Genome evolution and adaptation of a successful allopolyploid, *Capsella bursa-pastoris*

DMYTRO KRYVOKHYZHA
The term allopolyploid refers to an organism that originated through hybridization and increased its ploidy level by retaining the unreduced genomes of its parents. Both hybridization and polyploidy usually have negative consequences for the organism. However, there are species that not only survive these modifications but even thrive and can outcompete their diploid relatives. There are many intuitive explanations for the success of polyploids, but the number of empirical studies is limited.

The shepherd's purse (Capsella bursa-pastoris) is an emerging model for studying a successful allopolyploid species. Capsella bursa-pastoris occurs worldwide, whereas its parental species, Capsella grandiflora and Capsella orientalis, have more limited distribution range. C. grandiflora is confined to Northern Greece and Albania, and C. orientalis is found only in the steppes of Central Asia. We described the genetic variation within C. bursa-pastoris and showed that it is not homogeneous across Eurasia but rather subdivided into three genetically distinct populations: one comprises accessions from Europe and Eastern Siberia, the second one is located in Eastern Asia and the third one groups accessions around the Middle East. Reconstruction of the colonization history suggested that this species originated in the Middle East and subsequently spread to Europe and Eastern Asia. This colonization was probably human-mediated. Interestingly, these three populations survive in different environmental conditions, and yet most gene expression differences between them could be explained by neutral processes. We also found that despite a common history within one species, the two subgenomes retained differences already present between the parental species. In particular, the genetic load was still higher on the subgenome inherited from C. orientalis than on the one inherited from C. grandiflora. The two subgenomes were also differentially influenced by introgression and selection in the three genetic clusters. Gene expression variation was highly correlated between the two subgenomes but the total level of expression showed variation in parental dominance across flower, leaf, and root tissues.

This thesis for the first time shows that the evolutionary pathways of allopolyploids may differ not only on the species level but also between populations within one species. It also supports the theory that alloploidy provides an increased amount of genetic material that enables evolutionary flexibility.

Keywords: allopolyploidy, population structure, selection, genetic drift, gene expression, parental legacy, genetic load
"If you thought that science was certain - well, that is just an error on your part."

– Richard P. Feynman
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


IV **Kryvokhyzha D.**, Duan T., Orsucci M., Milesi P., Glémin, S., Wright, S.I., and Lascoux, M., 2018. Towards the new normal: Genomic and transcriptomic changes in the two subgenomes of a 100,000 years old tetraploid, *Capsella bursa-pastoris*. (Manuscript)

* These authors contributed equally to the work.
Reprints were made with permission from the publishers.
# Contents

1 Introduction ................................................................................................................. 9  
   1.1 Evolutionary significance of polyploidy and hybridization .......................... 10  
   1.2 Population structure, genetic drift and adaptation ....................................... 14  
   1.3 Challenges of studying alloployploid organisms ........................................... 18  
   1.4 A promising model system to understand alloployploidy:  
       *C. bursa-pastoris* ............................................................................................ 19  

2 Present study ............................................................................................................... 24  

3 Conclusions and future perspectives ....................................................................... 35  

Svensk sammanfattning .......................................................................................... 37  

Резюме українською ............................................................................................... 39  

Acknowledgements ................................................................................................. 41  

References ................................................................................................................. 43
1. Introduction

If you ate oatmeal or a sandwich for breakfast today, you ate a polyploid. Oatmeal is made from oat *Avena sativa*, which is a polyploid. Bread flour is usually produced by graining seeds of either common wheat *Triticum aestivum* or durum wheat *Triticum turgidum*, both of which are also polyploids. Moreover, these three species are also hybrids. Polyploids and hybrids are quite common among domesticated plants, and with the recent advent of molecular tools, it has become evident that they are also abundant in the wild.

Polyploid is an organism that has more than two paired sets of chromosomes. The majority of macroorganisms have a diploid set of chromosomes (2n), but some animals and many plants have tri- (3n), tetra- (4n), hexa- (6n) and even octoploid (8n) sets of chromosomes. Such increase in the ploidy level may arise from a doubling of the same genome within a cell, resulting in an autopolyploid, or from merging two genomes from different species during hybridization, resulting in an allopolyploid (Fig. 1.1). Both autopolyploidy and allopolyploidy imply significant genetic changes that disrupt cellular functions and usually lead to extinction. However, hybrids and polyploids that manage to survive often become successful species and can even outperform their diploid relatives.

Studying the phenomena associated to polyploidization may shed more light on our understanding of the link between the content of a genome and fitness. The genetic variability of polyploids has often been used in the domestication process. More than half of all crop species are polyploids. The frequency of polyploids in the wild, while significantly lower, remains high (40%) [1]. There has been a lot of progress in studying the genomic consequences of polyploidy in crops [2, 3]. However, polyploidy was rather a precondition of domestication than a consequence [1]. So, we also need to study the origin and mechanisms of evolution in natural polyploid species.

The genome-wide analysis of allopolyploid species is, however, not easy. The significant size of polyploid genomes, the aggregate of two or more divergent genomes and the presence of multiple alleles and loci, as well as massive structural and functional rearrangements, make it challenging to assemble the genome, obtain accurate genotype data and analyze it with existing population genetics models that are primarily developed for diploid data. This is the main reason why genomic studies of polyploids are still lagging behind those of diploids.

This thesis is a contribution to the progress of genomic studies of allopolyploids. Here, I summarize our investigations of genome evolution and adaptation in one of the most successful natural polyploids, the shepherd’s purse
Figure 1.1. Scheme of origin of a homoploid hybrid, allopolyploid and autopolyploid species. Parental species contain two chromosomes (2n). Diploid hybrid (homoploid) species arises through merging of one chromosome from each parental species. Allopolyploid species originates the same way except that it combines the entire chromosomal sets of both parental species. Autopolyploid species emerges by chromosome doubling of the same species.

Capsella bursa-pastoris. The results show that evolution of allopolyploids is more complex than previously imagined. In particular, our results indicate that both demography and introgression played an important role in the evolution of the shepherd’s purse.

1.1 Evolutionary significance of polyploidy and hybridization

The evolutionary role of polyploidy has long been a subject of debate. In the 20th century, G. L. Stebbins stated that polyploidy contributed little to progressive evolution [4]. W. H. Wagner also denied the significance of polyploidy and regarded it as ”evolutionary noise” [5]. On the other hand, R. J. Schultz insisted that polyploidy was far from playing a secondary role in evolution and had actually contributed to major steps in evolution [6]. Despite important progress in our understanding of polyploids since these early studies [7], it is still debated whether polyploidy is an evolutionary dead-end or a major driver of diversification [8, 9, 10, 11].

The evolutionary significance of hybridization has also been controversial. Charles Darwin himself changed his point of view from discounting hybridiza-
tion as a source of new forms ("I am very far from believing in hybrids...", Darwin’s letter to Joseph Hooker in 1856) to emphasizing its significance ("the several kinds of dogs are almost certainly descended from more than one species, and so it is with cattle, pigs and some other domesticated animals.", Variation of Animals and Plants under Domestication, 1890). The progressive role of hybridization was totally discarded during the era of the biological species concept that viewed species as discrete reproductively isolated units [12]. During the last two decades, the view of hybridization as a source of diversification and adaptation has been revived [13]. Moreover, with the advent of DNA sequencing and genome-scale analyses, it has become evident that hybridization resulting in genomic introgression and speciation is widespread [14, 15].

The evolutionary significance of polyploidy is usually attributed to its ability to rapidly produce new vigorous forms. An increase of ploidy results in nearly instantaneous reproductive isolation from parental species [16]. For example, if a newly emerged tetraploid (4n) individual crossed with a diploid (2n) ancestor, it would produce a triploid (3n) offspring. Having an odd number of chromosomes, such triploid progeny would be sterile due to problems with chromosome pairing during meiosis. Reproductive isolation may also arise from switching to the asexual reproduction mode, e.g. reproduction from maternal tissues bypassing sexual fusion of an egg and a sperm. Asexual reproduction in polyploids is often realized through parthenogenesis, a process when an embryo develops from an unfertilized egg cell. This is a common reproduction mode in polyploid animals [17]. Polyploids may also reproduce by selfing. The change of ploidy level often disrupts the self-incompatibility system and allows self-fertilization, which is especially common among plants [18, 19]. Thus, polyploidy is one of the few mechanisms that may lead to sympatric speciation (speciation without geographic isolation).

New polyploid lineages often demonstrate superiority relative to diploid species, a phenomenon known as polyploid vigor. This vigor is generally attributed to increased genetic variability. First, polyploid genomes are usually unstable at early stages and experience massive rearrangements and fragment loss [20]. These major genomic changes introduce more variation into a polyploid population that survives initial stages of formation and thus provides more material that can be selected for. Second, polyploids have several copies of each gene. This gene redundancy also benefits a polyploid because genes copies can gain new functions while retaining the old function. Third, polyploids have greater ability to mask deleterious mutations and suffer less from inbreeding depression. Because polyploids have several genomic copies, deleterious mutations that are usually recessive have smaller chance to influence a phenotype. For instance, take a diploid that has a genotype with two dominant alleles AA. If one of its two alleles mutates to the deleterious recessive allele a, the resulting genotype is Aa, and a allele needs to replace only one dominant allele A to produce a genotype aa and change the phenotype. Whereas one recessive allele would need to replace three dominant alleles in a tetraploid to
achieve the same phenotypic change because the genotype of a tetraploid is $AAAa$. Several genomic copies also ensure that crossing between closely related individuals, known as inbreeding, rarely leads to homozygous offspring and thus does not have a deleterious effect. Thus, deleterious mutations need to reach higher frequency in polyploids than in diploids to achieve a homozygous state and influence the phenotype. Fourth, polyploids often experience rapid changes in patterns of gene expression due to gene duplication. Duplicated genes may gain new functions (neo-functionalization), partition their original ancestral function (subfunctionalization) or completely lose their function (pseudogenization) [21, 22, 23]. Finally, these genetic changes also influence the proteome, metabolome and phenotypic variation. A few studies that have investigated the effect of polyploidy on proteome and metabolome show that the variation in proteome and metabolome may substantially differ between polyploids and parental species [24, 25]. Polyploids also tend to possess enlarged organs and overall larger size [26]. All these genetic and phenotypic modifications of polyploids provide more variability that may facilitate survival.

Hybridization also increases genetic variance and can produce new vigorous forms. The genetic variance is boosted by genomic re-patterning because of the conflict between two divergent genomes and release of transposable elements from the suppression mechanisms established within each parent. The genomic conflict between two divergent genomes also disrupts gene expression regulation networks and lead to non-additive gene expression patterns when expression levels are not non-intermediate between parental species. This new variation and mixture of different parental traits may result in transgressive phenotypes, whereby a hybrid exhibits extreme traits relative to its two parents. Such traits may facilitate colonization of new ecological niches that are not accessible to parental species. Hence, hybridization may result into hybrid speciation. The hybrid species is then isolated from its parental species by its novel ecological niche. The most famous examples of such a speciation scenario include *Helianthus* sunflowers, *Rhagoletis* flies, *Cottus* sculpins, *Lycaeides* butterflies [27, 28, 29, 30]. Isolation from parental species can also arise due to assortative mating when hybrids have different mating seasons or phenology from the parental species and mate only among themselves as in some hybrid *Heliconius* butterflies [31]. These are scenarios of homoploid hybrid speciation because the chromosome number of a hybrid does not increase (Fig. 1.1).

However, hybridization can also be accompanied by chromosomal doubling, i.e. allopolyploidy (Fig. 1.1). Allopolyploid forms can emerge from chromosome doubling in a formed diploid hybrid or from a fusion of two unreduced gametes during the hybridization event [32, 33]. There is also a way through a triploid bridge when an unreduced diploid gamete fuses with a haploid gamete from a diploid species to form a triploid, and later this triploid produces an unreduced triploid gamete that fuses with a haploid gamete of a
diploid. There are numerous extant allopolyploid species but in most cases it is still unclear how they originated. A few documented examples include *Primula kewensis* that originated from somatic chromosome doubling in a diploid hybrid [34], a tetraploid strain of silk moths (*Bombyx*) that was obtained through a triploid bridge [35], and fusion of diploid gametes was used in the production of synthetic polyploids of red clover and potato [36, 37]. Allopolyploids combine features of both hybrids and polyploids. They are more diverse than autopolyploids because they comprise two divergent genomes instead of one duplicated genome. They also suffer less from aneuploidy than homoploid hybrids and autopolyploids because each divergent subgenome is doubled and thus can find a pair during meiosis. This is why allopolyploids usually have even chromosome number, most frequently 4n, and exhibit disomic inheritance when divergent subgenomes do not recombine. Allopolyploids also benefit from having two divergent genomes by differentially expressing genes from the two subgenomes (homeologue-specific expression).

Hybridization may also contribute to adaptability and diversification through genomic introgression when one species just acquires a fragment of genetic material from another species. It is well known that bacteria thrive on introgression that often confers adaptive traits [38]. There is also growing evidence of such benefits among plants and animals. For example, transfer of herbivore resistance in sunflower *Helianthus* [39] and insecticide resistance in malaria mosquitoes [40] have been attributed to introgressive hybridization. Introgressive hybridization has also contributed to color pattern diversity of *Heliconius* butterflies [41] and adaptive radiations in Darwin’s finches [42]. There are even evidence for archaic adaptive introgression in humans [43].

All these factors that increase genetic diversity and were described above as advantageous can have negative effects as well. Assuming that the genome structure and regulatory networks were optimized under selection in the parental species, both increase of ploidy level and hybridization would disrupt this long-established design and result in fitness loss. Moreover, an increase in the copy number of chromosomes causes problems with mitosis and meiosis during cell divisions and changes cellular architecture due to an increased nucleus size [44]. Forms that switch to self-fertilization suffer from the negative effect of inbreeding depression. Hybrid forms experience conflicts between divergent genomes and may suffer from outbreeding depression. Maybe this is the reason why some allopolyploids evolve towards functional diploidization when subgenomes experience differential gene loss, deletion, and down-regulation of gene expression [45, 46, 47, 48]. On balance, hybridization and polyploidy are more often deleterious than advantageous but those individuals that survive frequently become evolutionary successful lineages.
1.2 Population structure, genetic drift and adaptation

Species are highly dynamic entities: their size, density, and geographical location fluctuate over time; they are also often highly fragmented over space. This seriously complicates the study of species and the reconstruction of their histories. Even if one is primarily interested in a trait that, at first glance, is not related to demographics, for instance, the relation between the two subgenomes of a polyploid, the details of the demographic history of the polyploid species will need to be considered. Demography will for instance influence the evolution of the trait through the local and overall amount of random genetic drift. Structured populations will also face different environmental conditions and be under different selection pressures. A textbook example of the impact of environmental changes and local adaptation is the Peppered moth (*Biston betularia*). It was found that dark moths were prevalent in the polluted environment around Birmingham, UK, because they were less conspicuous on dark tree trunks, while light moths were common in a clean environment such as Devon, UK, where tree trunks were covered with lichens and thus were light in color [49]. This difference is proven to be due to differential selection [50].

However, as hinted above, evolutionary differences between populations are not necessarily the result of selection. Differences may also arise simply due to neutral processes. Suppose there are two well-isolated populations and a new mutation arises in one of these populations. Its probability to disappear from the population is high, even if it is positively selected. This is nicely illustrated by J. Haldane’s [51] early result on the probability of fixation of a positively selected allele: the probability of fixation of a mutation with a selected advantage $s$ is $2s$. If $s = 0.01$, a large value, the probability of fixation will be a mere 2%! Of course, a neutral mutation generally has an even bleaker future. If by chance the mutation does survive and increases in frequency, a population with this new allele will become divergent from a population without this allele because the two populations do not exchange alleles. Over time there will be many mutations, some will go to fixation, primarily neutral or nearly neutral, and populations will diverge through random genetic drift [52].

The concept of random genetic drift stemmed from the work of J. Haldane, S. Wright, and R. Fisher but gained prominence through M. Kimura’s neutral theory of molecular evolution [53]. The neutral theory of molecular evolution states that most changes observed at the molecular level reflect a balance between mutations that are functionally neutral or nearly neutral and random genetic drift acting on these mutations in a population of finite size. The neutral theory of evolution does not state that adaptive selection does not occur but assumes that it plays a minor part in changes at the genome level. The neutral theory has been challenged by a selection theory of evolution which states that truly neutral mutations represent a minority of all changes and that selection is a major force of evolution [54]. Whether the majority of changes are neutral or adaptive is still hotly debated [54, 55, 56]. Independently of this debate,
The neutral theory laid a foundation for statistical frameworks to distinguish natural selection from random genetic drift at the molecular level [57].

The difficulty in telling apart different theories of molecular evolution stemmed in great part on the difficulty to measure random genetic drift. Since the seminal work of S. Wright, this is done through estimation of the so-called effective population size, $N_e$ [58]. The effective population size of a population, $N_e$, is the size of an idealized population, the Wright-Fisher population, that has experienced the same amount of random genetic drift than the population of interest. The Wright-Fisher population is a population in which individuals mate randomly and where the only source of change in allele frequencies is allele sampling. In other words, a population without structure and without selection or migration. Of course, this is a highly idealized situation. At neutral site, the effective size is directly and simply related to genetic diversity and this is captured by the quantity $\theta = 4N_e\mu$, where $\mu$ is the mutation rate and $\theta$ is genetic diversity that can be estimated by the number of pairwise differences between DNA sequences (Tajima’s $\pi$) [59] or by the number of segregating sites in DNA sequences (Watterson’s $\theta$) [60]. Hence, assuming equal or similar mutation rate, a more genetically diverse population has a larger $N_e$ than a less diverse one. The relationship between random genetic drift and $N_e$ is well illustrated in Figure 1.2. In a large population of 2000 individuals, an allele does not deviate too much from its original frequency of 0.5 over the time of 50 generations. However, when the population size is as small as 20 individuals, there is an increasing effect of randomness and an allele can get fixed (frequency = 1) or get lost (frequency = 0) even within 10 generations. If the three populations from the Figure 1.2 were real populations, they would strongly differ in their allele frequency simply due to stochasticity.

The effectiveness of selection also depends on $N_e$. Selection is more effective in populations with large $N_e$ because they are less affected by genetic drift, and the contrary is true for populations with small $N_e$. More precisely, the probabilities of an advantageous mutation to go to fixation and a deleterious mutation to be eliminated is a function of its initial frequency $1/2N_e$ and the product $N_e s$, where $s$ measures the strength of selection. The fate of new mutation is largely determined by selection when $N_e s$ is much greater than one, and by random genetic drift when $N_e s$ is much smaller than one. Thus, if the strength of selection is the same in two populations but their $N_e$ differ, levels of adaptation as the outcome of this selection will be different in these two populations.

The level of $N_e$ is affected by several factors. A population may go through a bottleneck, i.e. a strong decrease in size over a short period of time. This can greatly reduce $N_e$ and such populations will evolve largely under random genetic drift. A population bottleneck can be caused by some environmental event (e.g. drought, fire, overhunting), or when a small fraction of the original populations colonizes a new location (founder event). For example, selfing species likely go through repeated bottlenecks when they expand across a new
Figure 1.2. **Effect of population size on genetic drift.** Ten simulations each of random change in the frequency distribution of a single hypothetical allele over 50 generations for different sized populations; first population size n=20, second population n=200, and third population n=2000. (Image by Professor marginalia, distributed under a CC BY-SA 3.0 license).
Ne can also decrease with the switch of reproductive mode from outcrossing to self-fertilization, which frequently occurs in plants [61]. Ne can gradually increase when a population is growing or when new variation is introduced by migrants from other populations.

It can be concluded from the considerations above that natural populations are constantly under selection, leading to the evolution of adaptive traits, and genetic drift, leading to neutral variation. Distinguishing the two processes is one of the greatest challenges of evolutionary biology. This problem becomes particularly obvious in structured populations. Samples from two divergent populations exhibit differences that may look adaptive, but in fact, they are caused by genetic drift. One of the ways to separate the two processes is to assume that populations evolved largely neutrally if the variation between them can be explained by population structure. This is done through various mathematical models that take into account the population structure during tests for adaptive differences [62, 63, 64]. This control for population structure, however, may also mask real effects if the distribution of adaptive variation coincides with population structure. For example, the first association study between variation in Dwarf8, a gene regulating flowering, and flowering time in maize revealed that the association was due to population structure [65], but a later study with larger sampling proved that it was significant [66]. Thus, knowledge on the population structure of a species is extremely important if the aim is to find adaptive variation.

With growing evidence of inter-specific hybridization [14], admixture with other species in structured populations also need to be considered. One population of a species may be in contact with other not reproductively isolated species and this may lead to genomic introgression specifically in this population. These introgressed genomic regions can be neutral, but they may appear as the regions under selection in some selection tests (e.g. selective sweeps scans) because they stand out from the genome-wide and species-wide variation. Without the information on possible admixture with other species, such results may spuriously be interpreted as a local adaptive change because it is present only in one population. Admixture can also be strong enough to homogenize the difference between species and disrupt the whole-genome phylogeny for some populations. For example, a phylogenetic analyses of Heliconius elevatus and Heliconius pardalinus butterflies revealed that they are more genetically related to each other in the area where their distribution ranges overlap than to conspecific populations where these two species are geographically isolated [67]. An introgression can, of course, be under selection and truly adaptive. For example, in Spain, where house mice (Mus musculus domesticus) can hybridize with the Algerian mouse (Mus spretus), the majority of house mice have resistance to anticoagulant rodenticides that they acquired through introgression of the gene vkorc1 from the Algerian mouse [68]. House mice from other parts of Europe (Greece, Italy, UK), where the Algerian mouse does not occur, do not have vkorc1 that provide such resistance. Thus, know-
ing whether some populations of a species can potentially be admixed with any other species is extremely important.

The population genetics analyses of *Capsella bursa-pastoris* presented in this thesis illustrate well how a species can be structured into several populations and how these populations differ in their environments, effects of genetic drift, targets and strengths of selection. This thesis also demonstrates how different populations of one species can be influenced by admixture with sister species. It also provides a good example of challenges associated with analyses requiring a correction for population structure.

### 1.3 Challenges of studying allopolyploid organisms

To unravel the complex evolutionary history of allopolyploid organisms, one has to solve many practical and theoretical issues. Many allopolyploid species do not have a reference genome or the quality of their genomes is poor. This is primarily because genome assembly tools that reconstruct a whole genome from short DNA reads coming from a sequencing machine are not designed to resolve genome complexity of allopolyploids that have large genomes with many repeats, paralogs, and high heterozygosity. Genome assembly tools also generally do not handle properly the hybrid nature of allopolyploid genomes - regions of low divergence between two subgenomes would be collapsed to a single sequence by an assembler. Even such economically important species as bread wheat (*Triticum aestivum*) had only 61% of its genome sequence assembled [69] and only recent developments in both genomics and bioinformatics led to near-completion of its genome [70].

A general approach to deal with the absence of a reference genome for an allopolyploid species is to use an available reference genome of the closest diploid species. However, this approach has some issues that a researcher should be aware of. First, one of the subgenomes of the allopolyploid may be more related to the reference genome than the other subgenome and this will increase the success of mapping genomic reads from the more closely related subgenome. This mapping bias is not that critical for DNA data because at sufficient coverage most of the variation will be captured correctly. This is, however, a serious problem for gene expression analyses because mapping bias in RNA reads would directly influence the results. To reduce the severity of this mapping bias, one has to specify the correct divergence from the reference genome during mapping. Bias in the RNA data can also be reduced by removing sites that show strong biases in the DNA data. Second, to separate the two subgenomes, the mapping data needs to be phased. There are several programs that can phase the data into haplotype blocks [71], but the concatenation of these blocks into continuous sequences representing two subgenomes is only possible if the origin of an allopolyploid is known and parental genomic data is available. Allopolyploids usually exhibit disomic inheritance.
(two subgenomes do not recombine) and this allows the use of the information provided by fixed differences between the two subgenomes and the two parental species to perform the concatenation. Third, most of genome analysis software, as well as population genetic models, are developed for haploid or diploid data and they expect bi-allelic variation (two alleles per variable site). Allopolyploids, however, have four alleles at each genomic site and because of their hybrid nature, some variable sites may include more than two alleles. To deal with this issue, one usually has to treat tetraploid data as diploid data and remove all non-biallelic sites. This simplification is based on the fact that selfing and disomic inheritance that are common in allopolyploids lead to the strong reduction of heterozygosity within each subgenome and hence variation within subgenomes can be neglected.

The nature of allopolyploid data also makes it complicated to compare gene expression data between the two subgenomes of allopolyploid and the genomes of its diploid parental species. Traditionally, the gene expression level of a given gene is estimated by counting all reads that map to a gene. This is how the data is usually obtained for diploid parental species. On the other hand, gene expression data for the two subgenomes of an allopolyploid species can be obtained only by phasing the count data per variable site. For example, if a site is heterozygous AATT, one can obtain the number of reads with allele A and T correspondingly. Reads that are mapped to the region without polymorphism cannot be distinguished between two subgenomes and thus are skipped. This results in the reduction of the total read counts and under-estimation of expression level for genes with a low number of heterozygous sites. To correct this bias, one can use the unphased reads counts of an allopolyploid and transform them into the phased data by using the ratio between the two subgenomes that was estimated on the phased data. Suppose 30 reads map to a gene, 16 of which map to a polymorphic site AATT, 10 with allele A and 6 with allele T. Hence, the proportion of one subgenome is $6/(6 + 10) = 0.375$ and the other one is $1 - 0.375 = 0.625$. Extrapolating this proportion to total read count per gene gives $0.375 \times 30 = 11$ reads for the first subgenome and $0.625 \times 30 = 19$ reads for the second one. This method is not ideal because allopolyploid genes without any polymorphism are still skipped, but it makes the diploid and allopolyploid data more comparable.

All these challenges have been faced during the analyses of *C. bursa-pastoris* in this thesis. Some of the solutions presented above also have been developed during the work on this thesis.

1.4 A promising model system to understand allopolyploidy: *C. bursa-pastoris*

Genetic studies on the shepherd’s purse *Capsella bursa-pastoris* were started in the early twentieth century by G. Shull. G. Shull, then one of the most
prominent geneticists of his time [72] (among other things he introduced the concept and word "heterosis"), wanted to "carry Mendelian analysis beyond the field of domesticated plants and animals into the realm of wild nature where relationships have been unmodified by artificial breeding and selection by human agency". He and his associates carried out a very large amount of crosses (4000 pedigreed families!) from a worldwide collection and also performed some cytological studies. These studies led to a delineation of the species of the Capsella genus [73, 74, 75]. In particular, S. Hill showed that the species could be divided into two groups based on their ploidy level, a tetraploid and a diploid one [74]. The Capsella genus is today considered to consist of 4-5 species: C. grandiflora, C. orientalis, C. rubella, C. bursa-pastoris, and possibly C. thracica [76]. Among these species only C. bursa-pastoris is a polyploid with a worldwide distribution (Fig. 1.3). Actually, C. bursa-pastoris is one of the most common weeds in the world [77]. C. thracica is also a polyploid that is endemic to Bulgaria and was described as a separate from C. bursa-pastoris species [76]. However, given the recent information on the origin of C. bursa-pastoris [78], the status of C. thracica needs to be verified. It could as well be a divergent population of C. bursa-pastoris.

The widespread allotetraploid C. bursa-pastoris has recently emerged as a promising system for studying the evolution of allopolyploidy. First, information on its two progenitor diploid species and their current distribution is readily available. C. bursa-pastoris, a selfing species, originated from the hybridization of the Capsella orientalis and Capsella grandiflora/rubella lineages some 100-300 kya [78] (Fig. 1.3). Nowadays, C. orientalis, a genetically depauperate selfer, occurs across the steppes of Central Asia and Eastern Europe, while C. grandiflora, an extremely genetically diverse obligate outcrosser, is primarily confined to a tiny distribution range in the mountains of Northern Greece and Albania (Fig. 1.3). Second, self-fertilization, small size, yearly reproduction and low requirements for growing conditions make it easy to use C. bursa-pastoris in laboratory experiments (Fig. 1.4). Third, genomic data of C. bursa-pastoris can easily be phased into subgenomes [78, 79] due to a disomic inheritance [80]; thus, meiotic pairing happens between homologous chromosomes from the same ancestor and subgenomes inherited from different ancestors do not recombine. Fourth, genomic data of C. bursa-pastoris can be analyzed by mapping to the reference genome of C. rubella, a selfer recently derived from C. grandiflora, that was sequenced, assembled and edited in 2013 [81]. Moreover, high quality genome assemblies of C. bursa-pastoris, C. grandiflora and C. orientalis should be available soon (D. Weigel, personal comm.). Fifth, genomic studies of C. bursa-pastoris and other Capsella species benefit from their close relatedness to the best studied plant model species Arabidopsis thaliana. The estimated divergence time between the Arabidopsis and Capsella genera is about 10 million years [82, 83]. Sixth, the composition of the genus Capsella provides a good set up to study consequences of selfing and polyploidy. There have already been interesting studies on the genetic ba-
Figure 1.3. **Origin and distribution of the Capsella species.** The origin scheme shows the most likely scenario of the origin of *C. bursa-pastoris* through the hybridization between *C. orientalis* and *C. grandiflora/rubella* lineages and the increase of ploidy level to 4n. The approximate time of species splits is shown in thousands of years ago (kya). Reduced flowers in *C. bursa-pastoris*, *C. rubella*, and *C. orientalis* demonstrate the consequence of selfing when there is no need to attract pollinators. The map shows the limited distribution ranges of the diploid *Capsella* species in Eurasia, but *C. bursa-pastoris* is distributed worldwide because of its invasiveness.
sis of the shift to selfing and the so-called selfing syndrome [84, 85] and on the establishment of reproduction barriers [86, 87] in the diploid species of the genus. Our group also recently started a series of growth chamber experiments comparing the competitive ability of the different Capsella species. The first studies indicate that the two diploid selfing species have a much lower competitive ability than C. grandiflora, but C. bursa-pastoris has a competitive ability as high if not higher than C. grandiflora [88, 89]. It is, hence, as if the genome doubling associated with polyploidization restored the competitive ability lost because of the shift in the mating system. Seventh, worldwide distribution of C. bursa-pastoris exemplifies colonizing success [90]. Eighth, this worldwide distribution also enable studies of clinal variation in such traits as flowering time [91, 92, 93, 94]. Finally, evidence for gene flow between C. bursa-pastoris and its close relatives [95, 79] indicate that C. bursa-pastoris and the genus Capsella can also be used to investigate genomic introgression between diploids and polyploids.

In addition, this thesis demonstrates that C. bursa-pastoris provides a model to study alloploid evolution in structured populations [96, 94]. In Eurasia, C. bursa-pastoris is divided into three genetic clusters - Middle East, Europe, and Asia - with low gene flow among them and strong differentiation both at the nucleotide and gene expression levels (Papers I and II). This thesis also exemplifies that the genomic data of C. bursa-pastoris can easily be phased into subgenomes [79] (Paper III). It also reveals sophisticated regulation of gene expression that varies across different tissues (Paper IV). So, C. bursa-pastoris can be also used to study alloploid gene expression evolution across tissues and populations.

The puzzle posed by C. bursa-pastoris had already fascinated G. Shull, who wrote in 1929 "It is considered a matter of fundamental significance that the increase in the number of chromosomes in the bursa-pastoris group is correlated with greater variability, greater adaptability, greater vigor and greater hardiness" [75]. The results presented in this thesis further highlight the intricacy of C. bursa-pastoris and more generally the complexity of alloploidy evolution. I believe that after reading this thesis C. bursa-pastoris will fascinate you too.
Figure 1.4. *C. bursa-pastoris* growing in laboratory conditions.
2. Present study

The four papers of this thesis are sequential studies of the shepherd’s purse *C. bursa-pastoris*. First, we determined its population structure and colonization history. Then we compared phenotypic variation in the discovered populations to assess how much of this variation was due to local adaption and genetic drift. Finally, we investigated the evolution of the two subgenomes of *C. bursa-pastoris* by comparing their genomic and gene expression variation with each other and with the parental species overall and in different populations. The summaries of each of these studies are provided below.

**Paper I. Population structure and colonization history**

Polyploid species usually inhabit larger geographic areas and have wider ecological niches than their diploid progenitors [97] but our understanding of how they gained the capacity to colonize different localities with wide ranges of biotic and abiotic conditions remains limited. The first step to address this question is to characterize the population structure of a polyploid species and determine where it originated and how it spread. To that end, one has to obtain a wide-range sampling of a polyploid species, generate molecular data for many samples, and use fast and powerful computational tools to infer demographic histories. *C. bursa-pastoris* is a polyploid species that grows under a large range of environmental conditions world-wide [90] and its global population structure, place of origin and colonization history was not known. In this study, we genotyped a large sample of *C. bursa-pastoris* accessions with individuals sampled across Eurasia and a few sites from North America. We then analyzed its genetic structure and demographic history.

We genotyped 261 accessions of *C. bursa-pastoris* with genotyping-by-sequencing (GBS) and obtained 4,274 SNPs after quality control. The GBS technique uses restriction enzymes to selectively target DNA sequences outside of repetitive regions and to obtain SNPs in multiple samples with high efficiency [98]. Phasing of the GBS data was not feasible due to low coverage, so we analyzed the unphased data. To avoid bias in estimates of polymorphism, we removed sites showing fixed heterozygosity because most of them represented differences between the two subgenomes and not real polymorphism. We determined the population genetic structure with model-based ancestry inference implemented in *ADMIXTURE* [99]. The demographic history
was inferred by comparing alternative demographic models in the coalescent-based approximate Bayesian computation framework (ABC) [100]. To further characterize the inferred populations, we also estimated pairwise genetic differentiation between them [101] and their genetic diversity with Watterson’s theta ($\theta_W$) [60] and mean pairwise nucleotide differences ($\pi$) [59].

The ABC approach is a likelihood-free method based on summary statistics of the data that allows estimating several parameters (e.g. divergence time, migration rates, effective population size) of a given demographic model. For the sake of simplicity, let us assume a standard coalescent model with a single parameter, $\theta = 4N_e\mu$. Briefly, one first starts by calculating summary statistics of the data, for instance, the nucleotide diversity in the sample, $\pi_{obs}$. In a second step, values of the model parameter, in our example, values of $\theta$, are randomly sampled from a prior distribution and used in coalescent simulations under the demographic model which parameters we want to estimate. For each simulation run, we then compute the simulated summary statistic, $\pi_{sim}$, and compare it to the observed estimate, $\pi_{obs}$. If the difference between $\pi_{sim}$ and $\pi_{obs}$ is less than a threshold value (tolerance) we keep the value of $\theta$ used in the simulation run. Otherwise, we discard it. We repeat this a certain number of times until we obtain a posterior distribution of the parameter. ABC can handle complex demographics models and methods have been developed to compare the probability of the different models [102].

*C. bursa-pastoris* could be divided into three genetically distinct clusters: one cluster grouped accessions from Western Europe and Southeastern Siberia (hereafter EUR), a second included samples from the Middle East and Northern Africa (hereafter ME), and the third one united accessions from Eastern Asia (hereafter ASI) (Fig. 2.1). The ABC analysis supported the hypothesis that *C. bursa-pastoris* most likely originated in the Middle East and then spread towards Europe and later into Asia. The ME population diverged 167,250 years ago from an ancestral population, the EUR population diverged from ME 7,967 years ago, and the Asian population diverged from the Middle Eastern population 942 years ago. The estimated current effective population sizes ($N_e$) were 14,748, 21,130 and 25,376 for ASI, EUR and ME, respectively. The gene flow between these populations was moderate and each population underwent a bottleneck at its origin with the strongest bottleneck in ASI. The ME population, as the putative origin population, was the most genetically diverse and the ASI population was the least diverse.

A recent colonization of the world by *C. bursa-pastoris* was in agreement with earlier studies. Previous estimates of the origin time of *C. bursa-pastoris* were around 67,770 ya [95], 80,000 ya [103] and 128,000 ya [78], which were close to our estimate of the origin of the ME population some 167,250 year ago.
Figure 2.1. Genetic structure and ancestry of *C. bursa-pastoris* accessions in Eura-sia and Northern America. **A.** Map with the mean membership proportions inferred by *ADMIXTURE* for the three detected clusters in samples collected from the same site; $F_{ST}$: pairwise genetic differentiation [101]. **B.** Population structure of *C. bursa-pastoris* inferred with *ADMIXTURE*. Each vertical line represents an individual. Colors represent the inferred ancestry from three ancestral populations.

The discovered genetic differentiation between ASI, EUR and ME was also in line with previous studies [104, 95]. We further proposed that the colonization history of *C. bursa-pastoris* was human-mediated. This was primarily supported by the accessions from the Russian Far East that were geographically closer to the Asian cluster but genetically belonged to the European one. This long-distance dispersion was likely carried out by humans along the Trans-Siberian Railway, or through human colonization of Siberia in the 16th century. Similarly, the origin of the North American populations from the EUR and ME populations was best explained by human-mediated movement (see [76] and references therein).

As illustrated by the following three papers, this strong population structure and the recent multistage colonization history played a key role in the evolution of *C. bursa-pastoris*.

**Paper II. Adaptation and neutral processes**

The complex population structure and demographic history of species add to the difficulty of telling apart the effects of natural selection and genetic drift. In structured populations, observed phenotypic differences can simply be due to genetic drift, and if the genetic distance between populations is not con-
Figure 2.2. Differentiation between the three populations of *C. bursa-pastoris* in climatic parameters, gene expression levels and genetic variation. A. A PCA of 20 climatic variables for 65 populations. B. A multidimensional scaling plot of the expression levels in 19 484 genes from 24 accessions. Distances can be interpreted in terms of log2-fold-change (logFC). C. A PCA of 261 accessions based on 4 274 nuclear SNPs obtained by GBS.

sidered, the differentiation may be falsely interpreted as adaptive. The prevalence of neutral evolutionary processes has been suggested for gene expression evolution in various organisms, including humans [105], plants [106] and fish [107]. The variation in gene expression was often found to be correlated with genetic distance and thus was more likely a consequence of genetic drift than adaptive divergence. In the present study, we assessed the effect of population genetic structure on variation in gene expression and flowering time in *C. bursa-pastoris*.

To assess population structure, we used the GBS data generated for 261 accessions in Paper I [96]. The gene expression data was generated with RNA-Seq for a subset of 24 accessions. We also measured the variation in flowering time and circadian period length in the 261 accessions. We used flowering time and circadian rhythm as potentially adaptive traits because the onset of flowering at the right time ensures reproductive success, and the regulatory network of flowering time overlaps with the regulatory pathways of circadian rhythm [108, 109]. To compare the environmental differences between populations, we retrieved climatic data for the collection sites (20 bioclimatic variables) [110, 111]. The variation in climatic variables, gene expression, and GBS data was explored with a principal component analysis (PCA), multidimensional scaling plot (MDS), correlation tests and linear models. The correction for the population structure was done by using the phylogenetic distance between samples as a random effect in generalized linear mixed models.

The Asian (ASI), European (EUR) and Middle Eastern (ME) populations of *C. bursa-pastoris* showed clear differentiation in climatic conditions, gene expression variation and genetic polymorphism (Fig. 2.2). Variation in flowering time and circadian period length was significantly associated with these genetic groups. Controlling for population structure did not affect this associ-
ation indicating that variation in these traits could be adaptive. We also found significant differences between the three populations in gene expression. Out of 19 484 analyzed genes, 2 457 genes were differentially expressed between ASI and EUR, 1 317 between ASI and ME, 903 between EUR and ME. Some of these genes control flowering time variation and thus these differences could be adaptive. However, all significant expression differences were removed when correcting for the population genetic structure. This indicated that most differences were due to genetic drift. We assumed that these results could also be due to the confounding of population structure and distribution of adaptive traits. To bypass this effect, we also compared gene expression differences between early and late flowering ecotypes within the ASI and EUR clusters. We found different sets of genes showing differential expression in each of these two populations. Among these genes, \textit{FLOWERING LOCUS C (FLC)}, a gene known to regulate flowering time, was the best candidates for local adaptation in EUR.

These results exemplify how a demographic history that results in a strong population structure may complicate the interpretation of gene expression differences within a species. On the one hand, gene expression differences between populations of \textit{C. bursa-pastoris} detected without the correction of population structure could be interpreted as adaptive. However, all these differences were explained by the population structure. Thus, the potentially adaptive changes were likely false positives. It has been shown how easy it is to develop biologically sensible explanations of the adaptive importance for the difference detected in the data simulated under a purely neutral model [112]. On the other hand, the three populations inhabit regions with different climatic conditions and variation in gene expression between these populations is unlikely to be purely neutral. For example, variation in flowering time and circadian rhythm was not purely neutral. Thus, either the distribution of adaptive traits in \textit{C. bursa-pastoris} coincides with population structure or the effect of local adaptation, if present, is weak. Finally, local adaptation may get even more complicated if selection varies over the entire geographical range but acts at a very local scale. For example, it seems that neither humans nor \textit{Arabidopsis thaliana} have experienced large-scale local adaptation across their geographical range, but rather local selection within different populations [113, 114]. Evidence of selection on \textit{FLC} and some other genes in EUR and ASI also supports this hypothesis of a more local adaptation for \textit{C. bursa-pastoris}.

Paper III. Differential evolution of two subgenomes across populations

It remains largely unknown how the two genomes evolve once they have fused in an allopolyploid species and how strongly their evolutionary trajectories depend on the initial differences between the two parental species and the specific
degenerate history of the newly formed allopolyploid species. For example, species that went through repeated bottlenecks during their range expansion are expected to have reduced genetic variation and higher genetic load than more ancient central populations [115, 116]. Populations can also admix with different related species and such admixture could shift the evolutionary path of the local population. Finally, range expansion will expose newly formed allopolyploid populations to divergent selective pressures, providing the possibility of differentially exploiting the two subgenomes, and creating asymmetrical patterns of adaptive evolution in different populations. In this paper, we tested whether the two subgenomes of *C. bursa-pastoris* have similar or different evolutionary trajectories in term of hybridization, selection and gene expression in three known populations.

We phased 31 genomes and 24 transcriptomes for variation between the two subgenomes, one descended from the outcrossing and highly diverse *C. grandiflora* (hereafter *Cbp_{Cg}* ) and the other one descended from the selfing and genetically depauperate *C. orientalis* (hereafter *Cbp_{Co}* ). To perform the phasing, we used genomic data of 10 *C. orientalis* and 13 *C. grandiflora* as references. For each subgenome, we assessed its phylogenetic relationship with the diploid relatives, temporal changes of effective population size ($N_e$), signatures of positive selection, amount of accumulated deleterious mutations (genetic load), genetic diversity, and levels of gene expression. We also performed tests for admixture of the phased subgenomes of *C. bursa-pastoris* with other *Capsella* species.

Figure 2.3. **Genetic load in the subgenomes of *C. bursa-pastoris* and its parental species.** The proportion of deleterious nonsynonymous changes was estimated on derived alleles, i.e. alleles accumulated after the speciation of *C. bursa-pastoris*. Co and Cg are the two subgenomes of *C. bursa-pastoris*. ASI, EUR, ME, CO, CG indicate Asian, European and Middle Eastern populations *C. bursa-pastoris*, and parental species *C. orientalis* and *C. grandiflora*, respectively.
Figure 2.4. Phylogenetic relationships between subgenomes of *C. bursa-pastoris* and other *Capsella* species. **A.** Whole genome neighbor-joining phylogenetic tree. The bootstrap support based on 100 replicates is shown only for the major clades. The root (*Neslia paniculata*) is not shown. **B.** Overlapping of 1002 neighbor-joining trees reconstructed with 100 Kb sliding genomic windows. ASI, EUR ME, CO, CG, CR indicate Asian, European and Middle Eastern populations of *C. bursa-pastoris*, *C. orientalis*, *C. grandiflora*, and *C. rubella*, respectively. The *CbpCo* and *CbpCg* subgenomes are marked with Co and Cg, respectively.
We found that during $\sim$100,000 generations of co-existence within the same species, the effective population sizes, $N_e$, of the two subgenomes decreased gradually and converged in all three regions. However, the two subgenomes retained part of the initial difference between the two parental species. The $Cbp_{Cg}$ subgenome, inherited from a genetically more diverse parent, was also more diverse than the $Cbp_{Co}$ subgenome inherited from a selfing parent with a very low genetic diversity. The amount of genetic load also differed between subgenomes and indicated that purifying selection removed deleterious mutations more efficiently from $Cbp_{Cg}$ (Fig. 2.3) that had a larger effective population size than $Cbp_{Co}$. However, purifying selection was not as efficient as in outcrossing $C. grandiflora$, but it was more efficient than in selfing diploid $C. orientalis$. These differences in diversity and load, as well as differences in patterns of positive selection and levels of gene expression, also strongly depended on the specific histories of the three populations considered. The Asian population that was the most distant from the tentative place of origin of the species was the least genetically diverse and had the highest number of deleterious mutations. This population also experienced more selective sweeps on the $Cbp_{Co}$ subgenome, whereas the European and Middle Eastern populations experienced positive selection mostly on the $Cbp_{Cg}$ subgenome. Most strikingly, and unexpectedly, the three populations hybridized with different diploid relatives. There was strong hybridization between $C. orientalis$ and $C. bursa-pastoris$ in Asia and this admixture profoundly affected the phylogenetic relationship between subgenomes and parental species (Fig. 2.4). It looked like $C. orientalis$ was derived from the Asian $C. bursa-pastoris$. This pattern could also be retrieved if one assumes a scenario of multiple origins of $C. bursa-pastoris$, but it seems more likely to be the result of admixture. Admixture was also detected between $C. rubella$ and $C. bursa-pastoris$ in Europe. There were no major changes in gene expression between subgenomes, though the $Cbp_{Cg}$ subgenome was slightly more expressed than $Cbp_{Co}$ in Europe and the Middle-East, whereas there was rather equal expression between the subgenomes in Asia.

In summary, the shift to selfing and polyploidy had a global impact on the two subgenomes of $C. bursa-pastoris$ resulting in a sharp reduction of the effective population size in all populations, that was accompanied by relaxed purifying selection and accumulation of deleterious mutations. However, the two subgenomes still retained a strong signature of parental legacy and were differentially affected by introgression and selection in different geographical regions. Hence, this study illustrates that, first, the parental legacy has a long-term impact on the evolution of allopolyploid species, and second, the evolutionary trajectories of allopolyploid populations may differ depending on the ecological arena.
Paper IV. Convergence and parental legacy in the subgenomes

The impact of allopolyploidy on genome diversity and expression pattern still remains to be understood and put in a clear temporal perspective. Some studies have described genome upheaval and ”transcriptomic shock” [117, 118, 119], whereas others detected a strong legacy from the parental species and much more subtle and concerted changes [120, 121]. These differences across studies could result from many factors such as age and demographic history of the allopolyploid species or extent of the divergence between the parental species. In the present study, we investigated genomic and gene expression changes in the two subgenomes of the recent allopolyploid *C. bursa-pastoris*.

We sequenced the transcriptomes of three tissues (root, leaf, and flower) of *C. bursa-pastoris*, its two parental species, *C. grandiflora* and *C. orientalis*, and its close relative *C. rubella*. We compared the overall and homeologue-specific levels of expression between *C. bursa-pastoris* and its parental species. To control for mapping bias in gene expression data, we used the genomic data generated in the previous study (Paper III) [79]. We compared both the total expression level (\(C_{Cbp}Co + C_{Cbp}Cg\)) as well as homeologue-specific expression with correlation, differential expression analyses and a similarity index (\(S\)), which measured relative expression deviance of each subgenome from the average parental expression level. We also compared the distribution of synonymous and deleterious mutations between the two subgenomes with a model-based approach that accounts for an inherent bias towards one of the subgenomes. We then accounted for this bias and tested if the deleterious mutations were randomly distributed between the two subgenomes or whether they occurred more frequently on one of them than expected by chance. Finally, we checked whether the proportion of deleterious mutations per gene and gene expression levels were correlated in any way.

The gene expression ratio between subgenomes was overall equal (0.496) and not different from the ratio inferred in the DNA count data (0.497), where no difference between subgenomes is expected. Levels of gene expression were strongly correlated between the subgenomes (Spearman’s \(\rho\) 0.87-0.94), and expression in each subgenome was even more correlated with the corresponding parental species (Spearman’s \(\rho\) 0.88-0.97). These strong correlations were further supported by the differential gene expression analysis, which revealed that 55-80% of the genes depending on the tissue were of the same level in all three species (Table 2.1). However, a small proportion of genes showed the expression dominance of one of the parental species in the total level of expression. Notably, this dominance was opposite between the two tissues: the total expression level was more similar to *C. orientalis* in flower, and to *C. grandiflora* in root (Fig. 2.1). Interestingly, the expression of the selfing species, *C. rubella*, was the most similar to the expression pattern of its closest relative, the outcrossing *C. grandiflora* in all tissues including flower. This
Table 2.1. Gene expression levels in C. bursa-pastoris and the parental species

<table>
<thead>
<tr>
<th>Expression pattern</th>
<th>Flower</th>
<th>Leaf</th>
<th>Root</th>
</tr>
</thead>
<tbody>
<tr>
<td>No difference</td>
<td>8 805</td>
<td>12 656</td>
<td>10 278</td>
</tr>
<tr>
<td></td>
<td>(55.6%)</td>
<td>(80.0%)</td>
<td>(65.0%)</td>
</tr>
<tr>
<td>Additivity</td>
<td>1 498</td>
<td>713</td>
<td>941</td>
</tr>
<tr>
<td></td>
<td>(9.5%)</td>
<td>(4.5%)</td>
<td>(5.9%)</td>
</tr>
<tr>
<td>Dominance</td>
<td>2 164</td>
<td>506</td>
<td>687</td>
</tr>
<tr>
<td></td>
<td>(13.7%)</td>
<td>(3.2%)</td>
<td>(4.3%)</td>
</tr>
<tr>
<td></td>
<td>1 229</td>
<td>720</td>
<td>1 297</td>
</tr>
<tr>
<td></td>
<td>(7.8%)</td>
<td>(4.6%)</td>
<td>(8.2%)</td>
</tr>
<tr>
<td>Transgressive</td>
<td>1 139</td>
<td>565</td>
<td>1 226</td>
</tr>
<tr>
<td></td>
<td>(7.2%)</td>
<td>(3.6%)</td>
<td>(7.7%)</td>
</tr>
<tr>
<td></td>
<td>989</td>
<td>664</td>
<td>1 395</td>
</tr>
<tr>
<td></td>
<td>(6.2%)</td>
<td>(4.2%)</td>
<td>(8.8%)</td>
</tr>
</tbody>
</table>

CO, CG, and Cbp correspond to C. orientalis, C. grandiflora, and C. bursa-pastoris, respectively. The y-axes show levels of expression. The levels were considered different if they showed significant differential expression (FDR < 0.05). The most striking differences showing reversed levels of expression between tissues are highlighted in bold.
suggested that the detected similarity of *C. bursa-pastoris* and *C. orientalis* in flower tissues was not purely due to selfing. The variation between tissues was further confirmed by computing a similarity index (*S*), which was systematically biased toward the corresponding parental species in each subgenome but this bias varied between tissues. In addition, the correlation of *S* values of homeologous genes of two subgenomes suggested that these genes were generally regulated into the same direction. The slope of the regression between the expression ratio of the subgenomes over the expression ratio of the parental species was always less than one, indicating the presence of co-regulation of the two subgenomes through a mixture of *trans* and *cis*-regulation [122, 123]. The slope was less in flower tissue than in leaf and root tissues, suggesting a stronger *trans*-regulation in the former than in the latter. Finally, the deleterious mutations accumulated more on the *Cbp* subgenome relative to the synonymous mutations, the distribution of deleterious mutations between the two genomes had a smaller variance compared to the distribution of synonymous mutations. This suggests that both copies are needed and prevent to accumulate too many deleterious mutations. We also found no association between levels of expression and the number of deleterious mutations in genes.

To conclude, since the speciation of *C. bursa-pastoris* around 100,000 years ago, some of the major differences between the ancestral genomes started to be erased and the two subgenomes have started to act in a coordinated manner. However, a significant legacy effect on the number of deleterious mutations carried by the two subgenomes, and on the total expression level in the three tissues can still be detected.
3. Conclusions and future perspectives

In the twenty-first century, the research on polyploidy has become particularly molecular-centered due to the availability of huge amount of whole-genome sequencing data. Although this thesis is also primarily based on the whole-genome data, it differs from many modern works by considering not only molecular but also demographic and ecological aspects of polyploidy evolution. In Paper I, we have shown that *C. bursa-pastoris* is structured into three genetically distinct populations that have different effective population size and thus are unequally affected by random genetic drift. In Paper II, we further bolster the differentiation between these populations by showing the differences in climatic, phenotypic and gene expression variation between them. In Paper III, we reveal that selection targets different subgenomes across these populations and that the populations carry a different amount of genetic load. Moreover, *C. bursa-pastoris* hybridizes with different species depending on geographical region and this hybridization has a profound effect on molecular variation in some populations. All these results emphasize the importance of studying polyploid species across their distribution range because evolutionary processes may substantially differ across populations and sampling only one population could lead to biased results and misleading conclusions.

The major global changes relative to the parental species would, of course, be evident in all populations of an allopolyploid. In *C. bursa-pastoris*, these changes include a steep decrease in the effective population size, relaxed purifying selection and no strong expression dominance between subgenomes. There is also a long-lasting effect of parental legacy in the amount of genetic load between subgenomes. From these findings, it can be concluded that *C. bursa-pastoris* experienced some events shared by many polyploids. For example, it went through a demographic bottleneck and shifted to selfing which led to reduced efficacy of natural selection. Gene expression in *C. bursa-pastoris* as in many other allopolyploids was not disrupted, though some allopolyploids experience profound expression changes (so-called “transcriptomic shock”). However, I still would like to point out that smaller scale variation in these effects across population was significant and cannot be ignored.

This thesis also addresses the question of adaptation in polyploids. Our results on local adaptation seem to strongly depend on the levels of integration and geographical scale that are investigated. Signs of global adaptation in flowering time and circadian rhythm variation are observed in Paper II. On the other hand, gene expression variation does not depart from neutral expectations. In Paper III, variation in positive selection across subgenomes depends
on population and could be considered as indirect evidence of local adaptation. However, this inference is not straightforward as the observed patterns could also be due to complex demographic history, especially admixture with different sister species. The general view of low local adaptation, or possibly maladaptation in some subpopulations, is also supported by a large-scale, multi-site, common garden experiment that is not included in this thesis (Manuscript in prep.). Hence, despite being distributed across diverse environments and climates around the globe, the evolution of *C. bursa-pastoris* seems to have been dominated by neutral rather than adaptive processes.

In addition, this thesis further exemplifies how *C. bursa-pastoris* can be used as a model species to study polyploid evolution. In Paper III and IV, we successfully phased a major part of *C. bursa-pastoris* genome. Such phasing is only possible because *C. bursa-pastoris* is of known origin, it exhibits a strictly disomic inheritance, the genomic data of parental species is available, it has a small genome, and the fully assembled and annotated genome of a closely related species is available. These conditions are not met in many allopolyploid species. Paper III shows how this phased data sheds new light on the differential evolution of the subgenomes across populations of *C. bursa-pastoris*. Paper IV, that is based on the phased expression data of *C. bursa-pastoris* and its parental species, demonstrates that gene expression regulation can vary across different tissues and that there is a strong legacy from the parental species.

Finally, these thesis has already shed light on many questions of polyploid evolution, but if I had a chance to continue working on this project, I would work in the following directions. First, I would further extend Paper IV with additional gene expression analyses and a stronger focus on different populations. Second, I would verify the demographic history reconstructed in Paper II but this time with the ABC analysis of the phased genomic data. Third, I think one of the most intriguing questions would be to compare the absolute expression levels between *C. bursa-pastoris* and its diploid relatives at the single-cell level. Fourth, it would be most exciting to analyze genomic data with good quality reference genome assemblies of all four *Capsella* species to eliminate all technical issues I had to cope with during the work on this thesis. Finally, as a long time perspective, it would also be interesting to look at small RNAs and epigenetic regulation in *C. bursa-pastoris*. 
Svensk sammanfattning


asiatiska gruppen. Detta indikerade att denna spridning mest sannolikt utfördes av människor längs den transsibiriska järnvägen, eller genom kolonisering av Sibirien av människor under 1500-talet. På samma sätt är ursprunget för de nordamerikanska populationerna från Europeiska och Mellanöstern-populationer sannolikt också resultatet av människoformedlad spridning.

I Artikel II testade vi om det finns tecken på lokal anpassning i genuttryck och variation av blomningstid i *C. bursa-pastoris* samt om det finns någon effekt av den upptäckta populationsgenetiska struktur på variation av dessa två egenskaper. I strukturerade populationer kan de observerade fenotypiska skillnaderna helt enkelt bero på genetisk drift, och om det genetiska avståndet mellan populationer inte beaktas, kan differentieringen felaktigt tolkas som adaptiv. Vi upptäckte att populationerna av *C. bursa-pastoris* från Asien, Europa och Mellanöstern visade tydlig differentiering i klimatförhållanden, genuttryck och blomningstid. Variation i blomningstid, en potentiellt adaptiv egenskap eftersom blommingsstart vid rätt tidpunkt säkerställer reproduktiv framgång, påverkades svagt av populationens genetiska struktur vilket indikerade att variation i denna egenskap kunde vara adaptiv. Emellertid försvann alla skillnaderna i uttryck efter korrigerande för populationens genetiska struktur. Detta indikerade antingen att de flesta skillnaderna i genuttryck var neutrala eller att den lokala anpassningen, om den var närvarande, var svag och korrelerad med populationens strukturen. Vi upptäckte också några bevis för att anpassningen kunde vara mer lokal och variera inom varje genetisk grupp.

I Artikel III och IV fokuserade vi på utvecklingen av de två subgenomen hos *C. bursa-pastoris* och mängden kvarvarande föräldrarväv i dessa två subgenom. Vi upptäckte att båda subgenomen uttryckte kraftig minskning av genetisk mångfald efter artbildningen av *C. bursa-pastoris*, förmodligen på grund av självbefruktning. Trots deras gemensam historia som en art behöll de två subgenomen också skillnader i mängden genetisk belastning som ärvt av föräldraarterna och påverkades differentiellt av introgression och urval i de tre populationerna. Variationen i genuttryck var högt korrelerad mellan de två subgenomen men en liten del av generna visade differentiering över populationer och vävnader. Det fanns flera gener som visade överuttryck av subgenomet som ärvt av *C. grandiflora* i Mellanöstern och Europa, medan de två subgenomen uttrycktes lika i Asien. I blommor hade *C. bursa-pastoris* ett uttryck som liknar *C. orientalis* både i den övergripande nivån av transskript och på subgenomnivån, medan den i blad och rötter liknade *C. grandiflora*.

Sammanfattningsvis visar denna avhandling för första gången att allopolyploder kanske inte är homogen i sina distributionsområden och de kan också markant följa olika utvecklingsvägar i olika populationer. Denna avhandling visar också att trots om sammanslagningen av två divergerande genom i allopolyploder resulterar i deras konvergens, så kan vissa skillnader bibehållas och troligtvis möjliggör dessa skillnader evolutionär flexibilitet hos allopolyploida genom.
Алополіплоїд - це організм, який утворився шляхом гібридизації та одночасного збільшення набору хромосом (поліплоїдія). Гібридизація та поліплоїдія зазвичай мають негативні наслідки для організму. Однак є види, які не тільки виживають при цих модифікаціях, а навіть процівтівають та конкурують з батьківськими видами. Є кілька теоретичних пояснень цього явища, але кількість практичних досліджень обмежена.

Вид *Capsella bursa-pastoris*, відомий як грицики звичайні, є перспективною моделлю для вивчення таких успішних алополіплоїдних видів. *C. bursa-pastoris* поширення по всьому світі, тоді як її батьківські види, *Capsella grandiflora* та *Capsella orientalis*, мають значно менші ареали. Поширення *C. grandiflora* обмежується Грецією та Албанією, а *C. orientalis* зустрічається в Середній Азії. Генетичні дослідження грициків звичайних були розпочаті на початку XX століття. Проте дослідження еволюційної біології *C. bursa-pastoris* було розпочато значно пізніше. Наприклад, походження *C. bursa-pastoris* залишалося невідомим досить довго і тільки кілька років тому з’ясувалося, що *C. bursa-pastoris* є алополіплоїдом. Про всеєвітне поширення *C. bursa-pastoris* було відомо давно, але не було встановлено чи існує у цього виду певна генетична структура популяції, звідки цей вид походить і яким чином він поширився. Алополіплоїдне походження *C. bursa-pastoris* також ставить питання щодо еволюції її двох субгеномів, які були успадковані від досить різних батьків - генетично бідного пращура *C. orientalis*, який розмножується самозапиленням, та надзвичайно генетично різноманітного пращура *C. grandiflora*, який розмножується перехресним запиленням. Також що *C. bursa-pastoris* зустрічається в різних середовищах по всьому світі, як вона адаптується до цих різноманітних умов? Я проаналізував геномні, транскриптомні та деякі фенотипічні дані *C. bursa-pastoris* для вирішення всіх цих питань в цій дисертації.

У статьї 1 ми описали генетичні варіації *C. bursa-pastoris* і показали, що цей вид не є однорідними, а поділяться на три генетично відмінних популяцій на території Євразії: одна розповсюджена в Європі та на Сході Сибіру, друга - в Східній Азії, а третя - на Близькому Сході та Північній Африці. Реконструкція історії колонізації цих регіонів показала, що цей вид зародився на Близькому Сході й згодом поширився в Європу та Східну Азію. Ми також виявили, що зразки зі Сходу Сибіру генетично належали до європейської групи, але були розташовані ближче до азійської. Це дозволило припустити, що така далека колонізація швидше за все здійснилася за допомогою людей вздовж Транссибірської магістралі, або під
час колонізації Сибіру людьми в XVI столітті. Аналогічним чином, походження північноамериканської популяції від європейської та близькосхідної популяції імовірно є наслідком розповсюдження з людьми.

У статті II ми перевірили наявність ознак локальної адаптації в експресії генів та варіації часу початку цвітіння, а також перевірили чи варіації цих двох рис залежать від виявленої генетичної структури. Фенотипічні відмінності, що спостерігаються у структурованих популяціях, можуть бути пов’язані з дрейфом генів, і якщо не брати до уваги генетичну відстань між популяціями, відмінності можуть бути помилково інтерпретовані як адаптивні. Ми виявили, що всі три популяції демонструють чітке розмежування в кліматичних умовах, експресії генів та в часі початку цвітіння. На варіацію в часі цвітіння, як потенційно адаптивної риси, оскільки початок цвітіння в потрібний час забезпечує репродуктивний успіх, популяційна генетична структура мала слабкий вплив, що вказує на те, що варіація цієї ознаки може бути адаптивною. Проте всі відмінності в експресії генів зникли після корекції даних на генетичну структуру популяції. Це вказує на те, що більшість відмінностей в експресії генів є нейтральними або місцева адаптація, якщо вона присутня, є слабкою та збігатися зі структурою популяції. Ми також виявили деякі докази, що адаптація може бути більш локальною в межах кожної генетичної групи.

У статтях III та IV ми зосередили увагу на еволюції двох субгеномів та кількості батьківських рис, що залишилися в цих двох субгеномах. Ми виявили, що обидва субгенома проявляють зниження генетичної різноманітності після видоутворення C. bursa-pastoris, ймовірно через самозапилення. Проте, не зважаючи на спільну історію в межах одного виду, ці два субгеноми також зберегли відмінності в кількості генетичного тягаря, успадкованого від батьківських видів, а також зазнали різного впливу відбору та гібридизації з близькими видами. Варіація експресії генів сильно корелювала між субгеномами, але мала частка генів відрізнялася в експресії між популяціями та тканинами. На Близькому Сході та в Європі більша частка генів демонструвала надмірну експресію субгеному успадкованого від C. grandiflora, тоді як в Азії субгеноми були більш рівні між собою. У квітах експресія генів була більш подібна до C. orientalis, тоді як у листках та корінні вона була більш схожа з C. grandiflora.

На завершення варто зазначити, що ця дисертація вперше демонструє, що алополіплоїди можуть бути генетично неоднорідними по терitorії поширення й еволюційні шляхи можуть різнитися між популяціями. Ця робота також виявила, що хоча об’єднання двох різних геномів у алополіплоїдах призводить до їх еволюційного зближення, деякі відмінності можуть бути збережені довго, і ці відмінності ймовірно забезпечують еволюційну гнучкість алополіплоїдних видів.
Acknowledgements

There were many people who helped me during the course of this thesis as well as on the way to it and I would like to reflect on all of them here.

First of all, I would like to express my sincere gratitude to Martin Lascoux, my Ph.D. supervisor, for giving me a chance to do this Ph.D. project. I am very thankful for your patient guidance, enthusiastic encouragement, gentle criticism and constant positivity. These four years have been exciting, challenging and gratifying. You taught me to like negative and intricate results. I also have to admit that thanks to you I do appreciate neutral evolutionary processes now.

I would like to express my very great appreciation to Stephen Wright and Sylvain Glémin for their advice and assistance during the work on this Ph.D. thesis. Stephen, thank you for your hospitality during my visit to Toronto and for your prompt replies during our remote communication. Sylvain, your ideas were fascinating and inspiring. I would also like to thank Ulf Lagercrantz for co-supervising this project. Your assistance on the RNA-Seq experiments was very valuable.

Special thanks go to Karl Holm, Amandine Cornille, Jun Chen, Maria Guerrina, Clément Lafon-Placette and Marion Orsucci for their cooperation in some projects of this thesis. Karl, your time and willingness to help when I just started my Ph.D. studies is greatly appreciated. Amandine, it has been fun to share the office with you and to experience fickle Swedish weather during the common garden experiment. Jun, your expertise in genomic analyses was invaluable. Maria and Clément, thank you for helping with the crosses. Marion, I greatly appreciate your help at the final stage of this thesis when there was strong time pressure.

I also wish to thank Mimmi Eriksson and Tianlin Duan for doing their Master’s project with us. Your persistence and independence was impressive and the results you produced were a great contribution to this thesis. My grateful thanks also go to our collaborators in Toronto. Adriana Salcedo, Julia Kreiner, Tyler Kent, Huirun Huang and John Stinchcombe, working with you has been pleasant and productive.

I also wish to recognize the valuable help of Kerstin Jeppsson whose assistance in the lab was vital. Without your responsive help and timely provision of chemicals, none of my Ph.D. projects would be finished in time.

I express my very profound gratitude to my friend Glib Mazepa, who encouraged me to apply to this Ph.D. program and helped with the application process. Your contribution to this thesis was indirect but extremely valuable.
My sincere thanks also goes to Yevgen Ryeznik for his continuous encouragement. Talking to you in some down moments helped me to keep going. It was also a pleasure to work with you on our Data Science project.

My grateful thanks are also extended to all people at Evolutionary Biology Center and the Department of Plant Ecology and Evolution, in particular. I am thankful to Andres Cortes, Xiaodong Liu, Lili Li, Judith Trunschke, Elodie Chapurlat, Giulia Zacchello, Luis Leal, Anna-Malin Linde, Fia Bengtsson, Charles Campbell, Matthew Tye, Linus Söderquist, Froukje Postma, Thomas Ellis, Venkat Talla, Ludovic Dutoit, Nagarjun Vijay, Moos Blom, Foteini Spagopoulou, Homa Papoli, Paulina Bolivar, William Silva, Sergio Tusso, Lore Ament, Kristaps Sokolovskis, Josefine Stångberg, Ivain Martinossi, Matthias Weissensteiner, and Jesper Svedberg for stimulating scientific discussions and all the fun we had together.

I also would like to thank the organizers of all courses I took during my Ph.D. studies. I am especially grateful to Lucas Sinclair, Ludovic Dutoit, and Alexander Eiler for the Python course, Raazesh Sainudiin for the Data Science course, Andreas Dahlin for the Visualize your Science course, Terese Bergfors for the Writing Scientific English course, and Göran Arnegård for the Statistics course. These courses had the most profound impact on my scientific skills.

Probably, I would never be able to do this Ph.D. thesis without the knowledge and experience I gained during the Erasmus Mundus Master Program in Evolutionary Biology. I would like to pay my regards to people who run this program, in particular, Franjo Weissing, Irma Knevel and Jacob Höglund.

When I look back and think about the time I decided to pursue a scientific career, I wish to acknowledge my very first scientific supervisor, Gennadiy Shandikov. You kept telling me to do what I like and dispelled all my doubts. I am glad I followed your advice and I owe you a debt of gratitude for the very first step towards a Ph.D. degree.

I must express my very profound gratitude to my English teacher, Nikolay Yakymenko. You opened me a door to international science and made this Ph.D. thesis possible.

Finally, this accomplishment would not have been possible without unfailing support from my parents, sister, and wife. They have given up many things for me to do this Ph.D. thesis. Thank you for your patience, understanding and motivation during these last four years and my life in general. My sister deserves special thanks for designing the cover of this thesis. I also wish to thank my little daughter who was born just at the time of writing this thesis. I will blissfully remember restless days and sleepless night in the struggle to write this thesis and to calm you down.
References


H. B. D. Kettlewell, “Selection experiments on industrial melanism in the


[94] D. Kryvokhyzha, K. Holm, J. Chen, A. Cornille, S. Glémin, S. I. Wright,
U. Lagercrantz, and M. Lascoux, “The influence of population structure on
gene expression and flowering time variation in the ubiquitous weed *Capsella
bursa-pastoris* (Brassicaceae),” *Molecular Ecology*, vol. 25, no. 5,

not confer instant reproductive isolation in *Capsella* (Brassicaceae),”

[96] A. Cornille, A. Salcedo, D. Kryvokhyzha, S. Glémin, K. Holm, S. I. Wright,
and M. Lascoux, “Genomic signature of successful colonization of eurasia by
the allopolyploid shepherd’s purse (*Capsella bursa-pastoris*),” *Molecular

[97] C. Parisod, “Polyploids integrate genomic changes and ecological shifts,”

[98] R. J. Elshire, J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler,
and S. E. Mitchell, “A robust, simple genotyping-by-sequencing (GBS)
approach for high diversity species,” *PloS one*, vol. 6, no. 5, p. e19379, 2011.

ancestry in unrelated individuals,” *Genome research*, vol. 19, no. 9,

computation in population genetics,” *Genetics*, vol. 162, no. 4,


[102] M. A. Beaumont, “Approximate Bayesian Computation in Evolution and

from next-generation sequence data,” *Molecular ecology*, vol. 24, no. 5,

history from chloroplast microsatellites in the cosmopolitan weed *Capsella
bursa-pastoris* (Brassicaceae),” *Molecular ecology*, vol. 14, no. 14,


[106] M. Broadley, P. White, J. Hammond, N. Graham, H. Bowen, Z. Emmerson,
transcriptome evolution in plants,” *New Phytologist*, vol. 180, no. 3,

expression,” *Proceedings of the National Academy of Sciences*, vol. 103,

G. Coupland, “CONSTANS mediates between the circadian clock and the

[109] F. Valverde, A. Mouradov, W. Soppe, D. Ravenscroft, A. Samach, and


A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)