa-pastoris* genome (Table 1). Interestingly, a large proportion of these are fixed between homeologs, implying that one of the duplicate copies may be inactive in all individuals. Nevertheless, we observe an excess of rare variants relative to fourfold degenerate sites among the putatively inactivating mutations still segregating in *C. bursa-pastoris*, suggesting the continued action of weak purifying selection on at least some of these mutations (Figure S3). To understand the origins of the putatively deleterious mutations, we examined the frequency and abundance of putatively inactivating mutations occurring in the diploid *Capsella* genomes (Table 1; Figure S4). Strikingly, the majority of fixed deleterious SNPs between the homeologs in *C. bursa-pastoris* were also found in the diploid species; in particular, depending on the effect type,

40-60% of these were found in *C. orientalis*, 5-17% are in *C. grandiflora*, and 10-20% were found in both species. In contrast, 87% of frameshift indels are unique to *C. bursa-pastoris*. However, this could be related to the difficulty in accurately calling indel polymorphisms, since this trend is present in both diploids and for indels segregating in non-coding regions.

Nearly all (98%) of the putatively deleterious SNPs from *C. orientalis*, are fixed in *C. orientalis*, reflecting its very low levels of polymorphism (Figure S4). Since *C. orientalis* is highly selfing and has a low effective population size, this may reflect the fixation of slightly deleterious mutations that have been passed on to *C. bursa-pastoris*. In strong contrast, many of the putatively deleterious SNPs from *C. grandiflora* are found at low frequencies in this highly polymorphic outcrosser, and less than 1% are fixed in *C. grandiflora* (Figure S4). Furthermore, there is evidence that the frequencies of these mutations have increased in *C. bursa-pastoris* compared to *C. grandiflora*. Thus, although there is an apparent bias towards *C. orientalis* being a source of deleterious SNPs, it is likely that additional 'unique' deleterious SNPs in *C. bursa-pastoris* originated from rare SNPs from the *C. grandiflora*-like ancestor. In either case, it appears that the beginnings of gene inactivation in *C. bursa-pastoris* may have had a 'head start' from both fixed slightly deleterious inactivations in *C. orientalis* and low-frequency, possibly recessive deleterious variation from the outcrossing *C. grandiflora*-like progenitor.

Biased Gene Loss

To investigate bias in the types of genes being inactivated in all three *Capsella* species, we examined the functional categories of *Arabidopsis* orthologs using the Virtual Plant online server (49). Consistent with analyses of convergent ancient gene loss events (10), there was an

enrichment of *C. bursa-pastoris* SNPs affecting stop codon gain in genes associated with several functions associated with DNA, including *DNA metabolic process, DNA repair* and *DNA recombination* (Table S2). SNPs affecting splice site loss were also enriched in genes associated with *DNA metabolic process*. Interestingly, segregating *C. grandiflora* stop codon gained SNPs are enriched for similar functional categories related to DNA repair and metabolism (Table S2). These shared enrichments of functional categories are for mutations unique to each species as well as those shared between *Capsella* species, suggesting the similarities are not trivially due to shared polymorphisms. In contrast, putative loss-of-function mutations in *C. orientalis* were not enriched for any GO category. Thus, genes for which inactivating mutations segregate in the outbred diploid ancestor also tend to show independent, derived mutations causing gene inactivation in the recently formed tetraploid, but not in the highly homozygous selfing *C. orientalis*. Taken together, this suggests that mutations in these genes are tolerated both in the outcrossing diploid and the allotetraploid, perhaps due to their predominantly recessive effects. This is consistent with the hypothesis that dosage-insensitive genes are more likely to experience gene loss following whole genome duplication (50).

To test whether positive selection is driving the fixation of loss-of-function point mutations we scanned for signs of "selective sweeps" or dips in neutral diversity surrounding fixed inactivating mutations. Average nucleotide diversity at 4-fold synonymous sites was assessed in 50kb windows of coding regions up- and down-stream of fixed putatively deleterious point mutations, separately for homeologous chromosomes. These windows show a large amount of variation, but no clear trend of reduced neutral diversity near focal mutations in comparison

with synonymous substitutions (Figure S5). Thus, there is no clear evidence that inactivating mutations were fixed by positive selection.

Quantifying relaxation of purifying selection

Overall, we detect small numbers of gene loss events, and signs that many major effect mutations were inherited from parental species. This suggests that the earliest stages of polyploid gene degeneration may be subtle, and reflect a global shift in selection pressures genome-wide rather than rampant gene loss. To investigate the strength of selection acting on *C. bursa-pastoris* compared to its close diploid relatives, we estimated the distribution of fitness effects of deleterious mutations for all three *Capsella* species using allele frequency spectra and separating the two homeologous genomes of *C. bursa-pastoris* (51). Both *C. bursa-pastoris* homeologs and *C. orientalis* have significantly higher proportions of effectively neutral sites ($N_e s < 1$, where N_e is the effective population size and s is the strength of selection against deleterious mutations) than *C. grandiflora*, for both 0-fold nonsynonymous and conserved noncoding sites ($p < 0.005$; Figure 4). This suggests that the combination of high rates of selfing and reduced effective population size may be causing a reduction in the efficacy of selection unrelated to ploidy. In fact, the proportion of *C. orientalis* sites in the effectively neutral category is significantly greater than on the *C. orientalis*-descended *C. bursa-pastoris* homeolog for 0-fold sites ($p < 0.01$), but is not significantly greater for conserved noncoding sequences ($p > 0.05$). In contrast, both *C. bursa-pastoris* homeologs show a significant reduction in the proportion of strongly selected nonsynonymous sites compared to both diploid progenitors ($N_e s > 100$) ($p < 0.05$) although this is not the case for conserved noncoding sequence, and the difference between *C. orientalis* and *C.*

bursa-pastoris is small (Figure 4). Between *C. bursa-pastoris* homeologs there is no significant difference in the distribution of fitness effects for any site type, providing no evidence for early signs of 'biased fractionation' in terms of the efficacy of selection on the two homeologs. Overall, we observe strong evidence for a relaxation of purifying selection in *C. bursa-pastoris* relative to *C. grandiflora*, while the estimated strength of purifying selection is comparable to the highly selfing *C. orientalis*, which has experienced at least a 10-fold lower effective population size relative to the tetraploid species (Figure 3). Thus, genome-wide relaxation of selection in *C. bursa-pastoris* likely reflects the combined effects of selfing, effective population size and a shift in average selection coefficients due to tetraploidy.

Conclusions

We have investigated the evolutionary origin and selective consequences of polyploidization in the widespread tetraploid weed *Capsella bursa-pastoris* using whole genome sequencing. Our results provide comprehensive evidence for recent allopolyploid origins of *C. bursa-pastoris* from two parental lineages ancestral to present-day *C. orientalis* and *C. grandiflora*. By contrasting patterns of functional polymorphism and divergence genome-wide, we find evidence for significant relaxation of purifying selection, but no evidence for massive gene loss rapidly upon polyploidization. Many large-effect SNPs in *C. bursa-pastoris* are shared with *C. orientalis* and *C. grandiflora*, suggesting that gene degradation following allopolyploidization can get a head start from standing variation in progenitors. Taken together, and in contrast with patterns observed in other systems (30) our results suggest that the early stages of allopolyploid evolution

in *C. bursa-pastoris* were characterized by relaxed purifying selection rather than large-scale "genomic shock" and rapid gene loss.

Materials and Methods

Seeds from ten *Capsella bursa-pastoris* plants were collected from one individual per population across the species range in Eurasia, and seeds from ten *C. orientalis* individuals (35) were collected from five populations in Russia, China and Mongolia (Table S3). Seeds were sterilized, plated on half-strength Murashige-Skoog nutrient medium, vernalized at 4°C for three weeks, and germinated at room temperature. Seedlings were planted in a standard potting mix in 1L round pots and grown at the University of Toronto glasshouse. Ploidy was confirmed using flow cytometry with an internal standard (Plant Cytometry Services, Schijndel, the Netherlands).

Nuclear genomic DNA was extracted from leaf material for the 10 *C. bursa-pastoris* and 10 *C. orientalis* individuals using a modified CTAB protocol (52). Whole genome sequencing of each individual was completed at the Genome Quebec Innovation Centre using the Illumina GAII or Illumina Hiseq platforms (Illumina, San Diego, California, USA). Paired-end 100 or 108 base pair reads were generated for *C. bursa-pastoris* and *C. orientalis*, with median depth coverage of 20 and 12 reads per site, respectively. *C. orientalis* individuals were sequenced to lower coverage since they are largely homozygous and so SNP calls should be less ambiguous. Sanger sequencing of a more extensive sample of *C. bursa-pastoris* was used in combination with published sequences from other *Capsella* species (Table S4) to validate the conclusions about hybrid origins.

De novo fragment assemblies for *C. bursa-pastoris* were generated using Ray v 1.4 (53). Read mapping of Illumina genomic reads to the *C. rubella* reference genome (37) was conducted using Stampy version 13 (54), and phasing of read-mapped samples was conducted using Hapcut (55). SNP calling and polymorphism analyses were conducted using the Genome Analysis Toolkit (56). We constructed phylogenies using RAxML's (57) rapid bootstrap algorithm to find the best-scoring ML tree. For demographic inferences, we used fastsimcoal2.1 (45) to infer demographic parameters based on the multidimensional site frequency spectrum for *C. grandiflora*, *C. orientalis*, and the two *C. bursa-pastoris* homeologous genomes. To evaluate the fit of the best demographic model (Table S5), simulated datasets were compared with the observed site frequency spectra (Figure S6). The Distribution of Fitness Effects (DFE), was estimated for each species using the maximum likelihood approach designed by Keightley and Eyre-Walker (51). SnpEff (58) was used to predict the effects of SNPs called using GATK. See SI Text for detailed methods on *de novo* assembly, population genomics analyses and validation.

ACKNOWLEDGEMENTS

This project was funded by an NSERC Discovery Grant and a Genome Quebec/Genome Canada grant to S.I.W, and grants from the Swedish Research Council to T.S and to M.L.. G.D. was supported by an NSERC scholarship, and EBJ by an NSF graduate fellowship. Fastsimcoal2.1 computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project b2012190.

Literature Cited

1. Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet* 14:262–274.

2. Hough J, Williamson RJ, Wright SI (2013) Patterns of Selection in Plant Genomes. *Annual Review of Ecology Evolution and Systematics* 44(1):31-49.

3. Slotte T (2014) The impact of linked selection on plant genomic variation. *Brief Funct Genomics*. *doi:10.1093/bfgp/elu009*

4. Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Annu Rev Genet* 34:401–437.

5. Glémin S, Ronfort J (2013) Adaptation and maladaptation in selfing and outcrossing species: new mutations versus standing variation. *Evolution* 67:225–240.

6. Glémin S (2007) Mating systems and the efficacy of selection at the molecular level. *Genetics* 177:905–916.

7. Ohno S (1970) *Evolution by gene duplication*. Springer-Verlag, New York.

8. Hegarty MJ, Hiscock SJ (2008) Genomic clues to the evolutionary success of polyploid plants. *Curr Biol* 18:R435–44.

9. Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. *Proc Natl Acad Sci USA* 97:7051–7057.

10. 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* 110:2898–2903.

11. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.

12. Sankoff D, Zheng C (2012) Fractionation, rearrangement and subgenome dominance. *Bioinformatics* 28:i402–i408.

13. Vision TJ, Brown DG, Tanksley SD (2000) The Origins of Genomic Duplications in Arabidopsis. *Science*. 290: 2114-2117.

14. Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* 101:9903–9908.

15. Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–

438.

16. Barker MS et al. (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* 25:2445–2455.

17. Jiao Y et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:97–100.

18. D'Hont A et al. (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* 488:213–217.

19. Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci USA* 108:4069–4074.

20. Lai J et al. (2004) Gene loss and movement in the maize genome. *Genome Research* 14:1924–1931.

21. Schnable PS et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115.

22. Langham RJ et al. (2004) Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166:935–945.

23. Woodhouse MR et al. (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS Biol* 8:e1000409.

24. Woodhouse MR et al. (2014) Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc Natl Acad Sci USA* 111:5283–5288.

25. Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van de Peer Y (2002) The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci USA* 99:13627–13632.

26. Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16:1667–1678.

27. Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K (2011) Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci* 16:108–116.

28. Wood TE et al. (2009) The frequency of polyploid speciation in vascular plants. *Proc Natl Acad Sci USA* 106:13875–13879.

29. Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV (2010) Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and

network partners. *Genome Biol* 11: R125.

30. Buggs RJA et al. (2012) Rapid, Repeated, and Clustered Loss of Duplicate Genes in Allopolyploid Plant Populations of Independent Origin. *Curr Biol* 22:248–252.

31. Hurka H, Friesen N, German DA, Franzke A, Neuffer B (2012) "Missing link" species Capsella orientalis and Capsella thracica elucidate evolution of model plant genus Capsella (Brassicaceae). *Mol Ecol* 21:1223–1238.

32. 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* 29:1721–1733.

33. Hurka H, Neuffer B (1997) Evolutionary processes in the genus Capsella (Brassicaceae). *Plant Systematics and Evolution* 206:295–316.

34. Foxe JP et al. (2009) Recent speciation associated with the evolution of selfing in Capsella. *Proc Natl Acad Sci USA* 106:5241–5245.

35. 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* 106:5246–5251.

36. Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G (2013) Genomic Identification of Founding Haplotypes Reveals the History of the Selfing Species Capsella rubella. *PLoS Genet* 9. e1003754

37. Slotte T et al. (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* 45:831–835.

38. Paetsch M, Mayland-Quellhorst S, Hurka H, Neuffer B (2010) in *Evolution in Action* (Springer Berlin Heidelberg, Berlin, Heidelberg), pp 77–100.

39. Paetsch M, Mayland-Quellhorst S, Neuffer B (2006) Evolution of the self-incompatibility system in the Brassicaceae: identification of S-locus receptor kinase (SRK) in self-incompatible Capsella grandiflora. *Heredity (Edinb)* 97:283–290.

40. 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.

41. Haudry A et al. (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* 45:891–898.

42. Hurka H, Freundner S, Brown A, Plantholt U (1989) Aspartate-Aminotransferase Isozymes in the Genus Capsella (Brassicaceae) - Subcellular Location, Gene Duplication,

and Polymorphism. *Biochem Genet* 27:77–90.

43. Williamson RJ et al. (2014) *Evidence for widespread positive and negative selection in coding and conserved noncoding regions of Capsella grandiflora*. in revision, preprint available on BioArXiv: http://dx.doi.org/10.1101/002428

44. Mummenhoff K, Hurka H (1990) Evolution of the Tetraploid Capsella-Bursa-Pastoris (Brassicaceae) - Isoelectric-Focusing Analysis of Rubisco. *Plant Systematics and Evolution* 172:205–213.

45. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC (2013) PLOS Genetics: Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet*. 9:e1003905.

46. Clark RM et al. (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* 317:338–342.

47. 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* 25:2865–2871.

48. Long Q et al. (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet* 45:884–890.

49. Katari MS et al. (2010) VirtualPlant: A Software Platform to Support Systems Biology Research. *Plant Physiol* 152:500–515.

50. Conant GC, Birchler JA, Pires JC (2014) Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol* 19C:91–98.

51. 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* 177:2251–2261.

52. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* 19:11–15.

53. Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol* 17:1519–1533.

54. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21:936–939.

55. Bansal V, Bafna V (2008) HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* 24:i153–9.

56. DePristo MA et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.

57. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

58. 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)* 6:80–92.

Figure Legends

Figure 1. Results of RAxML tree analysis from whole genome alignments of the four *Capsella* species, using *Neslia paniculata* as an outgroup. Numbers show the number of trees with greater than 80% bootstrap support showing each possible topology, from genome-wide alignments.

Figure 2. Principal components 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 subgenome of origin. The first two axes are shown. B and C, analysis of SNP data from phased subgenomes of *C. bursa-pastoris*, with the first two axes (B) and the first and third axis shown.

Figure 3. Demographic parameter estimates with 95% confidence intervals for four models of allopolyploid speciation in *Capsella*. Four models were investigated: A) Stepwise population size change, no gene flow, B) Stepwise population size change, asymmetric gene flow, C) Exponential population size change, no gene flow and D) Exponential population size change, asymmetric gene flow. Model C was preferred based on AIC and Akaike's weight of evidence (w). Estimates of effective population sizes for *C. grandiflora* (*Cg*), *C. orientalis* (*Co*), *C. bursa-*

pastoris (*Cbp* A and *Cbp* B) are given in thousands of individuals, and estimates of the timing of the origin of *C. bursa-pastoris* (*T1(Cbp)*) and the split between *C. grandiflora* and *C. orientalis* (*T1(Cg-Co)*) are given in kya. Note that for models with exponential population size change, effective population sizes correspond to current effective population sizes. Confidence intervals are given in parentheses.

Figure 4. Estimates of the distribution of fitness effects (DFE) of deleterious mutations at 0-fold non-synonymous sites (A) and conserved non-coding sites (B), based upon the site frequency spectra of SNPs at these sites. The strength of selection is measured as $N_e s$, where N_e is the effective population size and s is the strength of selection. Error bars correspond to 95% CI of 200 bootstrap replicates of 10 kb blocks.

SI Figure Legends

Figure S1. Neighbor-joining trees for 9 loci amplified using homeolog-specific PCR and sequenced in 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 (Ch: China vs WEu: Western Eurasia) and homeolog designation (A or B), where applicable. As outgroups (not shown) we used *N. paniculata* or *A. thaliana* sequences. Only bootstrap values over 50% are shown.

Figure S2. Sliding windows of heterozygosity and coverage in *C. bursa-pastoris*. Values shown are the number of heterozygous relative to homozygous genotype calls in 5000 SNP windows,

and the average number of reads covering a genomic position. 5000 SNP windows correspond to approximately 170,000bp on average.

Figure S3**:** *C. bursa-pastoris* site frequency spectra of deleterious SNPs (green) and 4-fold synonymous SNPs (grey) segregating A) uniquely in *C. bursa-pastoris,* B) in both *C. bursa-pastoris* and *C. grandiflora,* C) in both *C. bursa-pastoris* and *C. orientalis,* D) segregating in all three species. The deleterious SNP categories included are stop codon gained, stop codon lost, start codon lost and splice site lost.

Figure S4: Joint site frequency spectra of shared variants for *C. bursa-pastoris* and *C. grandiflora* (A and C), and *C. bursa-pastoris* and *C. orientalis* (B and D). Putatively deleterious SNPs are shown in panels A and B, while variation at 4-fold synonymous sites is shown in panels C and D.

Figure S5. Neutral 4-fold synonymous diversity surrounding fixed putatively deleterious mutations in the two homeologs of *Capsella bursa-pastoris*: A) *C. bursa-pastoris* homeolog *A* (descended from *C. grandiflora*-like ancestor) B) *C. bursa-pastoris* homeolog *B* (descended from C. orientalis). Diversity is normalized by divergence with *Neslia paniculata.* Shaded regions correspond to bootstrapped (n=1000) 95% CI of neutral diversity surrounding fixed 4-fold synonymous mutations.

Figure 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 maximum likelihood parameters inferred by fastsimcoal2.1 for this model. C. Model residuals across joint SFS. D. Histograms of model residuals.

Figure 1

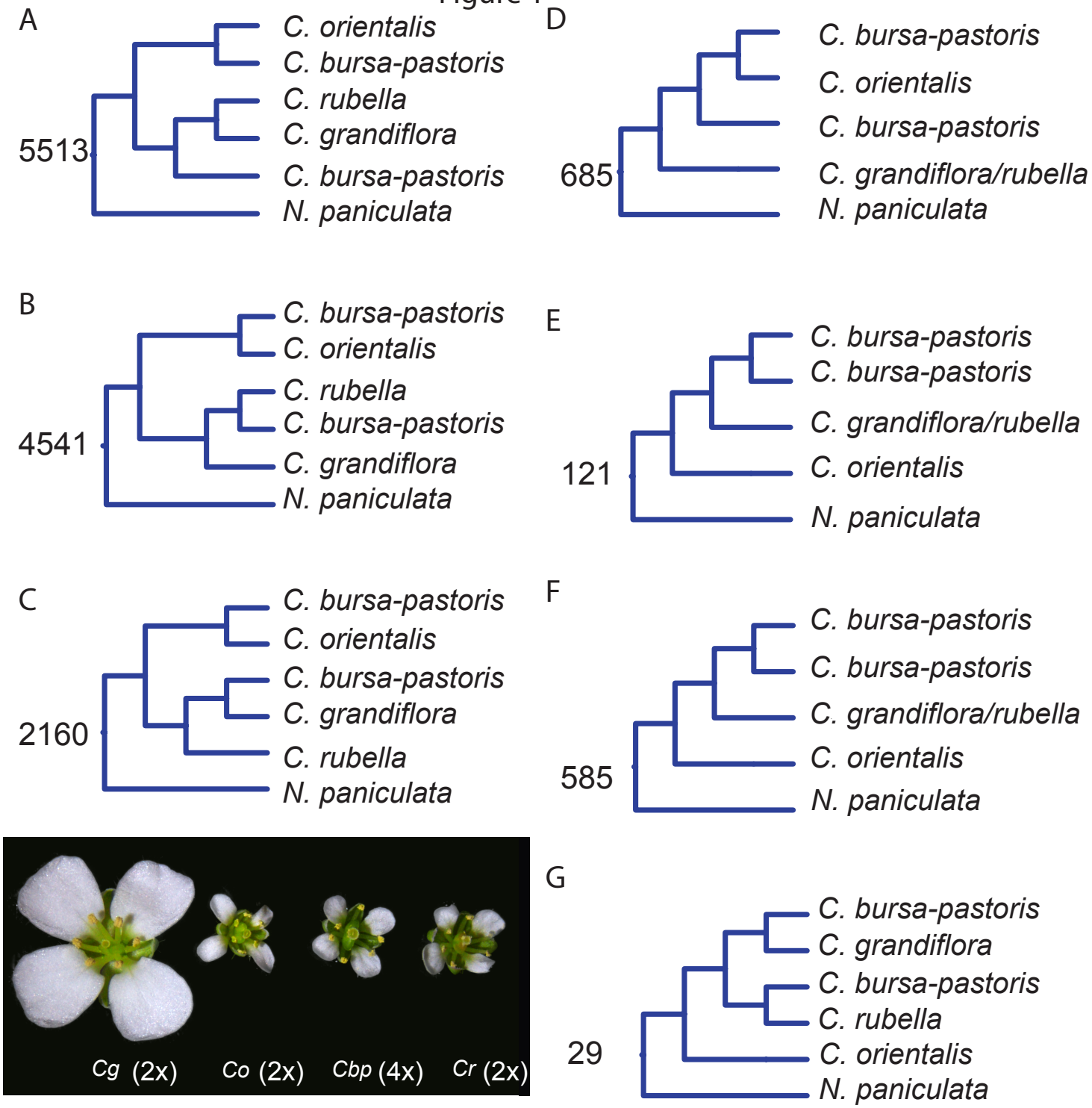

Figure 2

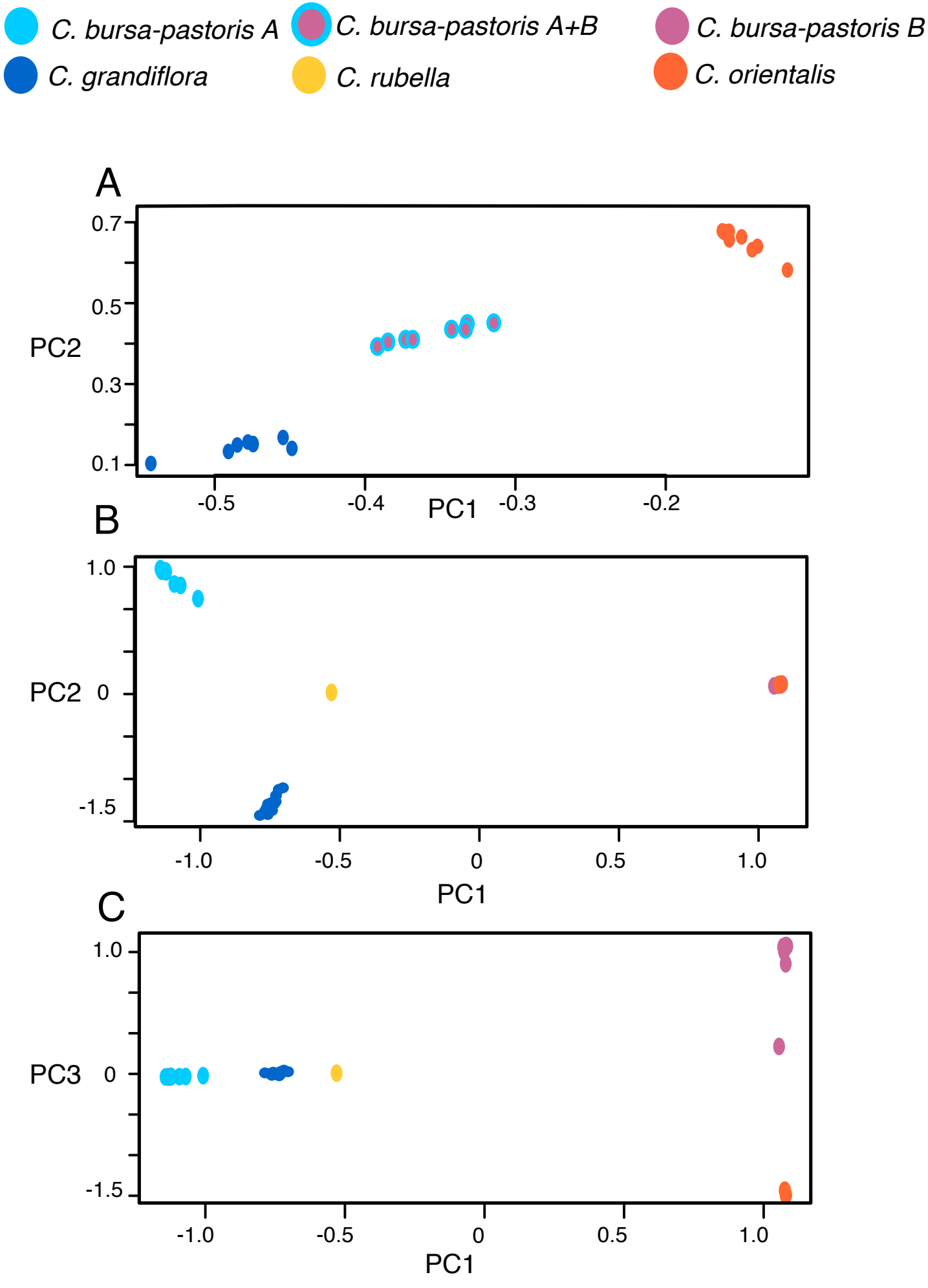

Figure 3

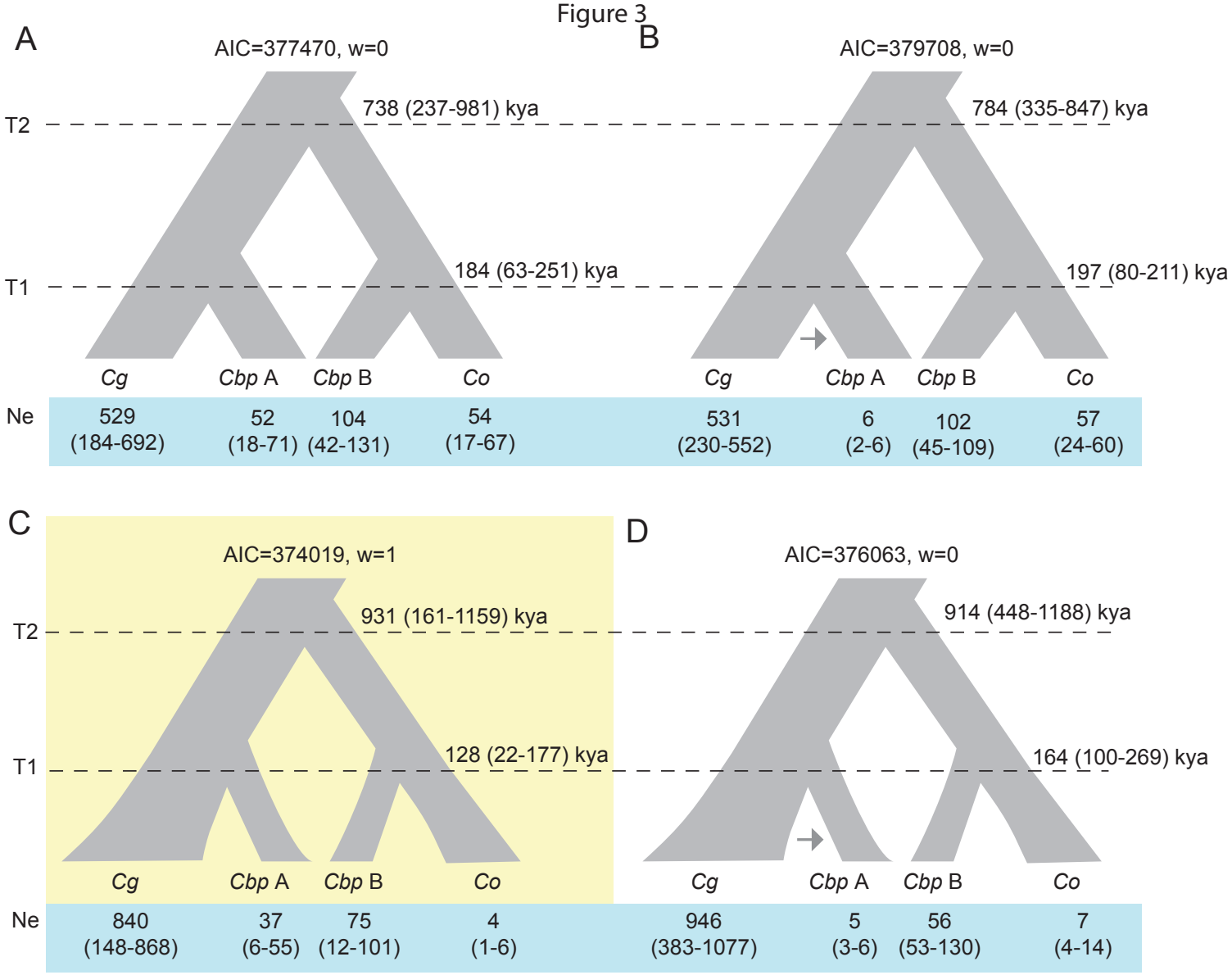

Figure 4

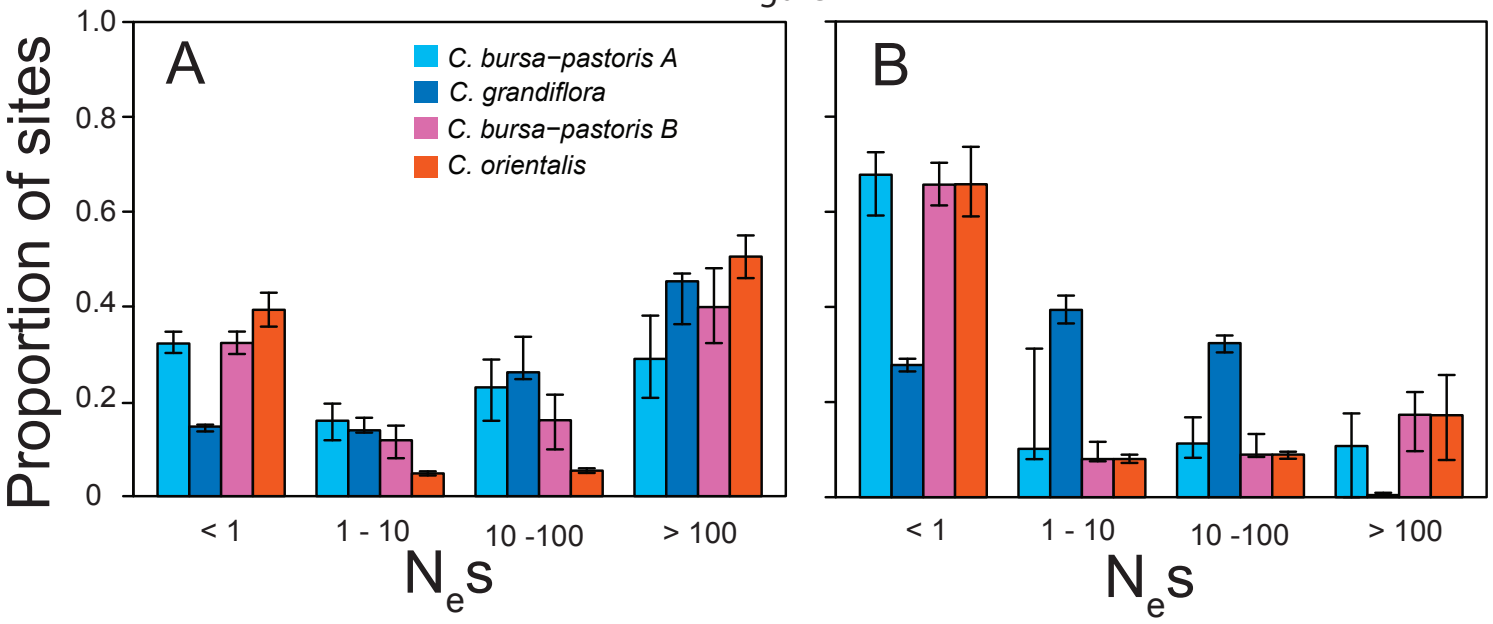

Table 1: Counts of the different categories of SNPs segregating in *C. bursa-pastoris*[1]

SNP effect	Unique to *C. bursa-pastoris*	Shared with *C. grandiflora*	Shared with *C. orientalis*	Shared with both
Four-fold synonymous	10767 (1071)	9203 (6875)	9325 (5027)	4106 (2755)
Start codon lost	347 (22)	75 (11)	199 (71)	125 (21)
Stop codon gained	2829 (105)	650 (55)	1093 (348)	485 (122)
Stop codon lost	312 (18)	138 (25)	252 (105)	220 (37)
Splice site lost	1652 (74)	394 (56)	743 (271)	454 (77)

1- values in parentheses indicate fixed heterozygous SNPs likely reflecting fixed differences between homeologues