Taking the long way around:

Worldwide geographical structure of the cosmopolitan weed *Capsella bursa-pastoris* (Brassicaceae)

UNIVERSITÄT OSNABRÜCK

Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)
des Fachbereichs Biologie/Chemie
der Universität Osnabrück

vorgelegt
von

Christina Wesse

aus
Bad Zwischenahn

Osnabrück, 15. April 2019

Tag der mündlichen Prüfung: 11. Juli 2019

“The Trees are so much taller
And I feel so much smaller
The Moon is twice as lonely
And the Stars are half as bright.”

Meinem lieben Vater,
der immer stolz auf mich war.

Abstract

The study of population biology and genetic diversity provides insights to the potential for colonization and can detect geographic patterns of invasion and range expansion, which is essential to predict how species might react to dynamic environments and the global climate change. An outstanding example for a successful colonizer is the Shepherd's Purse (*Capsella bursa-pastoris* (L.) Medik.). It is closely related to *Arabidopsis thaliana*, the "lab rat" of plant scientists, and originated 100-300 kya from the hybridization between an ancestral *C. orientalis* and an ancestor from the *C. grandiflora/rubella* lineage according to the current literature (Douglas et al., 2015). Many species invasions are the direct or indirect consequence of human activities, and the worldwide distribution of the Shepherd's Purse is partially associated with prehistorical human migration (e.g. Neuffer & Hurka, 1999; Cornille et al., 2016).

With the novel genome-wide restriction site-associated DNA sequencing (RADseq) it is possible to perform population genetic studies of unprecedented depth and complexity and allowed the exploration of evolutionary history, range expansion and invasion patterns of this plant species. I will show here that a large number of loci and a wide global sampling area, using seed collections from nearly all over the world covering a large part of the whole distribution area of this ubiquitous weed, reveal finer-scale population structure of *C. bursa-pastoris* than has previously been detected.

The work proposed here generates a comprehensive picture of phenotypic diversity in relationship to genetic variation within *C. bursa-pastoris*. Genetic variation is clearly geographically structured and split into two lineages apparently adapted to different environments, with one population predominantly distributed in Mediterranean climate regions and the other predominantly in temperate climate regions. The worldwide distribution patterns of the genetic variation can be explained by intra- and intercontinental migration, but environmental filtering due to climate pre-adaption seems also involved. The two clusters point to an early diversification into two lineages or may even suggest multiple origins of the species.

This dissertation consists of three papers and manuscripts written during my time as a doctoral student at the Osnabrück University.

Zusammenfassung

Das Fachgebiet der Populationsgenetik liefert Erkenntnisse über das Kolonisierungspotenzial von biologischen Arten und kann angewendet werden geografische Muster von Invasionen und Reichweitenausdehnungen zu erkennen, die zum Beispiel für die Vorhersage, wie Arten auf dynamische Umgebungen und den globalen Klimawandel reagieren könnten, unerlässlich sind. Ein herausragendes Beispiel für einen erfolgreichen Kolonisator ist das Hirtentäschelkraut (*Capsella bursa-pastoris* (L.) Medik.). Diese weit verbreitete krautige Pflanze ist eng mit der Ackerschmalwand (*Arabidopsis thaliana*), der „Laborratte" der Pflanzenwissenschaften, verwandt und entstand laut aktueller Literatur vor rund 100 bis 300 tausend Jahren aus der Hybridisierung zwischen einer angestammten *C. orientalis* und einem Vorfahren aus der *C. grandiflora/rubella*-Linie (Douglas et al., 2015). Viele Arteninvasionen sind die direkte oder indirekte Folge menschlicher Aktivitäten, und die weltweite Verbreitung des Hirtentäschels wird zum Teil mit vorhistorischer menschlicher Migration in Verbindung gebracht (z.B. Neuffer & Hurka, 1999; Cornille et al., 2016).

Mit der neuen Methode der genomweiten Restriktionsstellen-assoziierten DNA-Sequenzierung (RADseq) ist es möglich, populationsgenetische Studien von beispielloser Tiefe und Komplexität durchzuführen und die Evolutionsgeschichte, Expansion und Invasionsmuster dieser Pflanzenart zu erforschen. Ich werde in dieser Dissertation zeigen, dass eine große Anzahl von Loci und ein großer globaler Probenahmebereich mit Saatgut aus allen möglichen Gebieten der Welt, der einen großen Teil des gesamten Verbreitungsgebiets dieses allgegenwärtigen Unkrauts abdeckt, eine feinere Populationsstruktur des Hirtentäschels offenbaren, als dies zuvor nachgewiesen wurde.

Die hier vorgeschlagene Arbeit erzeugt ein umfassendes Bild der phänotypischen Vielfalt in Bezug auf die genetische Variation bei *C. bursa-pastoris*. Die genetische Variation ist klar geografisch strukturiert und in zwei Hauptgruppen unterteilt, die anscheinend an unterschiedliche Umgebungen angepasst sind, wobei eine Population überwiegend in mediterranen Klimaregionen und die andere überwiegend in gemäßigten Klimaregionen verteilt ist. Die weltweiten Verteilungsmuster der genetischen Variation lassen sich durch intra- und interkontinentale anthropogene Migration erklären, aber auch Klimaanpassung scheint beteiligt zu sein. Die beiden Cluster deuten auf eine frühe Diversifikation in zwei Linien hin oder können sogar auf mehrere Ursprünge der Art hinweisen.

Diese Dissertation besteht aus drei Papern und Manuskripten, die während meiner Zeit als Doktorandin an der Universität Osnabrück entstanden sind.

List of Contents

List of Figures

List of Tables

List of Abbrevations

AFLP	Amplified fragment length polymorphisms
bp	Base pairs
BWA	Burrows Wheeler Aligner
CTAB	Cetrimonium bromide
CV	Cross-validation
DAPI	4′,6-diamidino-2-phenylindole
DNA	Deoxyribonuleic acid
EB	Elution buffer
EDTA	Ethylenediaminetetraacetic acid
FAMD	Factor analysis of mixed data
FCM	Flow Cytometry
GBS	Genotyping by sequencing
GWAS	Genome-wide association study
kb	Kilo base pairs
LD	Linkage disequilibrium
MAF	Minor allele frequency
Mbp	Mega base pairs
MEM	Max Exact Matches
MFA	Multiple factor analysis
MID	Molecular identifier
MMG	Mediterranean Multilocus Genotype
MPS	Massive(ly) parallel sequencing
NGS	Next-generation sequencing
PAGE	Polyacrylamide gel electrophoresis
PCA	Principal component analysis
PCR	Polymerase chain reaction
PE	Paired-end sequencing
RADseq	Restriction site-associated DNA sequencing
RFLP	Restriction fragment length polymorphisms
RRLS	Reduced-representation libraries sequencing
rcf	Relative centrifugal force
SBS	Sequencing by synthesis
SE	Single-end sequencing
SNP	Single nucleotide polymorphism
WGRS	Whole genome re-sequencing

1. Introduction

1.1 Aim of this Thesis

Many species invasions are the direct or indirect consequence of human activities. Exotic plants for example have been imported intentionally for medical purposes or ornamentation, but also accidental as by-catch in crop seeds or adhesion to domesticated animals (Sakai et al., 2001). A successful establishment of a species into a new habitat involves phenotypic plasticity and the potential for genetic changes through drift or selection (Sakai et al., 2001). Range expansion is a feature of the evolutionary history of all species, whether intercontinental or on a more local scale. The study of population biology and genetic diversity provides insights to the potential for colonization and can detect geographic patterns of invasion and range expansion, which is essential to predict how species might react to dynamic environments and the global climate change.

An outstanding example for a successful colonizer is the Shepherd's Purse (*Capsella bursa-pastoris* (L.) Medik.), a tetraploid and predominantly inbreeding flowering plant. This rather common plant was widely distributed throughout whole Eurasia and around the Mediterranean Sea in prehistoric times by early agricultural activities of humans and colonized other continents from the beginning of the 16th century (Mooney et al., 2005). These unintentional transports allowed *C. bursa-pastoris* to also reach South America, Australia, South Africa and nearly every other possible locality, avoiding merely the very hot and humid tropics and arctic climates (Neuffer & Hurka, 1999; Neuffer et al., 1999; Neuffer et al., 2011; Kryvokhyzha et al., 2016; Fig. 1).

It is fascinating how evoultionary processes drive species distribution, and an interesting question whether populations from newly colonized continents differ from the source continent and whether colonization success is primarily due to the introduction of pre-adapted genotypes, or due to new mutation and *in situ* genetic diversification. In order to determine the colonization history, a global survey of genetic diversity must be performed over the whole distribution area of this species. The colonization history of this plant has been traced in parts by molecular markers in previous studies (e.g. RAPDs in Neuffer, 1996; isozymes in Neuffer & Hurka, 1999; isozymes and RAPDs in Neuffer et al., 1999; chloroplast microsatellites in Ceplitis et al., 2005; GBS in Kryvokhyzha et al., 2016), but a worldwide over-all view is still missing. Therefore, I would like to report a more extensive sampling from sites from every continent except Antarctica in this thesis and describe the genetic variation in relationship to phenotypic plasticity. Furthermore, with the novel genome-wide marker analyses such as restriction site-associated DNA sequencing (RADseq), it is now possible to perform population genetic studies of unprecedented depth and complexity and allows the exploration of evolutionary history, range expansion and invasion patterns of this plant species. I will show here that a large number of loci and a wide global sampling area reveal finer-scale population structure of *C. bursa-pastoris* than has previously been detected.

The main part of this thesis is based on the following peer-reviewed papers and manuscripts, which are referred to in the text by their Roman numerals (Reprint was made with permission from the publisher):

I. **Neuffer, B., Wesse, C., Voss, I., & Scheibe, R. (2018):** The role of ecotypic variation in driving worldwide colonization by a cosmopolitan plant. *AoB Plants*, *10*(1), ply005.

   Ecotypic differentiation of plant species has been a major topic or research for almost 100 years now. Many studies demonstraded the ecotypic differentiation of *C. bursa-pastoris* in various regions but the adaptability of anatomy and physiology of rosette leaves so far remained less recognized. In this published paper, we highlight leaf adaptations of *Capsella bursa-pastoris* such as thickness of the mesophyll and epidermis, stomatal density, photosynthetic capacity and resistance to high light conditions.

II. **Wesse, C., Hurka, H., Welk, E., & Neuffer, B.:** Geographical structure of genetic diversity in Shepherd's Purse, *Capsella bursa-pastoris* – a global perspective. In prep.

   In this manuscript, we display and analyze global geographical distribution patterns of isozyme genotypes of *Capsella bursa-pastoris* and describe the driving forces of the resulting distribution patterns. We sampled 21,812 individuals randomly taken from natural provenances, covering a broad spectrum of the distribution range. Polyacrylamide gel electrophoresis was performed to assay different isozyme systems and the population structure was analyzed with STRUCTURE. We detected two clusters clearly adapted to different environments. We explain the worldwide distribution patterns of the genetic variation by intra- and intercontinental migration, but environmental filtering due to climate pre-adaption seems also involved.

III. **Wesse, C., Koenig, D., Neuffer, B., & Weigel, D.:** Taking the long way around - Worldwide geographical structure of the cosmopolitan weed *Capsella bursa-pastoris* (Brassicaceae). In prep.

   The work proposed here generates a comprehensive picture of phenotypic diversity in relationship to genetic variation within *C. bursa-pastoris*. A combination of phenotyping and SNP sequencing data from 1,273 individuals from 384 different collection sites was used. Most interestingly, the previously observed worldwide population structure obtained from isozyme data (see *II.*) is also found via RADseq analysis. The two clusters point to an early diversification into two lineages or may even suggest multiple origins of the species.

1.2 *Capsella bursa-pastoris*: The “Shepherd‘s Purse“

Five species belong to the genus *Capsella*: The three diploid (2n = 2x = 16) species *C. grandiflora*, *C. rubella* and *C. orientalis*, and the two tetraploid (2n = 4x =32) species *C. thracica* and *C. bursa-pastoris*. All species are of old-world origin and share their characteristic heart-shaped fruits, but only two species, *C. rubella* and *C. bursa-pastoris*, became successful intercontinental colonizers, whereas *C. orientalis* remained endemic to the steppe regions of Central Asia and Eastern Europe, the range of *C. grandiflora* remained limited to northern Greece and Albania, and *C. thracica* can only be found in Bulgaria (e.g. Hurka et al., 2012). The evolutionary history of this genus has been outlined recently (Hurka et al., 2012), and detailed knowledge of the colonization history and migration patterns particularly of *C. bursa-pastoris* can be detected by analysis of historical records:

North and South America: During the 16th century, *C. bursa-pastoris* was brought to the New World when the Spanish Crown conquered Middle and parts of South America. However, the invasion of Patagonia was not before 1840 as the first herbarium record from 1877 shows (Neuffer et al., 1999). During the first half of the 17th century, the Shepherd‘s Purse was introduced to parts of North America (and particularly California) with the gold seekers (Crosby, 1986; Hornbeck, 1983; Neuffer & Hurka, 1999; Neuffer & Linde, 1999).

South Africa: Thunberg reported *C. bursa-pastoris* in the Cape Colony in South Africa between 1772 and 1775 (Marais, 1970) and Sonder listed it almost 100 years later as a common weed introduced from Europe (Harvey & Sonder, 1860).

Australasia: *Capsella* populations established themselves probably in the 19th century in Australasia, when people from the British Isles and Mediterranean countries immigrated to Australia (Lamping, 1985). Although European weens already were common before 1840 (Crosby, 1986), the first *Capsella* herbarium records are from 1847 (Kloot, 1983).

The enormous expansion could be established with extraordinary ecotypic differentiation (e.g. Neuffer & Bartelheim, 1989; Neuffer, 2011; Neuffer et al., 2018), the predominantly selfing mating system, the production of thousands of seeds produced per individual (Hurka & Neuffer, 1991), the ability to survive in a soil seed bank for many years (Hurka & Haase, 1982), and the power for long distance dispersal via myxospermy (Neuffer & Linde, 1999). All these factors made the Shepherd’s Purse one of the most wide spread flowering species on earth (Coquillat, 1951; Zhou et al., 2001; Randall, 2012).

During this doctoral research project, a common garden experiment has been performed to record phenotypic data from specimens from a variety of locations from all over the world. The sampling locations cover a large area along a Eurasian west-east gradient including the origination area of the species, and an American north-south gradient, which represents the recently colonized range of this species. Together with additional populations from South Africa and Australia, this study covers almost the entire global range of this wide-spread plant (Fig. 1).

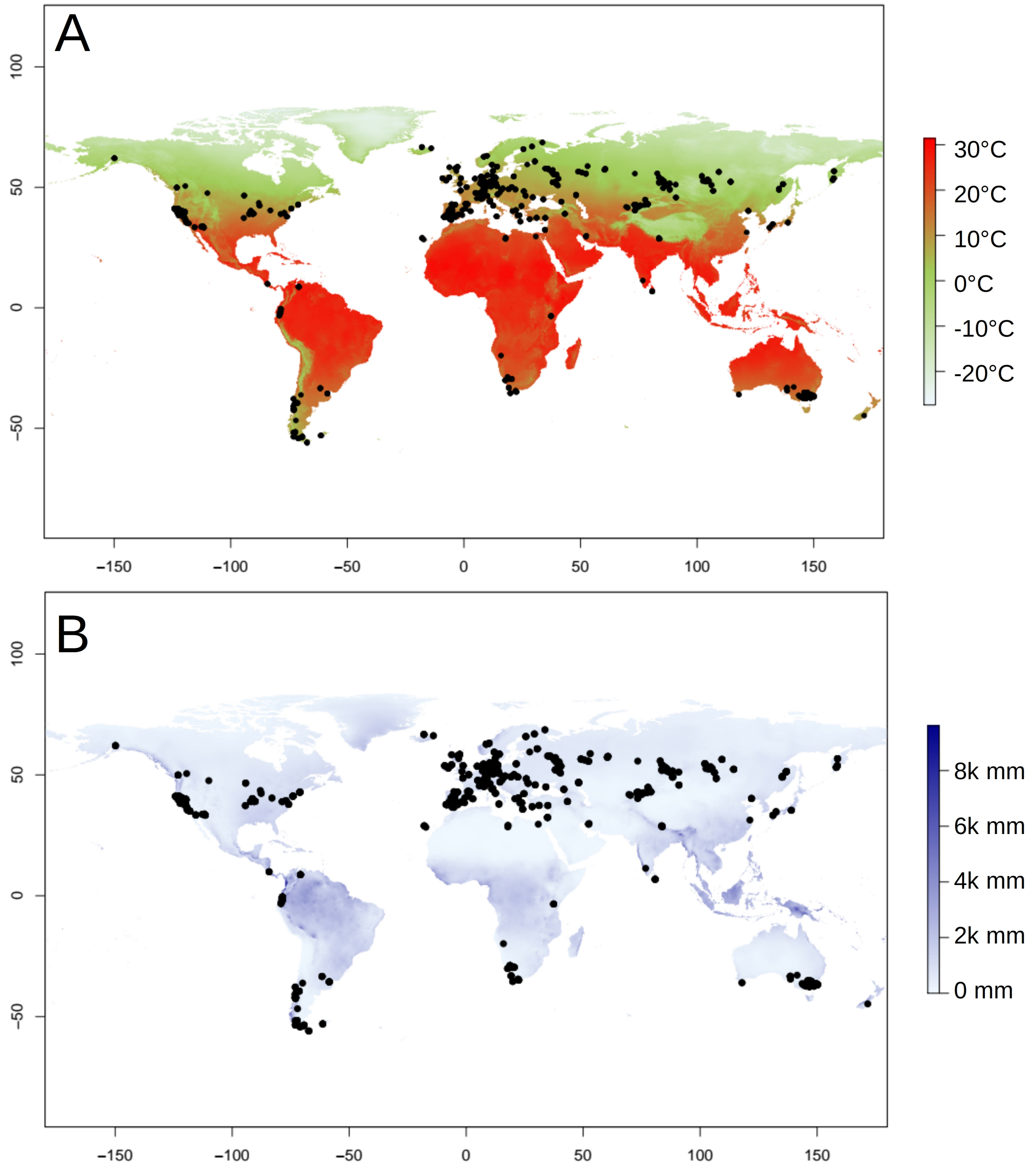

Figure 1: Sampling locations of *C. bursa-pastoris* individuals.
A: Colors refer to annual mean temperature of coordinates.
B: Colors refer to annual mean precipitation of coordinates.
Data derived from Worldclim, August 28th 2017 (http://www.worldclim.org/)

1.3 Restriction site-associated DNA Sequencing (RADseq)

Illumina sequencing belongs to the so-called massive(ly) parralel sequencing (MPS) techiques. MPS is one of several high troughput sequencing methods and is used as synonym for next-generation sequencing (NGS) in general. All these techniques have in common, that the templates – in contrast to the chain-terminating method of Sanger sequencing – are sequenced in parallel on a flow cell; often in folds of millions or billions per single sequencing run. The DNA bases are detected via image aquisition: Each template has unique coordinates on the flow cell and each fluorescent base built-in during synthesis is determined with photosensing methods ("Sequencing by synthesis", SBS).

Restriction site-associated DNA sequencing (RADseq; Miller et al., 2007) is a NGS method using Illumina technology which is often used in population genomics for single nucleotide polymorphisms (SNP) discovery and genotyping. Akin to analyses using restriction fragment length polymorphisms (RFLPs) and amplified fragment length polymorphisms (AFLPs), RADseq is a "complexity reduction" method: In contrast to whole genome re-sequencing (WGRS) methods, reduced-representation libraries sequencing (RRLS) does not investigate the whole genome but a more manageable set of sequences (Davey & Blaxter, 2010). Whereas the previous techniques RFLP and AFLP were used to detect polymorphisms within restriction cut sites, sequences adjacent to the restriction cut sites are obtained via RADseq (Baird et al., 2008). RADseq loci can occur in coding and non-coding regions (Davey et al., 2011).

RADseq genotyping detects SNPs, the most abundant type of genetic marker – this predestines them for studying the inheritance of genomic regions (Berger et al., 2001). RRLS protocols can produce up to several tens of thousands SNPs. RADseq is a comparatively low priced method for population genomic scans when a reference genome is available, at the cost proportionate up to 10 times the cost of RFLP or AFLP analyses with data outcome approx. 10,000 times bigger (Hohenlohe et al., 2010; Scaglione, 2012; de Villemereuil et al., 2016).

The RADseq procedure can roughly be summarized in the following way: During RAD library preparation, isolated DNA from individuals is digested with a chosen restriction enzyme (in this thesis *KpnI*), producing sticky-ended fragments to which adapters with molecular identifiers (MID) are ligated (Miller et al., 2007; Baird et al., 2008). The MIDs are unique "barcodes" within each sequencing pool, that enable *in silico* re-sorting of individuals when multiplexed for sequencing in parallel (Baird et al., 2008). The samples are then sheared with ultrasound and then ligated to a second adapter and amplified via polymerase chain reaction (PCR). The sequencer-ready libraries are then size selected and sequenced on the Illumina platform (Illumina Inc., United States). The resulting sequences downstream the restriction enzyme cut sites, called RAD tags, will include the

MID sequence, adapters and of course the individual‘s DNA sequence of interest neighbouring the restriction site. Illumina sequencers usually have sequence read length of 50 – 300 bp and can optionally sequence “single end“ (SE) or “paired end“ (PE). During SE-sequencing, one forward read (i.e. beginning from the restriction enzyme cut site) will be produced, whereas during PE-sequencing one forward and one reverse strand starting from the randomly sheared end will be generated. The PE-method is beneficial to reduce possible sequence errors in long reads. However, the two sequencing reads from PE can be of different size and longer than the complement read and therefore overlap the adaptor sequences, possibly causing errors within the contigs produced which makes differentiation between the samples difficult. Usually RAD tags are approximately 150 bases long (Davey & Blaxter, 2010), and since the reads are not expected to exceed this number in this study, the SE-method is sufficient and will be used. If a reference genome is available, sequence reads can be aligned to it and SNPs identified using NGS bioinformatics tools. After the genotyping is finished, the data is usually filtered to remove loci and samples with large amounts of missing data.

Both techniques, RADseq and genotyping by sequencing (GBS; Elshire et al., 2011), use restriction enzymes for cutting and adapt specific sequencing adaptors to the loci to be sequenced, therefore both terms are often used interchangeably albeit they describe particular methods in detail (Baird et al., 2008; Andrews et al., 2016). The choice of the restriction enzyme or enzymes determine the number and selection of DNA fragments to be sequenced, since restriction enzymes can be “rare cutters“ or “common cutters“ (Andrews et al., 2016): Roughly estimated, a so-called 8-cutter will cut every 4^8 = 65,536 bp, and a 6-cutter every 4^6 = 4,096 bp (Davey et al., 2011; Andrews et al., 2016), but there are also computational tools available to estimate the number of expected loci (e.g. Lepais & Weir, 2014). The optimal number depends on the particular study. Geographic population structure studies usually require several hundreds or thousands of loci, but if small populations are studied and/or bottlenecks are expetced, one should consider to increase the number of loci (Andrews et al., 2016).

After the genotypes have been obtained, many studies use computer programs like STRUCTURE or ADMIXTURE to reconstruct the demographic history of species from RAD/GBS data. Whereas STRUCTURE uses a Bayesian algorithm to define populations (Pritchard et al., 2000), ADMIXTURE was later implemented and uses an inference model (Alexander et al., 2009). Both are popular approaches using model-based genetic clustering algorithms and are commonly used alongside Principle component analysis (PCA). The main principle ist to estimate the number of *K* (i.e. clusters) and assign individuals to these defined populations (in percentage of cluster affiliation). For this analysis, it has to be assumed that the value of *K* is the true amount of ancestral groups which existed at some point in the past and that the investigated (modern) individuals were produced by mixing of these ancestral populations (Lawson et al., 2018). The results can be visualized in a barplot and biologically interpreted.

A short summary of the whole RADseq procedure can be found in table 1 and figure 2.

Table 1: Summarized procedure of the RADseq workflow.

Step		**Performance**	**Output**
Primary analysis	1. Laboratory work (Primary analysis)	Experimental design Library preparation Enrichment (Capture)	DNA
	2. NGS	e.g. Illumina	.fastq(.gz)
Secondary analysis	3. Quality assessment	Trimming, filtering Software: Trimmomatic	.fastq(.gz)
	4. Alignment to reference genome	Software: BWA	.sam/.bam
Tertiary analysis	5. Variant identification	Single nucleotide variants (SNVs), structural variants (e.g. indels) Software: SAMtools, freebayes	.vcf
	6. Population structure analyses	e.g. PCA, ADMIXTURE, GWAS Software: PLINK, SNPrelate, EMMAX	various

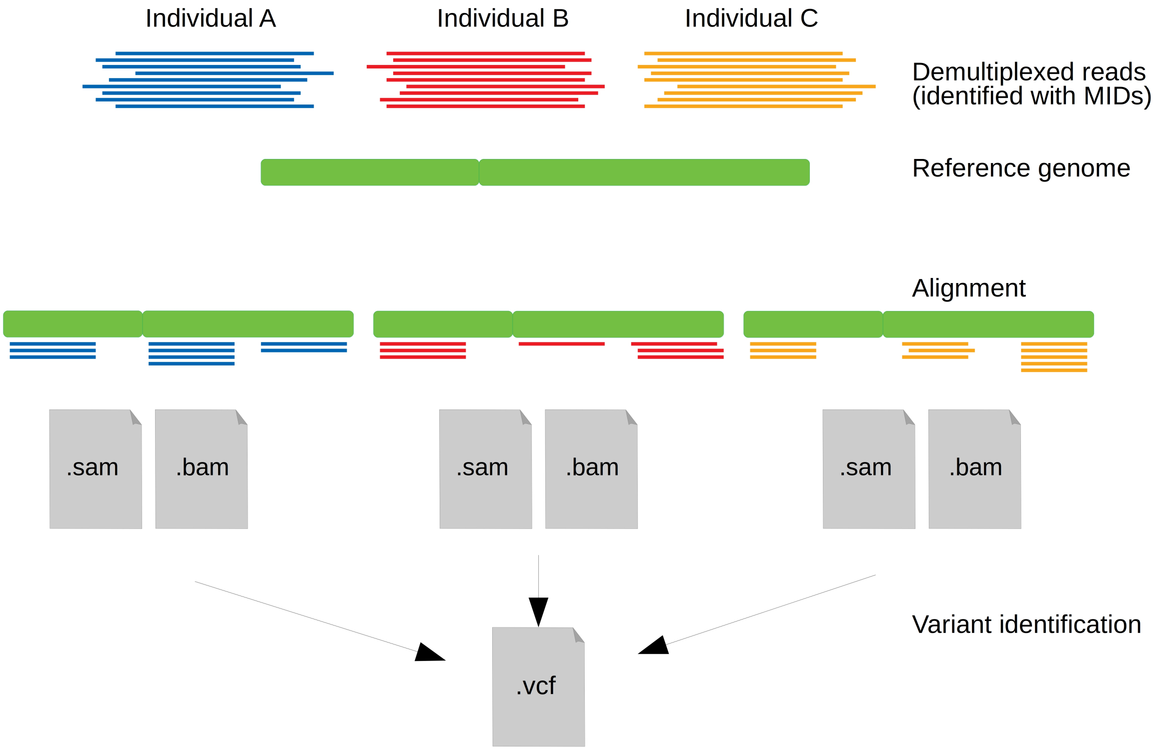

Figure 2: Visualization of the bioinformatic RADseq procedure and file formats.

1.4 Phenotyping and Genome-wide Association Studies (GWAS)

Phenotypic plasticity is the ability of organisms to express different phenotypes dependant on different environments (Bradshaw, 1965). Unsurprisingly, a high level of phenotypic plasticity is quintessential for widespread or even invasive species to adapt to a broad range of environmental conditions (Bradshaw, 1965), e.g. water-availability (regarding climate zone), UV radiation (relevant to altitude), and temperature seasonality (regarding latitude).

Common garden experiments, also known as transplant experiments, test the effect of the environment on phenotypes by growing specimens from primary diverse geographical origins in a common environment, generally under laboratory or seminatural conditions (de Villemereuil et al., 2016). In this connection it is appropriate to have a randomized controlled arrangement of the tested organisms to eliminate local or edge effects (e.g. disposability of light and nutrients). With a suitable amount of replicates, the colltected data (e.g. onset of flowering or plant height) from individuals of the same family (i.e. a group of individuals with known pedigree) can be averaged and reliably analyzed (de Villemereuil et al., 2016).

Using genome-wide association studies (GWAS) gives the possibility to combine information from phenotypic data and significant genetic variants (e.g. investigated via RADseq), as SNPs are tested for their association with each morphological trait of interest: If an adaptive signal is detected in genotypic and environmental data, as well as in the phenotypic data from a common garden experiment, certain alleles could be related to an ecological factor and local adaptation of the species is most likely (Holderegger et al., 2008). The output from GWAS analyses can be visualized in a scatter plot. This so-called “Manhattan plot“ shows the SNPs on the x-axis and the calculated association on the y-axis. The Bonferroni threshold shows the significance of detected associations. Usually, a Bonferonni corrected genome wide significance threshold of $-\log(5 \times 10^{-8})$ is used (Reed et al., 2015). This value is based on the fact that approximately one million independent SNPs occur across a genome (Reed et al., 2015). A less stringent suggestive association is a threshold of $-\log(5\times10^{-6})$ (Reed et al., 2015). Another way to confirm the association is to visualize the expected and observed values in a qq-plot. The qq-plots are used to show the relationship between expected and observed distributions of SNP level statistics. Each deviation from the ideal horizontal line (X = Y) shows a strong hint to “true“ association (λ-Statistics). In a convincing qq-plot, most of the observed values follow the horizontal line, with a few tremendous outliers showing the most probable associations. This results in the identification of molecular markers that are under natural selection.

I. The role of ecotypic variation in driving worldwide colonization by a cosmopolitan plant

Barbara Neuffer, Christina Wesse, Ingo Voss, Renate Scheibe

AoB PLANTS, Volume 10, Issue 1, 1 February 2018, ply005

https://doi.org/10.1093/aobpla/ply005

Published: 24 January 2018

Abstract

For almost 100 years now, ecotypic differentiation of plant species has been a major topic of research. In changing environments, the question needs to be answered as to how long it takes to adapt, and which parameters are subject to this fast adaptation. Short-living colonizing plant species are excellent examples, especially when they are selfing. Shepherd's Purse *Capsella bursa-pastoris* (Brassicaceae) is one of the most wide-spread flowering species on earth and avoids only the hot and humid tropics. Many studies demonstrated the ecotypic differentiation of *C. bursa-pastoris* in various regions of the world but ecotypic differentiation regarding adaptability of anatomy and physiology of rosette leaves so far remained less recognized. However, the leaves are relevant for subsequent seed set; in particular, winter-annual accessions require a robust rosette to survive adverse conditions. Leaf-related traits such as the thickness of the mesophyll and epidermis, stomatal density, photosynthetic capacity and the ability to withstand and even use high light conditions were therefore analysed in provenances from various climatic zones. Photosynthetic capacity depends on leaf anatomy and cellular physiological parameters. In particular, the ability to dynamically adjust the photosynthetic capacity to changing environmental conditions results in higher fitness. Here, we attempt to relate these results to the four Mendelian leaf types according to Shull.

Keywords: Anatomy, *Capsella*, chlorophyll fluorescence parameters, leaf types, photosynthetic capacity, Shepherd's Purse

Subject: Evolution, Ecophysiology

Issue Section: Research article

Introduction

Since Turesson’s pioneering work (1922a, b, 1930), the ecotypic differentiation of plant species has been a main interest in population biology: how and how quickly are plant species able to adapt to changing environmental conditions. Due to their property of self-fertilization, short-living colonizing plant species are suitable examples. A single seed is able to colonize a new habitat and to establish a new population. Which traits play a role in this highly successful colonizing history? For a short-living plant, several characters are critical, in particular, the physiology of germination and the determination of flowering time. In both developmental steps, the plant quits a status that is resistant to harsh conditions and changes to a highly sensitive status. If the environment is unfavourable, the individual will not complete its life cycle or it will produce only a low amount of immature seeds. Between germination and flowering, when they grow vegetatively in a rosette status, these plants are able to survive extreme conditions, e.g. cold winter. So far, numerous investigations have aimed at the control of flowering time, but less information is available concerning ecotypic differentiation with regards to a combination of morphological, anatomical and physiological leaf characters, which we are focussing on in this study.

The Shepherd’s Purse *Capsella bursa-pastoris* (Brassicaceae) belongs to the most prevalent flowering plants on earth (Coquillat 1951; Zhou *et al.* 2001), but it is not found in the hot and humid tropics. Their extraordinary colonizing success may be caused by the predominantly selfing mating system, rapid propagation by seeds as an annual to winter annual, the production of an enormous amount of seeds per individual (Hurka and Neuffer 1991), the ability to survive in a soil seed bank for many years (Hurka and Haase 1982) and the power for long-distance dispersal via myxospermy (Neuffer and Linde 1999).

In prehistoric times *C. bursa-pastoris* was distributed over the whole of Eurasia including the regions surrounding the Mediterranean Sea either along river shores or by early agricultural activities of humans. Later on, from the beginning of the 16th century, Europeans colonized all other continents, used the same agricultural techniques and crop plants and introduced many weeds as neophytes that in some cases turned out to be pests for the native biodiversity (e.g. Mooney *et al.* 2005). This unintentional transport paved the way for *C. bursa-pastoris* to reach the New World, Australia, South Africa, New Zealand, the Falkland Islands and other localities (Neuffer and Hurka 1999; Neuffer *et al.* 1999, 2011; Kryvokhyzha *et al.* 2016). This fast expansion of one weedy plant species was only possible due to its extraordinary capability of ecotypic differentiation. The differentiation of *C. bursa-pastoris* has been recorded for Europe (e.g. Neuffer and Bartelheim 1989, reviewed in Neuffer 2011), and such pre-adapted ecotypes have been able to find their niche elsewhere on the globe (Neuffer and Hurka 1999; Neuffer *et al.* 1999).

In many studies, the ecotypic differentiation of *C. bursa-pastoris* has been demonstrated for various regions of the world predominantly regarding germination, flowering time, rosette diameter and number of inflorescence branches (reviewed in Neuffer *et al.* 2011). Adaptive traits are frequently related to the rosette leaves that are responsible for the production of resources required for subsequent yield and abundant seed set. In the case of winter annuals, a robust rosette is required to survive suboptimal weather conditions. Important leaf traits are rosette diameter, the number of leaves in a rosette, leaf area, thickness of the leaf as well as of the epidermal cells, stomatal density, and photosynthetic capacity, as well as photosynthetic light utilization. Stomatal density and other epidermal characteristics strongly influence water-use efficiency (WUE), which is particularly important in dry habitats accompanied by high irradiation (reviewed in Körner 2003). The photosynthetic capacity depends on number, total area and anatomy of leaves and on cellular physiological parameters; in particular, the ability to dynamically adjust photosynthetic capacity to changing environmental conditions results in higher fitness (Athanasiou *et al.* 2010). The climatic adaptability and ecotypic intraspecies differentiation in a combination of morphological, anatomical and physiological characters have been shown for *Diplotaxis erucoides* populations from Sicily by Schleser *et al.* (1989).

The degree of leaf-margin dissection – from entire leaves with smooth margins to serrated and increasingly deeply lobed leaves – is likely to be functionally important. Leaf-margin dissection shows a very robust negative correlation with mean annual temperature, both at the within-species and at the community level. Such a wide-spread relationship between a morphological trait and an environmental parameter across different phylogenetic scales provides a strong argument that the trait in question is adaptive (Nicotra *et al.* 2011). Recently, the analysis of leaf size and compound leaves of a large number of species in relation to geography and climate was analysed by Wright *et al.* (2017) which is an indicator for global climatic change. The genus *Capsella* shows a high level of variation in all of the above-mentioned leaf-related traits, including the leaf shape, which ranges from entire leaves to very deeply dissected ones (Fig. I.1; Shull 1909; Sicard *et al.* 2014).

Here, we will shed light on a combination of morphological/anatomical and physiological variation that underlies rapid local adaptation in one of the world's most successful weeds.

Methods

For each of the four datasets different treatments and determinations were performed as described below. It was our aim to analyse all leaves at the same stage of development. Therefore, we started the leaf studies when the first flower bud appeared indicating the end of the vegetative phase. In the case of late flowering plants, we started no later than 3 months after sowing to avoid any senescence.

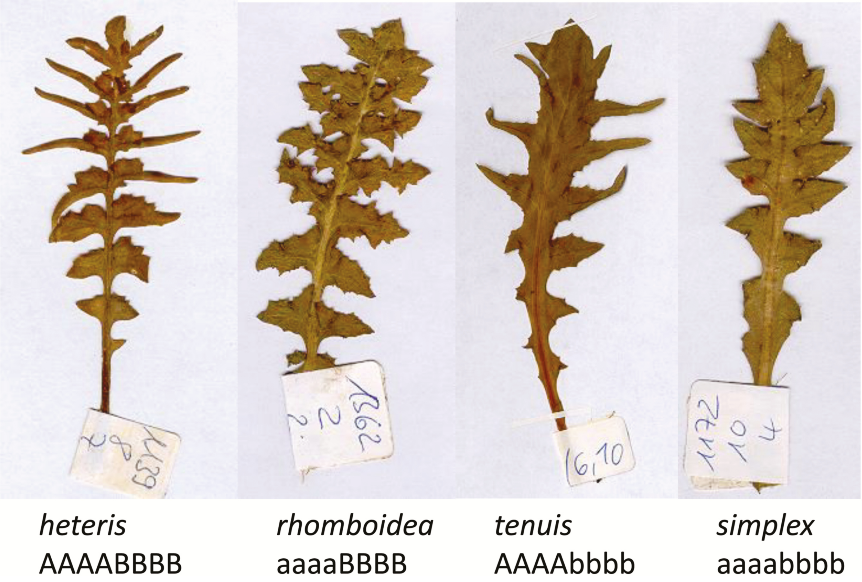

Figure I.1. Rosette leaf types and allele formulas of tetraploid *Capsella bursa-pastoris* after Shull.

Analysis 1: isotope analysis

To test the ability of progenies from different environmental habitats to react on drought stress we grew sister individuals under different conditions and analysed $\delta^{13}C$ values in combination with morphological and anatomical features (Fig. I.2, green colour; Table I.1). Plants were grown in a growth chamber with a daynight rhythm of 15:9 h with 15–25 °C. Progeny of individuals collected in the wild (Fig. I.2) were divided into two groups, one group was kept under water stress conditions with a maximum of 10 ml water daily per 1-L pot, the other group with at least 30 ml daily, for non-stress conditions. Each individual was planted in 1-L substrate with sand and slightly fertilized turf in a proportion of 1:2. Each population and each condition were represented by up to five sister individuals. For each individual (three individuals for each population and treatment), cell size, stomata density and the percentage of the volume of mesophyll cells compared with intercellular space were measured 30 times.

Furthermore, the $\delta^{13}C$ values were determined as follows: dried leaf material was combusted in an excess of oxygen at ~1000 °C, and the resulting CO_2 used for isotope analysis using the MAT 250 mass spectrometer (Schleser and Poling 1980). Carbon isotope values of leaves are based on total organic matter rather than on a selected chemical compound such as cellulose. Several test measurements have shown that relative variations of the isotope content led to similar results for total organic matter and cellulose. Only the absolute values differ by 2 to 2.5°/$_{00}$. Results are reported in terms of $\delta^{13}C$ relative to PDB (Belemnite from the Pee Dee formation in South Carolina; Craig 1957). The $\delta^{13}C$ value as the ratio between $^{13}CO_2$ and $^{12}CO_2$ was measured twice for each individual.

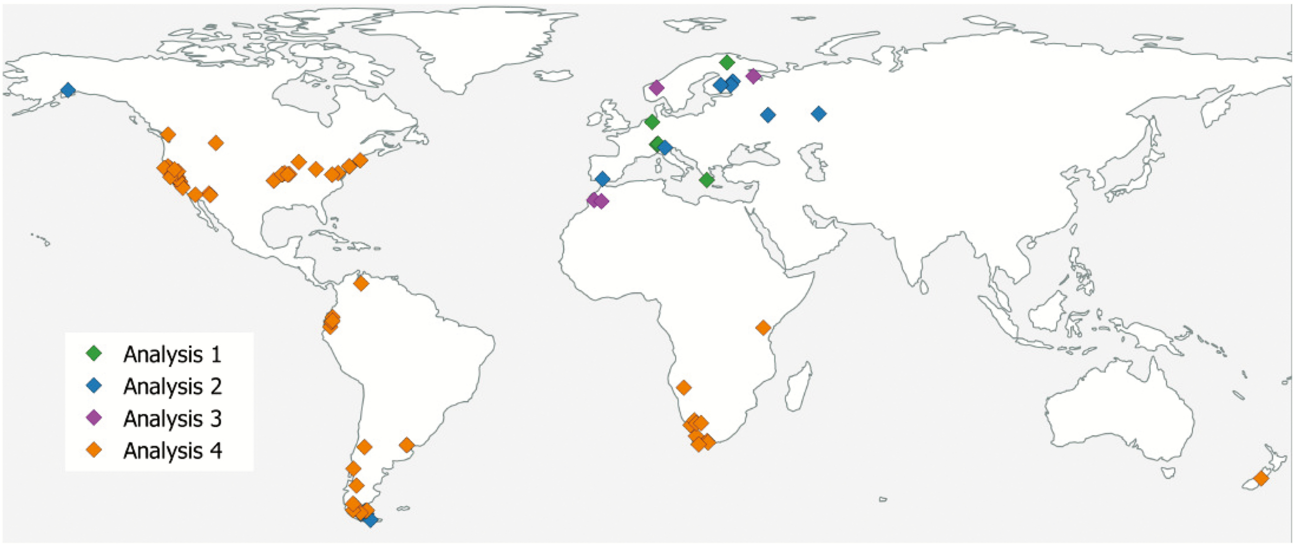

Figure I.2. Localities of provenances used in four different experimental designs (Analyses 1–4). Analysis 1: six populations (green colour); Analysis 2: 15 populations (blue colour), Analysis 3: four populations (purple colour); Analysis 4: 76 populations (orange colour).

Analysis 2: chlorophyll fluorescence and CO_2 gas exchange analysis

To test whether progenies from various environmental habitats differ in their ability for their photosynthetic activity, we analysed chlorophyll fluorescence and CO_2 gas exchange in combination with morphological and anatomical features (Fig. I.2, blue colour; Table 1). Of each accession (see Fig. I.2) three individuals were grown in a growth chamber with 12-h photoperiod and 15 °C day and 5 °C night temperature. For each individual, the thickness of the leaf and the epidermis cells, stomata density and leaf area were measured.

Using a FluorCam 800MF (Photon Instruments, Brno, Czech Republic), we determined different chlorophyll fluorescence emission parameters of the whole rosette: F_m: maximum fluorescence emission in light; F_0: ground fluorescence in light; F: fluorescence emission in light after light pulses of 6000 µmol m^{-2} s^{-1} (for 800 ms every 30 s). From these measured parameters, the photosynthetic light utilization was estimated (Schreiber et al. 1986; Genty et al. 1989; Scheibe et al. 2005; Hanke et al. 2009; Scheibe and Dietz 2012; Silva et al. 2012; Voss et al. 2013).

Furthermore, the CO_2 gas exchange was measured using the Lic400XT Portable Photosynthesis System (Li-Cor Biosciences, Lincoln, NE, USA). A/C_i curves enabled us to calculate the efficiency of RubisCO to fix CO_2 under limiting conditions. To determine the photosynthetic capacity of the secondary reaction in relation to the specific light intensity, light saturation curves were recorded. From the obtained parameter values the quantum yield for CO_2 uptake and the light compensation point (LKP) could be calculated.

Table I.1. Provenances of all studied *Capsella* populations; country labelled by KFZ.

Population	Country	Locality	Latitude	Longitude	Elevation (m)	Species	Leaf type after Shull	Collector	Analysis
83	D	Teglingen	52.65	7.35	14	*C. bursa-pastoris*	*rhomboidea*	Benneweg	1
147	FIN	Ivalo	68.65	27.57	160	*C. bursa-pastoris*	*rhomboidea*	Bosbach, K., Hurka, H.	1
257	CH	Disentis	46.68	8.83	1400	*C. bursa-pastoris*	*rhomboidea*	Hurka, H.	1
279	CH	Andermatt	46.63	8.6	1480	*C. bursa-pastoris*	*tenuis*	Hurka, H.	1
282	CH	Trun	46.75	8.98	850	*C. bursa-pastoris*	*heteris*	Hurka, H.	1
434	GR	Kalamata	37.04	22.12	1270	*C. rubella*	*simplex*	Bosbach, K., Hurka, H.	1
679	USA	Neosho	36.87	−94.37	323	*C. bursa-pastoris*	*sim/rho*	Borgwart, M.	4
680	USA	Neosho	36.87	−94.37	323	*C. bursa-pastoris*	*tenuis*	Borgwart, M.	4
681	USA	Chicago	41.88	−87.63	182	*C. bursa-pastoris*	*heteris*	Borgwart, M.	4
700	USA	Davis	38.53	−121.73	16	*C. bursa-pastoris*	*heteris*	Hurka, H.	4
701	USA	Davis	38.53	−121.73	16	*C. bursa-pastoris*	*sim/rho*	Hurka, H.	4
702	USA	Davis	38.53	−121.73	16	*C. bursa-pastoris*	*tenuis*	Hurka, H.	4
703	USA	Davis	38.53	−121.73	16	*C. bursa-pastoris*	*rhomboidea*	Hurka, H.	4
706	USA	Davis	38.53	−121.73	16	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
712	USA	Williams	39.15	−122.15	25	*C. bursa-pastoris*	*sim/rho*	Hurka, H.	4
713	USA	Stockton	37.95	−121.28	5	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
714	USA	Stockton	37.95	−121.28	5	*C. bursa-pastoris*	*tenuis*	Hurka, H.	4
715	USA	Coulterville	37.72	−120.2	544	*C. bursa-pastoris*	*rhomboidea*	Hurka, H.	4
717	USA	Fresno	36.75	−119.77	98	*C. bursa-pastoris*	*tenuis*	Hurka, H.	4
718	USA	Fresno	36.57	−119.62	98	*C. bursa-pastoris*	*sim/ten*	Hurka, H.	4
722	USA	Shafter	35.5	−119.27	106	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
723	USA	Wheeler Ridge	34.98	−118.93	111	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
726	USA	Tuttle	37.3	−120.38	62	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
727	USA	Willows	39.52	−122.3	67	*C. bursa-pastoris*	*heteris*	Hurka, H.	4
728	USA	Willows	39.52	−122.2	43	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
729	USA	Chico	39.78	−121.95	59	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
730	USA	Red Bluff	40.15	−122.25	103	*C. bursa-pastoris*	*heteris*	Hurka, H.	4
732	USA	Douglas City	40.65	−122.93	609	*C. bursa-pastoris*	*ten/rho*	Hurka, H.	4
733	USA	Weaverville	40.73	−122.93	636	*C. bursa-pastoris*	*ten/het*	Hurka, H.	4
736	USA	Myers Flat	40.27	−123.87	85	*C. bursa-pastoris*	*sim/het*	Hurka, H.	4
745	USA	Placerville	38.73	−120.67	610	*C. bursa-pastoris*	*rhomboidea*	Hurka, H.	4
746	USA	Davis	38.53	−121.73	16	*C. bursa-pastoris*	*sim/rho*	Hurka, H.	4
747	USA	Truckee	39.33	−120.18	1819	*C. bursa-pastoris*	*sim/ten*	Hurka, H.	4
748	USA	Berkeley	37.87	−122.25	112	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
750	USA	Bucks Lake	39.87	−121.17	1582	*C. bursa-pastoris*	*sim/het*	Hurka, H.	4

Table I.1, *continued (2/3)*

Population	Country	Locality	Latitude	Longitude	Elevation (m)	Species	Leaf type after Shull	Collector	Analysis
785	USA	Jefferson City	38.52	−92.07	207	*C. bursa-pastoris*	*sim/ten*	Koch	4
786	USA	Montgomery City	38.88	−91.45	266	*C. bursa-pastoris*	*ten/het*	Koch	4
846	USA	St Louis	38.63	−90.18	141	*C. bursa-pastoris*	*tenuis*	Neuffer, B.	4
847	USA	Jefferson City	38.57	−92.17	194	*C. bursa-pastoris*	*simplex*	Neuffer, B.	4
848	USA	St Louis	38.5	−90.63	191	*C. bursa-pastoris*	*tenuis*	Neuffer, B.	4
852	USA	Boston	42.35	−71.07	5	*C. bursa-pastoris*	*rho/het*	Neuffer, B.	4
853	USA	Boston	42.35	−71.07	5	*C. bursa-pastoris*	*sim/het*	Neuffer, B.	4
855	USA	Columbus	39.95	−83	230	*C. bursa-pastoris*	*simplex*	Crawford, D.J.	4
939	YV	Pico el Aguila	8.85	−70.82	3877	*C. bursa-pastoris*	*simplex*	Bosbach, K.	4
961	NAM	Etosha National Park	−19.17	15.92	1178	*C. bursa-pastoris*	*simplex*	Schröpfer, R.	4
966	EAT	Mt. Kilimanjaro Nat. Park	−3.07	37.37	5325	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
1137	FIN	Nurmes	63.55	29.12	120	*C. bursa-pastoris*	*rhomboidea*	Neuffer, B.	2
1139	FIN	Kuopio	62.58	28.59	95	*C. bursa-pastoris*	*heteris*	Neuffer, B.	2
1141	FIN	Suolahti	62.57	25.85	100	*C. bursa-pastoris*	*heteris*	Neuffer, B.	2
1141	FIN	Suolahti	62.57	25.85	100	*C. bursa-pastoris*	*tenuis*	Neuffer, B.	2
1198	RCH	Puerto Octay	−41	−72.88	153	*C. bursa-pastoris*	*sim/ten*	Hurka, H.	4
1273	E	Pilas	37.3	−5.7	80	*C. bursa-pastoris*	*simplex*	Neuffer, B.	2
1355	I	Malcesine	45.77	10.82	800	*C. bursa-pastoris*	*simplex*	Neuffer, B.	2
1357	USA	Anchorage	61.22	−149.88	20	*C. bursa-pastoris*	*simplex*	Handke	2
1376	RA	Buenos Aires	−34.67	−58.5	19	*C. bursa-pastoris*	*tenuis*	Damborenea, S.	4
1377	RA	Buenos Aires	−34.67	−58.5	10	*C. rubella*	*heteris*	Damborenea, S.	2
1377	RA	Buenos Aires	−34.67	−58.5	19	*C. bursa-pastoris*	*rhomboidea*	Damborenea, S.	4
1380	RCH	Punta Delgada	−52.45	−69.55	50	*C. bursa-pastoris*	*heteris*	Neuffer & Neuffer	4
1381	RCH	San Sebastian	−53.15	−69.4	100	*C. bursa-pastoris*	*heteris*	Neuffer & Neuffer	2
1385	RA	Ushuaia	−54.8	−68.3	20	*C. bursa-pastoris*	*tenuis*	Neuffer & Neuffer	2
1387	RCH	Punta Delgada	−52.22	−69.28	200	*C. bursa-pastoris*	*ten/rho*	Neuffer & Neuffer	4
1388	RCH	Porto Gregorio	−52.32	−69.74	10	*C. bursa-pastoris*	*heteris*	Neuffer & Neuffer	2
1389	RCH	Punta Arenas	−52.9	−70.97	34	*C. bursa-pastoris*	*heteris*	Neuffer & Neuffer	4
1390	RCH	Tehuelche	−53.15	−70.89	200	*C. bursa-pastoris*	*tenuis*	Neuffer & Neuffer	2
1393	RCH	Nationalpark Torres del Paine	−50.72	−72.7	578	*C. bursa-pastoris*	*rho/het*	Neuffer & Neuffer	4
1394	RCH	Nationalpark Torres del Paine	−52.18	−73	46	*C. bursa-pastoris*	*heteris*	Neuffer & Neuffer	4
1397	RA	Perito Moreno	−50.47	−73	500	*C. bursa-pastoris*	*ten/rho*	Neuffer & Neuffer	4
1412	RA	Las Lenas	−35.18	−69.9	2232	*C. bursa-pastoris*	*rho/het*	Hilger, H.	4
1461	N	Lom	61.83	8.55	400	*C. bursa-pastoris*	*Not scored*	Neuffer, B.	3
1475	RUS	Almejewsk	54.87	52.3	130	*C. bursa-pastoris*	*rhomboidea*	Neuffer, B.	2

Table I.1, *continued (3/3)*

Population	Country	Locality	Latitude	Longitude	Elevation (m)	Species	Leaf type after Shull	Collector	Analysis
1481	USA	Hobson	47	-110	1306	*C. bursa-pastoris*	*rho/het*	Hellwig, F.	4
1513	USA	Washington	38.9	-77.02	14	*C. bursa-pastoris*	*rhomboidea*	Desmarowitz, C.	4
1514	USA	Shenandoah	38.48	−78.62	315	*C. bursa-pastoris*	*simplex*	Desmarowitz, C.	4
1515	USA	Shenandoah	38.48	−78.62	315	*C. bursa-pastoris*	*simplex*	Desmarowitz, C.	4
1517	USA	New York	40.72	−74.02	0	*C. bursa-pastoris*	*sim/ten*	Desmarowitz, C.	4
1518	USA	New York	40.72	−74.02	0	*C. bursa-pastoris*	*tenuis*	Desmarowitz, C.	4
1519	USA	New York	40.72	−74.02	0	*C. bursa-pastoris*	*tenuis*	Desmarowitz, C.	4
1520	USA	New York	40.72	−74.02	0	*C. bursa-pastoris*	*heteris*	Desmarowitz, C.	4
1530	RUS	Kem	64.97	34.65	10	*C. bursa-pastoris*	*Not scored*	Hurka, Linde, Neuffer	3
1570	RUS	Uzunovo	54.53	38.62	150	*C. bursa-pastoris*	*rhomboidea*	Hurka, Neuffer, Pollmann	2
1570	RUS	Uzunovo	54.53	38.62	150	*C. bursa-pastoris*	*heteris*	Hurka, Neuffer, Pollmann	2
1581	EC	Quito	−0.22	−78.5	2850	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
1583	EC	Cuenca	−2.83	−79.15	3100	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
1584	EC	Provinz Chimborazo	−1.53	−78.8	3800	*C. bursa-pastoris*	*sim/het*	Hurka, H.	4
1586	EC	Pillaro	−1.17	−78.53	2843	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
1622	NZ	Double Hill	−43.62	171.63	433	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
1643	ZA	Clanwilliam	−32.22	19.2	509	*C. bursa-pastoris*	*sim/het*	Neuffer, B.	4
1648	ZA	Richtersveld	−29.25	17.73	358	*C. bursa-pastoris*	*simplex*	Neuffer, B.	4
1650	ZA	Goageb	−28.02	18.75	932	*C. bursa-pastoris*	*simplex*	Neuffer, B.	4
1652	ZA	Seeheim	−28.75	19.3	932	*C. bursa-pastoris*	*simplex*	Neuffer, B.	4
1655	ZA	Pofadder	−28.75	20.55	995	*C. bursa-pastoris*	*simplex*	Neuffer, B.	4
1668	ZA	Bergwater	−33.58	22.2	1176	*C. bursa-pastoris*	*heteris*	Neuffer, B.	4
1678	ZA	George	−33.95	22.45	223	*C. bursa-pastoris*	*simplex*	Neuffer, B.	4
1682	ZA	Caledon	−34.47	19.9	242	*C. bursa-pastoris*	*simplex*	Neuffer, B.	4
1759	USA	Phoenix	33.4	−111.83	374	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
1763	USA	Coolidge	32.98	−111.53	434	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
1764	USA	Rockwood	33.07	−115.52	−55	*C. bursa-pastoris*	*simplex*	Hurka, H.	4
1996	RCH	Coyhaique	−45.57	−72.07	15	*C. bursa-pastoris*	*heteris*	Klotz, St.	4
2030	CDN	Vancouver	49.27	−122.88	9	*C. bursa-pastoris*	*tenuis*	Hameister, S.	4
2069	MA	Marrakesch	31.63	−7.98	470	*C. bursa-pastoris*	*Not scored*	Hurka, H.	3
2072	MA	Boulmane	31.32	−6	1800	*C. bursa-pastoris*	*Not scored*	Hurka, H.	3

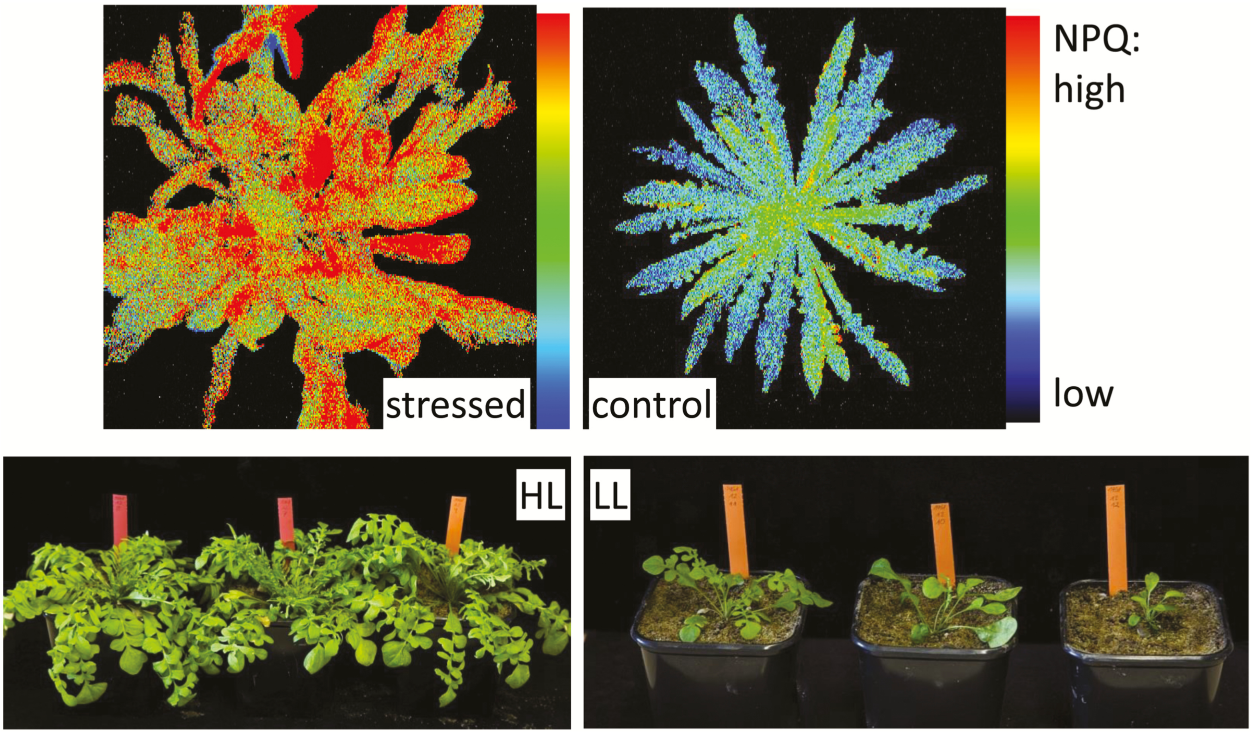

Figure I.3. Upper part: non-photochemical quenching (NPQ) of chlorophyll fluorescence as an indicator of energy dissipation in non-stressed (right) and stressed (left) conditions monitored with a FluorCam. Blue: low NPQ, indicating a low proportion of thermal dissipation. Red: high NPQ, indicating a higher proportion of thermal dissipation due to stress. Lower part: individuals of the same age grown under high light (HL, left) and low light (LL, right) conditions, respectively.

Analysis 3: CO_2 gas exchange analysis under different light stress conditions

In order to test the ability of progenies from different environmental habitats to light stress, we cultivated sister individuals under different conditions and analysed CO_2 gas exchange in combination with morphological and anatomical features (Fig. I.2, purple colour; Table I.1).

We used material from two very different vegetation zones, namely the boreal (1461, 1530) and the meridional (2069, 2072) climatic region to carve out the ecotypic adaptation of the leaves to different environmental conditions (Fig. I.2). For each population, up to 49 individuals of the progeny of two individual plants collected in the wild were used. The material was sown in a growth chamber with 12-h photoperiod and 15 °C day and 5 °C night temperature. The material was then divided into four experimental groups: 7.5-h photoperiod, 20 °C, high light setting (800 µmol m^{-2} s^{-1}, Fig. I.3, left); 7.5-h photoperiod, 20 °C, low light setting (100 µmol m^{-2} s^{-1}, Fig. I.3, right); 12-h photoperiod, 20 °C, medium light setting (150 µmol m^{-2} s^{-1}); 12-h photoperiod, 15 °C day and 5 °C night (‘cold’), medium light setting (100 µmol m^{-2} s^{-1}).

The anatomical and physiological analysis was performed as in Analysis 2.

Analysis 4: thickness of the leaf in a New World transect

As the thickness of the leaf is not only a general character in adaptation to sunny or shady orientation of leaves but seems also to be a character for ecotypic differentiation within Shepherd's Purse, we performed a large New World transect including populations from South Africa (Fig. I.2, orange colour; Table I.1). The individuals were grown in a common garden field experiment and planted randomly in the Botanical Garden in Osnabrück (Germany, May to July 2015). For anatomical analysis, material was taken directly from the field and stored in 70 % alcohol. After 1 day in tap water, the leaves became sufficiently soft for anatomical cuttings. The thickness of five rosette leaves as well as their upper and lower epidermis was determined for the terminal lobe of the leaf and for one lateral leaflet in one to two individuals of each population (Fig. I.1). We decided to study different positions of the leaf as the information might differ; also the leaflets may differ between the leaf types (Fig. I.1).

Statistical data evaluation

The data have been analysed statistically with the SPSS software package version 23. To test the normal distribution, we used the Kolmogorov–Smirnov test. In cases where data were significantly not normally distributed, we used the Spearman correlation for the correlation analyses; rho-value and significance are included in the figures. Only significantly correlated data are shown in the figures. The correlations are based on individual data. For testing significant difference (i) between treatments within a population or region, (ii) between populations within one treatment, (iii) between leaf types, we performed parameter-free Wilcoxon-test or the H-test of Kruskal and Wallis (Table I.2).

As in Analysis 4, when studying the leaf type of a progeny of 76 populations from various vegetation zones, we performed a post hoc Duncan test and an ANOVA.

Results

In this study, we correlate the results from the four described analyses with the different trait categories anatomy, physiology and morphology. A caveat in these analyses is, however, that we cannot ascertain the similarities of developmental maturation between the leaves analysed. Some of the differences observed, may, therefore, reflect intraspecific variation in life history.

Anatomical analyses

In Analysis 1, progeny from wild populations, when grown under water stress, developed denser mesophyll cells compared with the loose texture and large intercellular spaces in unstressed plants

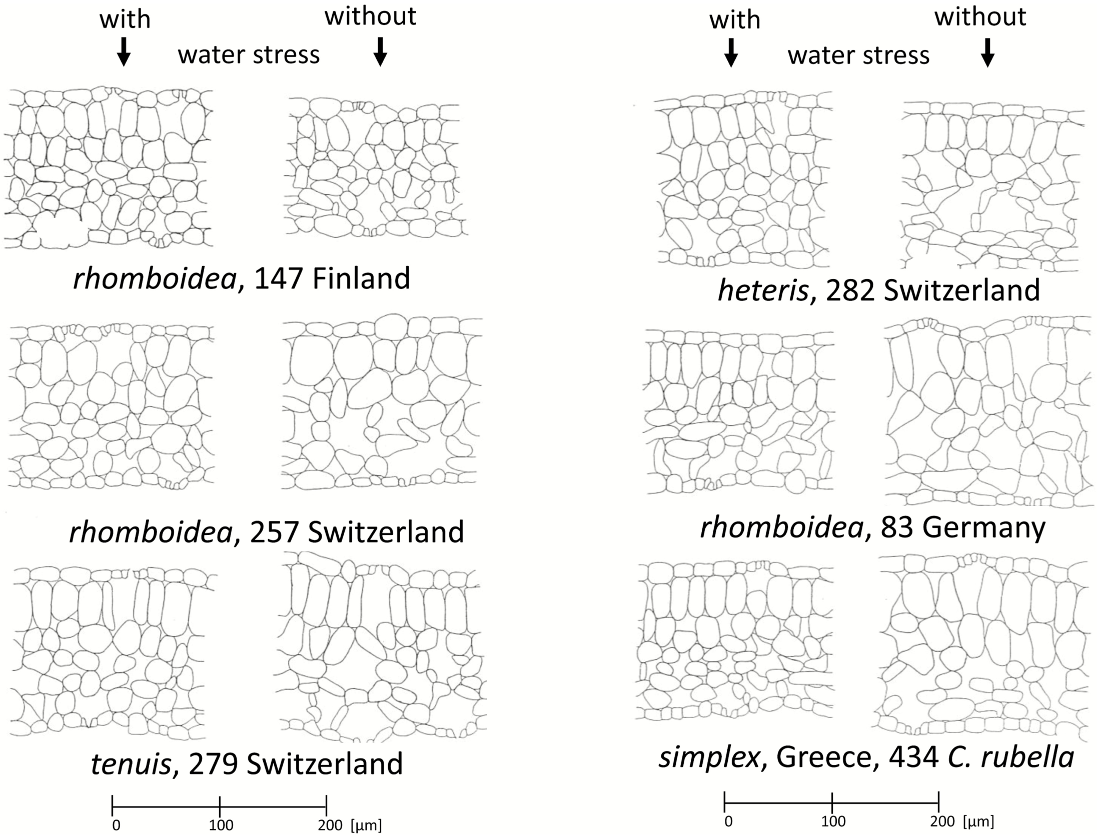
with
without
water stress
with
without
water stress
rhomboidea, 147 Finland
heteris, 282 Switzerland
rhomboidea, 257 Switzerland
rhomboidea, 83 Germany
tenuis, 279 Switzerland
simplex, Greece, 434 *C. rubella*
0
100
200 [µm]
0
100
200 [µm]

Figure I.4. Cross sections of leaves from plants grown under different water stress conditions (Analysis 1). Correlation analyses are given in Figs I.5A, I.8 and I.10.

(Fig. I.4). Furthermore, the palisade cells appeared narrower with a smaller diameter. *Capsella rubella* developed two palisade layers under both water-stressed and control conditions (Pop. 434), whereas *C. bursa-pastoris* exhibits two layers only under water stress (e.g. Pop. 147, 282). Pop. 257 showed only one palisade layer in both conditions.

Anatomical leaf parameters of various provenances grown under different conditions were correlated with geographical/elevational parameters at the places of origin (Fig. I.5): the whole leaves and, in particular, the epidermis cell layer became significantly thinner with a higher degree of latitude (Fig. I.5, all analyses, Fig. I.6). In Analysis 4, we differentiated between the terminal leaflet (Fig. I.5D1) and the lateral leaflet (Fig. I.5D2) and observed at both positions that the thickness was the same. Interestingly, with a higher elevation at the place of origin leaves became thicker in Analysis 2 (Fig. I.5C).

Even the epidermis layer itself varied with the degree of latitude and became thicker for populations originating from locations closer to the equator (Fig. I.6). Stomata became less dense when populations originated from northern latitudes (Fig. I.7). Cell sizes appear to decrease with the degree of latitude (Fig. I.8, left) and increase with higher elevation (Fig. I.8, right). However, only when grown under water stress conditions was the correlation highly significant.

Table I.2. Non-parametric tests for significant differences (Wilcoxon-test, H-test of Kruskal and Wallis). Probability values: P < 0.05: significant differences (light grey); P < 0.001 highly significant differences (dark grey).

Anatomy – low light versus high light Fig. I.9 above (Wilcoxon)

	Stomata below	Stomata above	Epidermis below	Epidermis above	Leaf Thickness
Norway	0.002	0.002	0.028	0.027	0.028
Russia	0.002	0.018	0.028	0.027	0.028
Morocco	0.000	0.000	0.012	0.027	0.028
Low Light	0.000	0.000	0.012	0.012	0.012
High Light	0.000	0.000	0.002	0.005	0.005

Anatomy – warm versus cold Fig. I.9 below (Wilcoxon)

	Stomata below	Stomata above	Epidermis below	Epidermis above	Leaf Thickness
Norway	0.002	0.002	0.008	0.007	0.008
Russia	0.003	0.003	0.017	0.018	0.018
Morocco	0.000	0.000	0.003	0.005	0.003
Low Light	0.000	0.000	0.001	0.001	0.001
High Light	0.000	0.000	0.001	0.002	0.001

Physiology – low light versus high light Fig. I.12 above (Wilcoxon)

	F_v/F_m	NPQ	qP
Norway	0.001	0.001	0.001
Russia	0.008	0.008	0.008
Morocco	0.000	0.000	0.000
Low Light	0.000	0.627	0.000
High Light	0.000	0.107	0.000

Physiology – warm versus cold Fig. I.12 below (Wilcoxon)

	F_v/F_m	NPQ	qP
Norway	0.001	0.007	0.005
Russia	0.008	0.225	0.043
Morocco	0.000	0.005	0.000
Low Light	0.000	0.945	0.000
High Light	0.000	0.152	0.001

Table I.2, *continued (2/2)*

Anatomy and physiology Figs I.5A, I.8 and I.10 (Wilcoxon)

Population	**Cell/mesophyll**	**Cell number**	**Cell size**	**Stomata density**	**Leaf thickness**	**$\delta^{13}C$**
83	0.028	0.027	0.028	0.028	0.028	0.005
147	0.042	0.042	0.043	0.043	0.042	0.008
257	0.042	0.043	0.043	0.043	0.043	0.005
279	0.027	0.027	0.028	0.028	0.028	0.005
282	0.027	0.027	0.027	0.028	0.027	0.005
434	0.027	0.028	0.028	0.043	0.028	0.008

Fig. I.5A H-test of Kruskal and Wallis

No water stress	0.061	0.034	0.019	0.031	0.043	0.041
Water stress	0.024	0.020	0.015	0.036	0.120	0.000

Leaf morphology Figs I.5B, C, I.7 and I.11 (H-test of Kruskal and Wallis)

Leaf area	**Stomata below**	**Stomata above**	**Leaf thickness**	**Latitude**	**Elevation**	**A_{max}**
0.000	0.037	0.006	0.015	0.014	0.000	0.204

Leaf morphology Figs I.5D and I.6 (H-test of Kruskal and Wallis)

Leaf terminal	**Leaf lateral**	**Epid term above**	**Epid lat above**	**Epid term below**	**Epid lat below**	**Latitude**
0.001	0.000	0.101	0.000	0.009	0.024	0.000

Leaf morphology Figs I.8 and I.10 (H-test of Kruskal and Wallis)

	Cell/mesophyll	**Cell number**	**Cell size**	**Stomata density**	**Leaf thickness**	**$\delta^{13}C$**
No water stress	0.020	0.357	0.477	0.026	0.336	0.009
Water stress	0.034	0.023	0.014	0.200	0.133	0.000

Comparing populations originating from very divergent local conditions: with dry and hot conditions in the summer and high irradiation in Morocco, and temperate humid conditions and very long days in summer in Norway and in Karelia (Russia) (Fig. I.9), the population from Morocco possessed thicker leaves and larger upper epidermis cells under all conditions, compared to the other populations. Whereas the Russian population showed a low variation of stomata density when grown under low or high light conditions and between cold versus warm temperature, respectively, the populations from Norway and Morocco increased stomata density when grown under high light and in cold conditions.

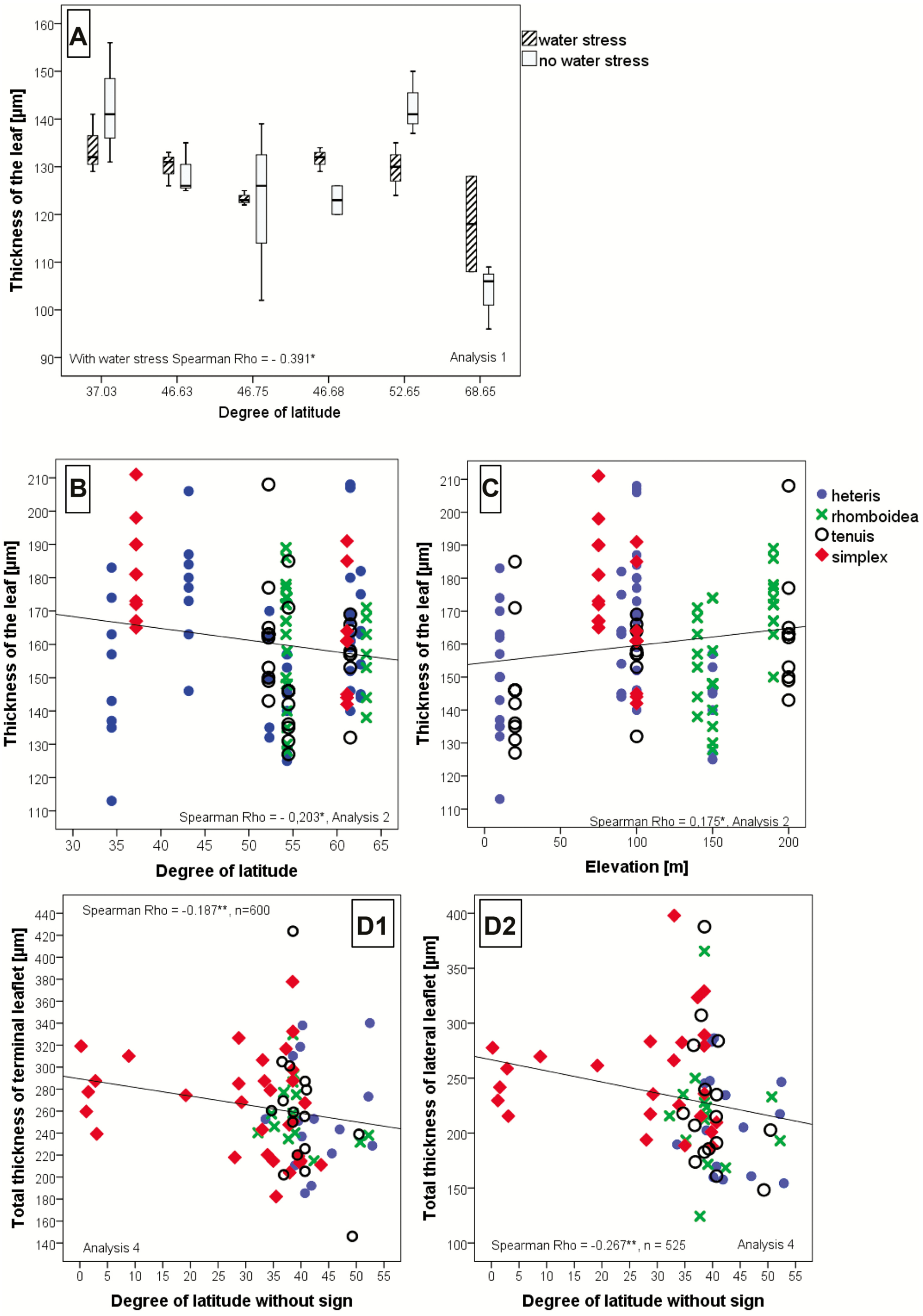

Figure I.5. Spearman correlation of leaf thickness with latitude/elevation. The thickness of the leaf is significantly negatively correlated with latitude and in one case positively with the elevation; (A) = 25:15 °C, 9-h photoperiod, populations arranged according to the degree of latitude, three individuals have been tested for each population and treatment, both treatment groups differed significantly or highly significantly when tested by H-test of Kruskal and Wallis (see Table I.2); (B and C) = 15:5 °C; 12-h photoperiod; (D)=common garden field experiment (n = number of individuals, in the figure are shown mean square values of the populations). Rho-value and significance (*α < 0.05, **α <0.01) are included in the figures. A pairwise correlation test between the terminal and the lateral leaflet thickness (Kendall-, Friedmanand Wilcoxon-test) evidenced for significant differences between both. Only significantly correlated data are shown in the figures. The correlations are based on individual data.

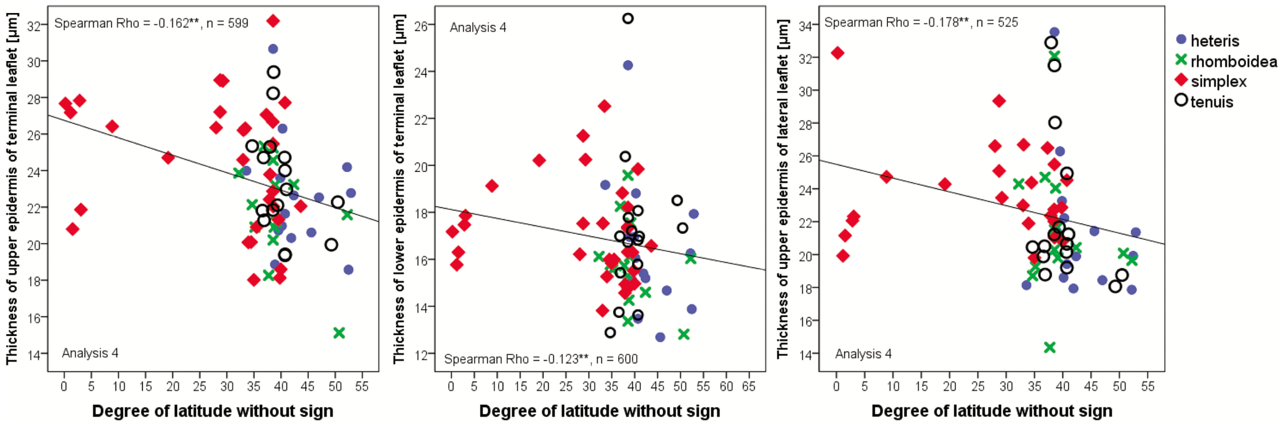

Figure I.6. Spearman correlation of epidermis thickness with the degree of latitude. At the terminal leaflet, the upper and the lower epidermis is significantly negatively correlated, at the lateral leaflet only the upper epidermis is significantly correlated with the latitude. As the material originated from North and South America, we evaluated the statistics for the degree of latitude without a signature (north or south of the equator) for calculating a linear correlation. Rho-value and significance (*α < 0.05, **α < 0.01) are included in the figures. Only significantly correlated data are shown in the figures. The correlations are based on individual data, n = number of individuals; mean square values of the populations are shown in the figure.

Physiological analyses

With water stress, the $\delta^{13}C$ values were higher, meaning that more carbon isotopes had been fixed and assimilated during photosynthesis. As RubisCO discriminates the isotopes in the case of unhampered CO_2 uptake from the atmosphere into the intercellular spaces and across the cell membranes, higher values are a result of partially closed stomata and increase CO_2 isotope concentration within the leaf. With water stress, the proportion of cell volume to intercellular volume increased (consequently the intercellular space decreased, Fig. I.10), and the stomata density also increased (see Fig. I.7). Although the differences between water-stressed and non-stressed individuals are apparent in leaf anatomy (Fig. I.4) and physiology (Fig. I.10), leaf anatomy and $\delta^{13}C$ values clearly differ between the provenances. The *tenuis* leaf type (provenance from the Alps) differed clearly from the other provenances by larger intercellular space compared to the cell volume. This trait coincided with lower $\delta^{13}C$ values which might be the result of reduced ability to close stomata under water stress conditions. The difference between the $\delta^{13}C$ values under water stress conditions might hint at an ecotypic differentiation with a high phenotypic plasticity for the provenances with the other leaf types (*heteris*, *rhomboidea* and *simplex*). The *C. rubella* individuals showed the highest stomata density without water stress. With water stress, these individuals intermingled in between the *C. bursa-pastoris* individuals, so that no differentiation between the two species was apparent from the characterized parameters. When excluding *C. rubella* from the analysis, only the correlation between the percentage cell/intercellular volume versus $\delta^{13}C$ under water stress condition remained significant (Spearman rho = 0.635*). To substantiate differences between the two species, which are often to be found in mixed populations in their common

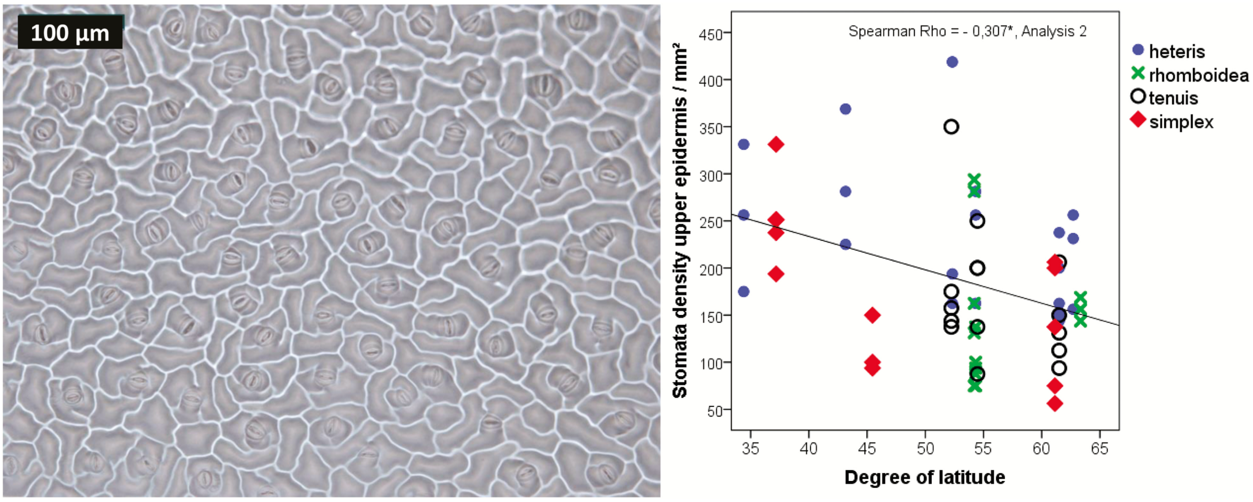

Figure I.7. Spearman correlation of stomata density with latitude in Analysis 2 (15:5 °C, 12-h photoperiod). An example of a nail varnish imprint of the lower epidermis (left) used for the measurements is shown. Rho-value and significance (*α < 0.05, **α < 0.01) are included in the figures. Only significantly correlated data are shown in the figures. The correlations are based on individual data.

distribution area, analysis of a larger number of populations is needed.

The CO_2-assimilation rate correlated highly significantly with the thickness of the leaf. Populations originating from higher latitudes develop thinner leaves under greenhouse conditions (Fig. I.5B), enabling higher CO_2-assimilation rates (Fig. I.11). The non-photochemical quenching (NPQ) of the Russian population increased significantly under high light which might indicate higher stress from the increased temperature for these plants (Fig. I.12). On the other hand, the NPQ of the Moroccan individuals is even lower under high light conditions, suggesting that these conditions are tolerated easily by these individuals (Fig. I.12). The NPQ values are highly significantly negatively correlated with the stomata density at the lower surface (Spearman rho = 0.352**), namely, the stomata density increased at higher NPQ values. Efficient light use for CO_2 assimilation as can be recognized by photochemical quenching (qP) was highest in the Moroccan population under high light conditions, whereas the Russian population was characterized by low qP values (Fig. I.12). Under all other environmental conditions, the populations displayed barely any differences. The qP values are significantly positively correlated with the stomata density at the lower leaf surface (Spearman rho = 0.284*) and with the area of a rosette leaf (Spearman rho = 0.281*).

Leaf types

The geographical distribution of the Mendelian leaf types according to Shull is apparent when regarding the measured leaf thickness from accessions along a transect through North and South America (Fig. I.6): the *simplex* leaf type occurred more frequently close to the equator and seemed to be nearly absent at higher latitudes. This appears to be confirmed in the isotope analysis (Fig. I.8),

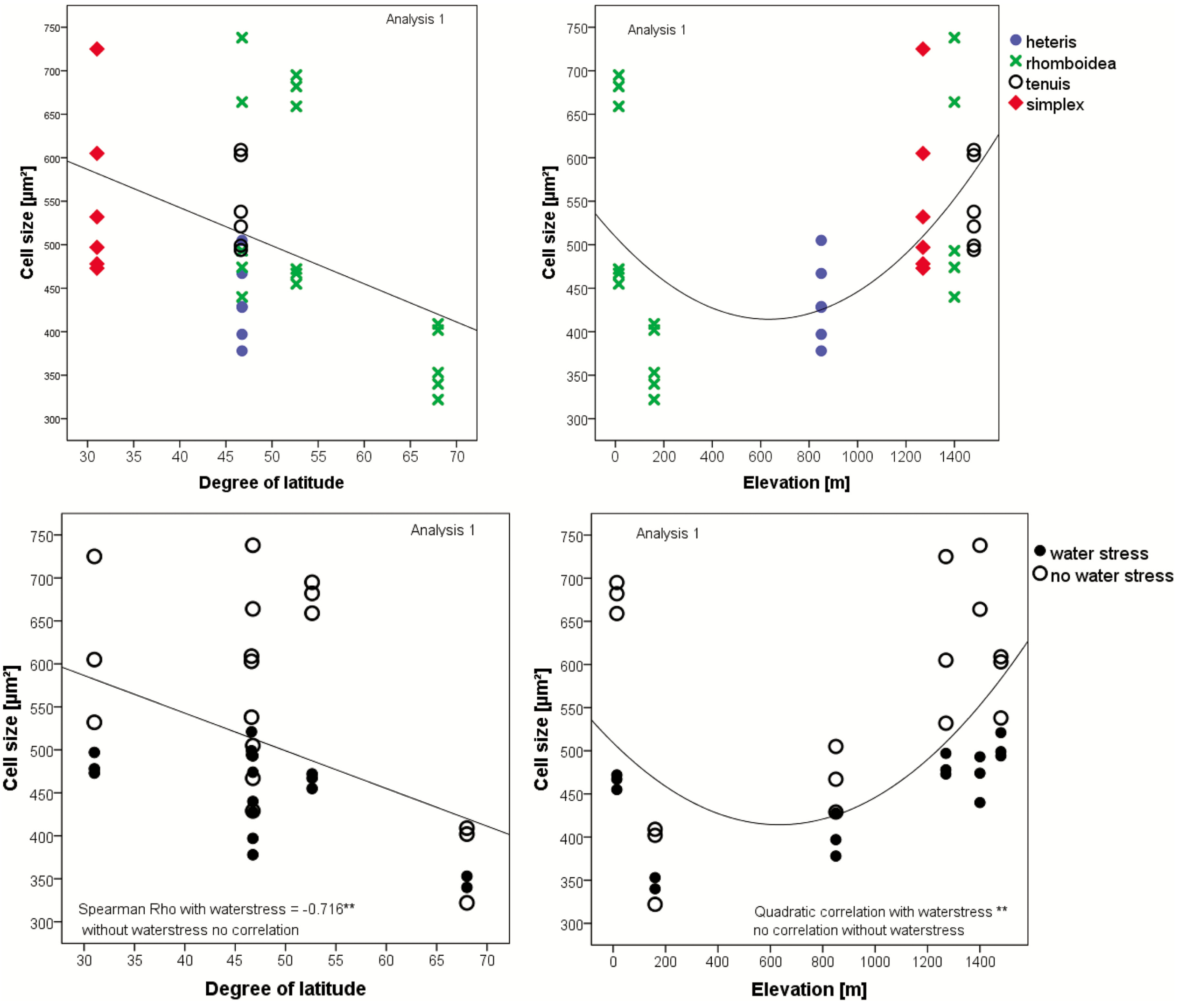

Figure I.8. Correlation of cell size with latitude (linear, Spearman) and altitude (quadratic). The correlation is highly significant under water stress conditions. Upper diagram: labels according to the leaf type; lower diagram: same correlation, but labels according to watering conditions (Analysis 1). Rho-value and significance (*α < 0.05, **α < 0.01) are included in the figures. Both treatment groups differed significantly when tested by Wilcoxon-test (see Table I.2). Only significantly correlated data are shown in the figures. The correlations are based on individual data.

whereas, in order to be able to make a clear statement, the number of studied populations in Analysis 1 is too small. On the other hand, in Analysis 2, this leaf type *simplex* did not correlate with latitude. In all analyses, the *tenuis* leaf type prefers temperate regions with adequate humidity during the vegetation period, even at higher altitudes (Analysis 1, Fig. I.8). The reduced plasticity of *tenuis* compared with the other leaf types is also evident for the physiological traits, as the δ^{13}C values were considerably lower compared with the values of the other leaf types, especially under water stress conditions (Analysis 1, Fig. I.10). However, in Analysis 1, the number of analysed individuals and populations was restricted, and therefore, a higher sample number is necessary to verify our interpretation.

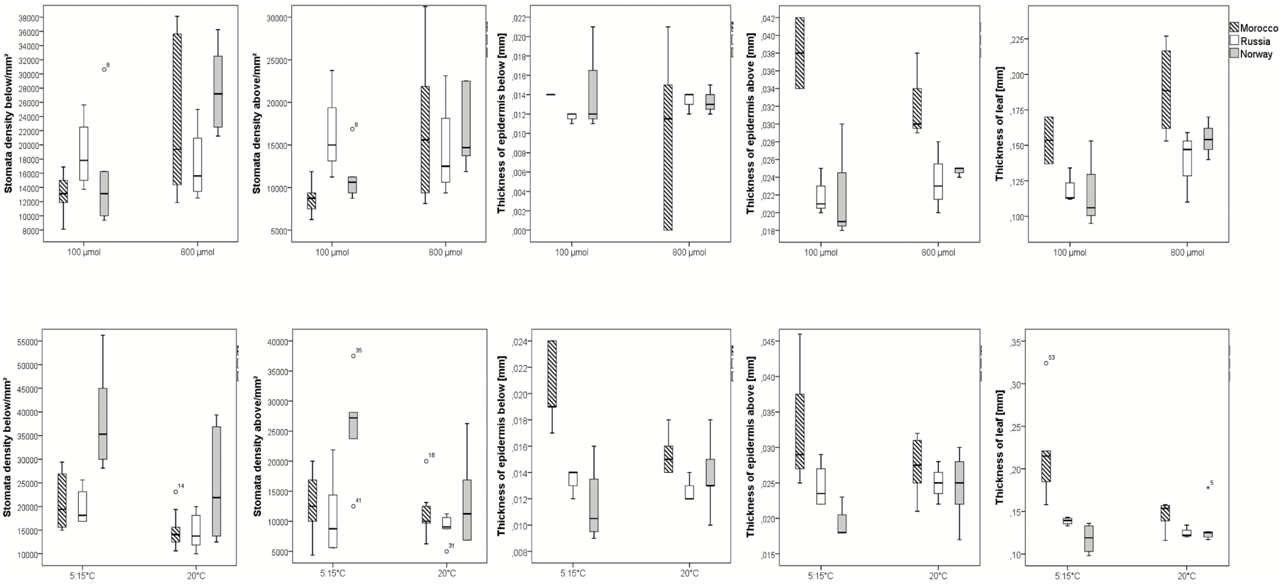

Figure I.9. Anatomical data of provenances from different climatic/geographic regions under low (100 µmol m^{-2} s^{-1}) versus high (800 µmol m^{-2} s^{-1}) light conditions (upper part: 7.5-h photoperiod, 20 °C, Analysis 3), and cold versus warm temperatures (lower part: 12-h photoperiod, 150 µmol quanta/m² s, Analysis 3), three individuals for each provenance and treatment. All treatment groups differed significantly or highly significant when tested by Wilcoxon-test (see Table I.2).

To summarize and generalize the results of our observations and experimental studies, the following statements are put forward:

- The thickness of the leaf, of the epidermis and the epidermal cell size are negatively correlated with the degree of latitude.
- The stomata density varies significantly between different light conditions and provenances.
- Physiological studies ($\delta^{13}C$ values) showed that the leaf types/ecotypes *heteris*, *rhomboidea* and *simplex* appear to be able to close the stomata more efficiently under water stress conditions than the *tenuis* leaf type which might be due to lower plasticity.
- Ecotypes with thinner leaves exhibit a lower maximal rate of CO_2 assimilation (A_{max}) at saturating light.
- Physiological parameters resulting in high photosynthetic capacity under stressful, strong light conditions are typically found when the plants originate from hot, dry and sunny regions.

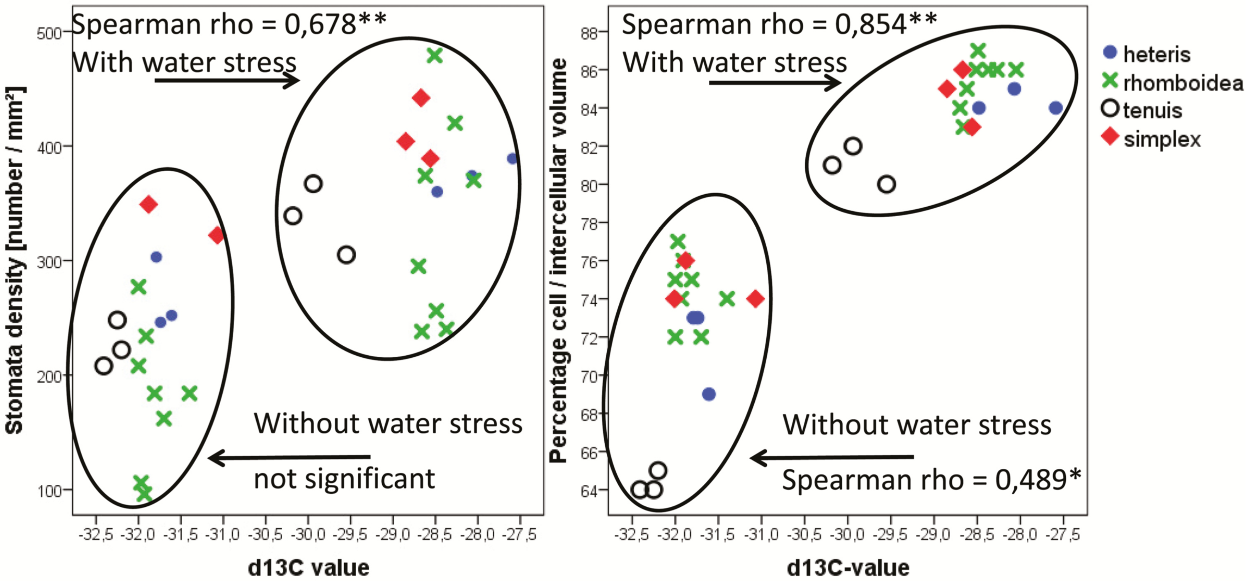

Figure I.10. Spearman correlation of δ^{13}C values and anatomical traits. The δ^{13}C values increased under water stress and were highly significantly correlated with anatomical traits under water stress conditions. The leaf type *tenuis* is characterized by lower δ^{13}C values under both conditions compared with the other types (see 25:15 °C, 9-h photoperiod). Except for *simplex* leaf type (*Capsella rubella*), all individuals belong to *Capsella bursa-pastoris*. Rho-value and significance (*$\alpha < 0.05$, **$\alpha < 0.01$) are included in the figures. Only significantly correlated data are shown in the figures. Without *C. rubella* (phenotype *simplex*) only 'percentage cell/intercellular volume: δ^{13}C' under water stress remained significant (Spearman rho = 0.635*). Both treatment groups differed significantly or highly significant when tested by Wilcoxon-test (see Table I.2). The correlations are based on individual data.

Discussion

The high degree of polymorphism of the leaves in the genus *Capsella* has been well known for more than 100 years. Almquist (1907, 1921) listed 200 elementary species, and in his opinion, this is a result of high variabilty of the genus in nature. In parallel, the geneticist Shull (1909, 1911) performed extensive inheritance studies which formed the basis for the hypothesis of the existence of two Mendelian loci with two alleles, each responsible for the four basic leaf types within the genus. Later on, Shull argued in favour of an additional factor 'I' for leaves with completely entire margins ('*simplissima*', Shull 1929). Particular plants with small rosettes of linear leaves which have a spider-like appearance have been designated by Hus (1914) as *xCapsella bursa-pastoris arachnoidea*, and a leathery appearance corresponds to the dominant allele '*K*' which was named '*coriacea*' factor by Shull (1929). Clausen and Hiesey (1958) confirmed at least four pairs of genes responsible for the leaf shape in *Capsella* and suggested even a higher number of loci that are responsible for other modifications of the leaf. Our observations evidenced that, in the case of heterozygotes of tetraploid *C. bursa-pastoris* individuals, the dominance might be incomplete (e.g. *AaaaBbbb*) and then the leaf type would be intermediate and not distinguishable.

So far, we determined rosette leaves of many provenances worldwide (Fig. I.13). In one case regarding provenances growing along an altitudinal cline, we detected an increase in the percentage

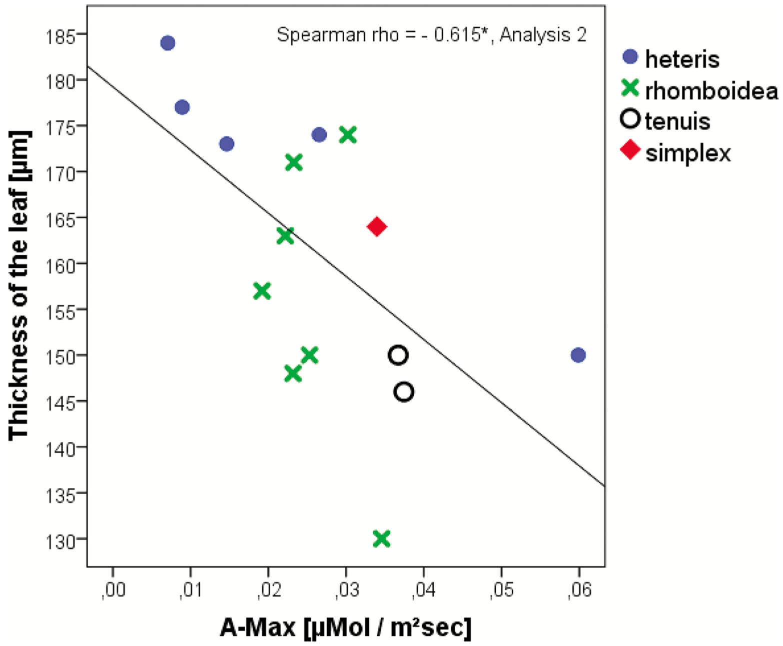

Figure I.11. Spearman correlation of leaf thickness with CO_2-fixation index. This value represents the CO_2-fixation index, i.e. the maximal rate of CO2 assimilation (A_{max}) at saturating light (Analysis 2). Rho-value and significance(*$\alpha < 0.05$, **$\alpha < 0.01$) are included in the figures. Only significantly correlated data are shown in the figures. The correlations are based on individual data.

of the B-allele with higher elevation (Neuffer and Bartelheim 1989). The B-allele is responsible for dividing the lobes to the midrib in the leaf types *heteris* and *rhomboidea*. In another analysis with provenances growing along a latitudinal cline, the variability in leaf types did not correspond to the north–south gradient. In general, out of 15,050 scored leaves (Fig. I.13), 19 % have been *heteris*, 51 % *rhomboidea*, 12 % *tenuis*, 11 % *simplex* and 7 % remained unscored.

The question is: why are the leaf types not randomly distributed in the case of selection neutrality and why is the percentage of the *rhomboidea* leaf type high? Are the leaf types adaptive by themselves, is the adaptation mirrored by anatomical or physiological characters, or is the leaf type linked to adaptive anatomical and/or physiological leaf characters? Dissection of the leaf has been described as being inversely correlated with the mean annual temperature at the community and species level for trees (Royer et al. 2009). Therefore, dissection has been used for paleoclimatic reconstructions (Royer et al. 2005; Little et al. 2010), and it is conceivable that the dominance of the *rhomboidea* leaf type is a consequence of better adaptation of dissected leaves to various climates. In *Capsella*, the morphology of the leaf type depends for some reason on the environmental conditions. *Capsella* appears to exhibit earlier flowering times the longer the day in a long-day photoperiod (Hurka et al. 1976; Neuffer 1990). Under long day and warm temperature conditions some ecotypes flower so early that only a few rosette leaves are able to develop (Neuffer and Hurka 1986), resulting in these leaves which do not attain a pronounced leaf morphology but remain

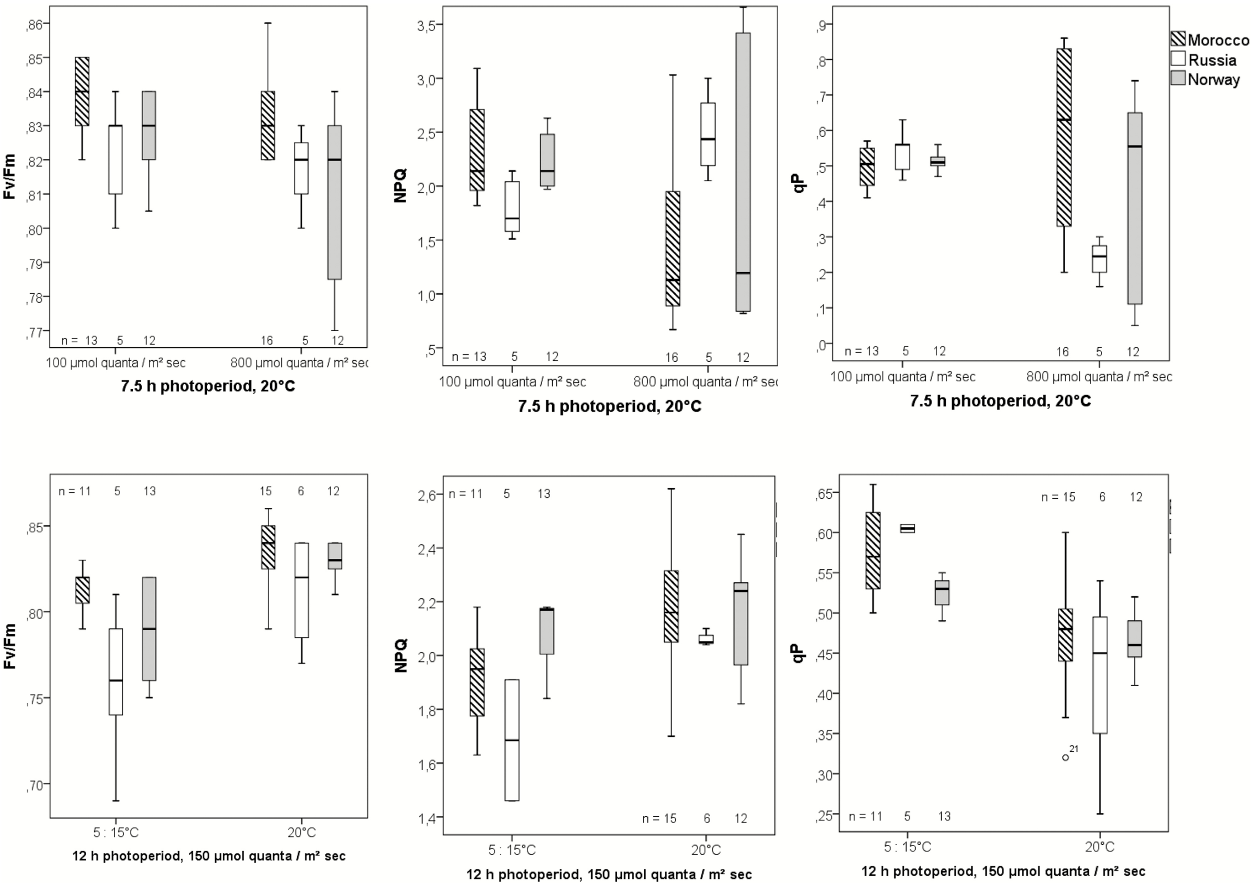

Figure I.12. Chlorophyll fluorescence parameters of provenances from three different regions under low (100 µmol m^{-2} s^{-1}) versus high (800 µmol m^{-2} s^{-1}) light conditions (upper diagrams: 7.5-h photoperiod, 20 °C, Analysis 3), and low versus higher temperatures (lower diagrams: 12-h photoperiod, 150 µmol m^{-2} s^{-1}, Analysis 3). Parameters were measured under steady-state light conditions. F_v/F_m = maximum quantum yield of PSII; NPQ = non-photochemical quenching; qP = photochemical quenching (*n* = number of individuals tested). Most treatment groups differed significantly or highly significant when tested by Wilcoxon-test (see Table I.2).

simple (Neuffer 1989). When grown under short day conditions and/or cold temperatures these provenances produce more rosette leaves (Neuffer and Hurka 1986), which enables them to reach the state which facilitates the development of the more pronounced leaf types (Neuffer 1989). In our study, the distribution of the leaf types was clearly divided into two subgroups according to the climax vegetation zone or the thermal vegetation zone according to Schroeder (1998) documented by the Duncan test in Analysis 4 (Table I.3): regarding climax vegetation, only the leaf type *tenuis* belonged to a second subgroup, whereas regarding thermal vegetation zones, both *tenuis* and *simplex* comprised a second subgroup.

The molecular basis of the leaf shape in Brassicaceae is beginning to be unravelled. At first, the leaf shape seems to have evolved from small, simple leaves (*Aethionema* spec.) to compound leaves (*Cardamine* spec.). First results for *Cardamine hirsuta* have been obtained in the Tsiantis group: they hypothesize that 44 genes are potentially implicated in the leaf development, e.g.

SHOOT MERISTEMLESS, *BREVIPEDICELLUS* or *CUP-SHAPED COTYLEDON* (Gan et al. 2016). For the leaflet formation in comparison with the simple leaves of *Arabidopsis thaliana*, the enrichment of the transcription factors of the *PLETHORAS* family is required, especially of *PLT7* (Gan et al. 2016). Furthermore, Hay and Tsiantis (2016) detected a duplication of the gene *LATE MERISTEM IDENTITY1* (*LMI1*) giving rise to *REDUCED COMPLEXITY* (*RCO*) in *C. hirsuta*. This duplication is lost again in *A. thaliana* and seems to be responsible for the reversal to simple leaves. In a detailed analysis of Sicard et al. (2014) with the two diploid species *Capsella grandiflora* and *C. rubella*, a second duplication which forms *RCO-A* and *RCO-B* has been detected. The difference between *C. grandiflora* with simple leaves (leaf type *simplex*) and *C. rubella* with dissected leaves (leaf type *rhomboidea*) was an allelic variation at the RCO-A locus. Furthermore, these authors detected four insertions of relatively recent origin in the RCO-A genomic organization which differed either in their absence or presence in various provenances. One future aim is to identify the molecular genetic background for the above-mentioned, already known genes and alleles that model the morphology of *Capsella* rosette leaves.

Finally, the adaptation of the rosette leaf is of highest importance, especially in the case of late flowering to biennial ecotypes overwintering with a rosette. In general, *C. bursa-pastoris* forms larger rosette leaves in later flowering plants under field conditions in common garden experiments (e.g. Neuffer and Hurka 1986; Neuffer 2011).

In this study, it is the first time that we report anatomical and physiological results of Shepherd’s Purse leaf types. Körner (2003) reviewed leaf anatomical and physiological characters and discussed how leaves are adapted to high mountain ecosystems. He observed a significantly thicker mesophyll and larger epidermis cells for plants from higher altitudes which is in accordance with our findings. Regarding the fact that the climate in high latitudes of Scandinavia might be similar to high elevations in the Alps, the result for *Capsella* seems to be contradictory at first glance. However, in more northern latitudes, the days in the summer are longer and irradiation less strong. Therefore, the occurrence of thinner leaves with smaller epidermis cells in northern latitudes can be explained as a logical adaptation.

The physiological adaptation of leaves to various environments is often characterized by WUE, as can be deduced from $\delta^{13}C$ values. In a comparison between different provenances of the grass *Leymus chinensis* from dry steppe regions of Asia, the differentiation under various conditions seemed to be more the result of plasticity rather than of ecotypic differentiation (Liu et al. 2016). The authors argue with the clonal propagation of this species which comes close to the general-purpose genotype in the sense of Baker (1974). In our case, the differentiation is apparently not only plastic but also ecotypic with a genetic background. We assume that ecotypes when growing under high light intensities at their places of origin are more adapted to high light, and are able to increase

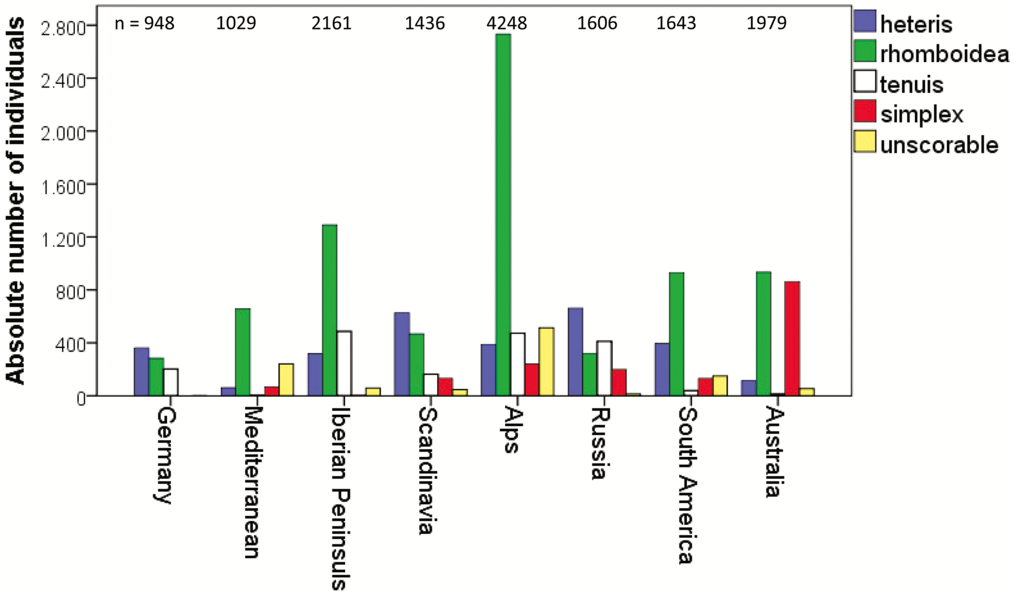

Figure I.13. Distribution of rosette leaf types in various regions worldwide, *n* = number of individuals.

their quantum-yield efficiency considerably, whereas non-adapted genotypes are not able to do so or even suffer from photoinhibition as can be deduced from maximum quantum yield of PSII (F_V/F_m) values when analysed with the FluorCam. This ecotypic differentiation might be the result of the mixed mating system with an outcrossing of up to 12 % under good field conditions (Hurka et al. 1989), whereas *L. chinensis* is a clonally propagating species (Liu et al. 2016).

Another aspect of the adaptation of the leaf are the qualitative and quantitative intraspecific variations of the main flavonoid pattern which was put forward by Eschmann-Grupe (1990) with populations of *C. bursa-pastoris*. The leaves appeared to reflect the adaptation of a population to the place of origin and varied with different environmental conditions. Five main and nine less prominent flavonoids were detected. Focusing on the main flavonoids the authors studied one population from high altitudes in the Alps, one from Norway and one from central Germany under various conditions in the growth chamber as well as in a reciprocal field experiment in 2000 m elevation in the Alps and in central Germany. The three populations varied qualitatively in their flavonoid composition only the population from the Alps contained all five main flavonoids. The Norwegian population contained no isoorientin, and the population from central Germany lacked diosmetin-7-O-β-D-glucoside. Under the various environmental conditions, the pattern did not differ qualitatively, but the quantity was increased significantly in field conditions, especially in the Alps. The composition and amount of secondary metabolites stored in the vacuoles of epidermal cells might be another physiological adaptation of *Capsella* ecotypes to high irradiation. These characteristics are possibly interesting for further studies of the ecotypic differentiation of *Capsella* rosette leaves besides the morphological, anatomical and photosynthetic parameters used in this study.

Table I.3. ANOVA and post hoc Duncan test to prove leaf type distribution to climax vegetation zones and thermal vegetation zones (Schroeder 1998, Analysis 4). Probability for subgroups $\alpha = 0.05$.

Leaf type	**N**	**Climate vegetation zone**		**Thermal vegetation zone**	
		Group 1	**Group 2**	**Group 1**	**Group 2**
heteris	125	4.72		2.92	
rhomboidea	90	4.67		3.00	
tenuis	125		5.92		3.40
simplex	260	4.25			3.35
Significance		0.224	1.000	0.518	0.663

Conclusion

Here, we present a first insight into the ecotypic differentiation of *Capsella* rosette leaves in a combination of morphological, anatomical and physiological characters. The geographical distribution and frequencies of specific morphological leaf types seem to mirror the adaptation to particular environmental conditions at the places of origin. However, the actual adaptation might be overlaid by anatomical and physiological adaptive traits which, with the numerous combinations of variations, point to a genetic background. To unveil the molecular background of the ecotypic differentiation of the *Capsella* rosette leaves, the knowledge of the molecular settings behind the leaf morphology is not yet sufficient. The clear geographic distribution pattern of the morphological leaf types might be partially adaptive by itself and therefore responsible for the frequency differences in various regions. However, the selection for specific morphological leaf types under different environmental conditions could be caused by genetic hitchhiking via anatomical or physiological adaptive characters with a genetic background. This linkage between the various traits is possible via closely linked loci on the same chromosome, and the effect is enhanced tremendously by the mating system which relies predominantly on selfing. It is, therefore, necessary to phenotypically and genetically elucidate all three aspects in combination: morphology, anatomy and physiology.

Contributions by the Authors

B.N. provided the material, supervised Analyses 1 and 4 as well as the anatomical and morphological part of Analyses 2 and 3, evaluated data and wrote draft versions of the manuscript. C.W. supervised Analysis 4 which is part of her PhD thesis, evaluated data and contributed to draft versions of the manuscript. I.V. performed the physiological lab work of Analyses 2 and 3, and evaluated and interpreted these data. R.S. supervised the physiological part of Analyses 2 and 3, and wrote parts of the manuscript.

Sources of Funding

Financial support for expeditions to collect plant material was given in the framework of various projects funded by the DAAD, DFG and BMBF. Support for Analysis 4 was provided by DFG (NE 314/11-2).

Conflict of Interest

None declared.

Acknowledgments

We would like to thank S. Sutmöller for performing Analysis 1 and G. Schleser for providing staff and experience at the KFA Jülich, Germany, for δ^{13}C analyses in this part. We would like to thank M. Behr for helping with experiments for Analysis 2, F. Lojewski for Analysis 3 and K. Nithyanandan for Analysis 4. We also thank the staff of the Botanical Garden of the University of Osnabrück for their help with the field experiment of Analysis 4, and R. Grupe, University of Osnabrück, for technical assistance. L. Schmieding we thank for proof reading and editing the English text. Financial support of the DFG is greatly acknowledged. Lastly, we thank two anonymous reviewers for many substantial comments which helped to significantly improve the manuscript.

Literature Cited

Almquist E. 1907. Studien über die *Capsella bursa-pastoris* (L.). *Acta Horti Bergiani* **4**:1–92.

Almquist E. 1921. Studien über die *Capsella bursa-pastoris* II. *Acta Horti Bergiani* **7**:41–95.

Athanasiou K, Dyson BC, Webster RE, Johnson GN. 2010. Dynamic acclimation of photosynthesis increases plant fitness in changing environments. *Plant Physiology* **152**:366–373.

Baker MG. 1974. The evolution of weeds. *Annual Review of Ecology, Evolution, and Systematics* **5**:1–24.

Clausen J, Hiesey WM. 1958. Experimental studies on the nature of species. IV. Genetic structure of ecological races. *Carnegie Institution of Washington Publication* **615**:171–175.

Coquillat M. 1951. Sur les plantes les plus communes a la surface du globe. *Bulletin mensuel de la Société linnéenne de Lyon* **20**:165–170.

Craig H. 1957. The geochemistry of stable carbon isotopes. *Geochimica et Cosmochimica Acta* **3**:53–92.

Eschmann-Grupe G. 1990. Licht- und Temperatureinflüsse auf die intraspezifische Variation der Flavonoidprofile bei *Capsella bursa-pastoris* (L.) Med. *Verhandlungen der Gesellschaft für Ökolologie* **19**:82–86.

Gan X, Hay A, Kwantes M, Haberer G, Hallab A, IoioRD, Hofhuis H, Pieper B, Cartolano M, Neumann U, Nikolov LA, Song B, Hajheidari M, Briskine R, Kougioumoutzi E, Vlad D, Broholm S, Hein J, Meksem K, Lightfoot D, Shimizu KK, Shimizu-Inatsugi R, Imprialou M, Kudrna D, Wing R, Sato S, Huijser P, Filatov D, Mayer KFX, Mott R, Tsiantis M. 2016. The *Cardamine hirsuta* genome offers insight into the evolution of morphological diversity. *Nature Plants* **2**:16167.

Genty B, Briantais JM, Baker NR. 1989. The relationship between the quantum yield of photosynthetic electron transport and quenching of chlorophyll fluorescence. *Biochimica et Biophysica Acta* **990**:87–92.

Hanke GT, Holtgrefe S, König N, Strodtkötter I, Voss I, Scheibe R. 2009. Use of transgenic plants to uncover strategies for maintenance of redox homeostasis during photosynthesis. *Advances in Botanical Research* **52**:207–251.

Hay A, Tsiantis M. 2016. *Cardamine hirsuta*: a comparative view. *Genetics and Development* **39**:1–7.

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

Hurka H, Haase R. 1982. Seed ecology of *Capsella bursa-pastoris* (Cruciferae): dispersal mechanism and the soil seed bank. *Flora* **172**:35–46.

Hurka H, Krauss R, Reiner T, Wöhrmann K. 1976. Das Blühverhalten von *Capsella bursa-pastoris* (Brassicaceae) [the flowering behaviour of *Capsella bursa-pastoris* (Brassicaceae)]. *Plant Systematics and Evolution* **125**:87–95.

Hurka H, Neuffer B. 1991. Colonizing success in plants: genetic variation and phenotypic plasticity in life-history traits in *Capsella bursa-pastoris*. In: Esser G, Overdieck D, eds. *Modern ecology: basic and applied aspects*. Amsterdam, New York: Elsevier, 77–96.

Hus H. 1914. The origin of x*Capsella bursa-pastoris arachnoidea*. *The American Naturalist* **48**:193–235.

Körner C. 2003 Uptake and loss of carbon. In: Körner C, ed. *Alpine plant life – functional plant ecology of high mountain ecosystems*, 2nd edn, Chapter 11. Berlin, Heidelberg, New York: Springer, 171–200.

Kryvokhyzha D, Holm K, Chen J, Cornille A, Glémin S, Wright SI, Lagercrantz U, Lascoux M. 2016. The influence of population structure on gene expression and flowering time variation in the ubiquitous weed *Capsella bursa-pastoris* (Brassicaceae). *Molecular Ecology* **25**:1106–1121.

Little SA, Kembel SW, Wilf P. 2010. Paleotemperature proxies from leaf fossils reinterpreted in light of evolutionary history. *PLoS One* **5**:e15161.

Liu Y, Zhang L, Xu X, Niu H. 2016. Understanding the wide geographic range of a clonal perennial grass: plasticity versus local adaptation. *AOB Plants* **8**:plv141; doi:10.1093/aobpla/plv141.

Mooney HA, Mack RN, McNeely JA, Neville LE, Schei PJ, Waage JK. 2005. *Invasive alien species*. Washington, D.C.: Island Press.

Neuffer B. 1989. Leaf morphology in *Capsella* (Cruciferae) – dependency on environments and biological parameters. *Beiträge zur Biologie der Pflanzen* **64**:39–54.

Neuffer B. 1990. Ecotype differentiation in *Capsella*. *Vegetatio* **89**:165–171.

Neuffer B. 2011. Native range variation in *Capsella bursa-pastoris* (Brassicaceae) along a 2500 km latitudinal transect. *Flora* **206**:107–119.

Neuffer B, Bartelheim S. 1989. Gen-ecology of *Capsella bursa-pastoris* from an altitudinal transect in the Alps. *Oecologia* **81**:521–527.

Neuffer B, Bernhardt K-G, Hurka H, Kropf M. 2011. Monitoring population and gene pool dynamics of the annual species *Capsella bursa-pastoris* (Brassicaceae) – initiation of a long-term genetic monitoring and a review of relevant species traits. *Biodiversity and Conservation* **20**:309–323.

Neuffer B, Hirschle S, Jäger S. 1999. The colonizing history of *Capsella* in Patagonia (South America) – Molecular and adaptive significance. *Folia Geobotanica* **34**:435–450.

Neuffer B, Hurka H. 1986. Variation of growth form parameters in *Capsella* (Cruciferae). *Plant Systematics and Evolution* **153**:265–296.

Neuffer B, Hurka H. 1999. Colonization history and introduction dynamics of *Capsella bursa-pastoris* (Brassicaceae) in North America: isozymes and quantitative traits. *Molecular Ecology* **8**:1667–1681.

Neuffer B, Linde M. 1999. *Capsella bursa-pastoris* – colonisation and adaptation; a globe-trotter conquers the world. In: van Raamsdonk LWD, den Nijs JCM, eds. *Plant evolution in man-made habitats*. Amsterdam: Proc. VIIth Symp. IOPB 1998, 49–72.

Nicotra AB, Leigh A, Boyce CK, Jones CS, Niklas KJ, Royer DL, Tsukaya H. 2011. The evolution and functional significance of leaf shape in the angio-sperms. *Functional Plant Biology* **38**:535– 552.

Royer DL, Meyerson LA, Robertson KM, Adams JM. 2009. Phenotypic plasticity of leaf shape along a temperature gradient in Acer rubrum. PLoS One **4**:e7653.

Royer DL, Wilf P, Janesko DA, Kowalski EA, Dilcher DL. 2005. Correlations of climate and plant ecology to leaf size and shape: potential proxies for the fossil record. *American Journal of Botany* **92**:1141–1151.

Scheibe R, Backhausen JE, Emmerlich V, Holtgrefe S. 2005. Strategies to maintain redox homeostasis during photosynthesis under changing conditions. *Journal of Experimental Botany* **56**:1481–1489.

Scheibe R, Dietz KJ. 2012. Reduction-oxidation network for flexible adjustment of cellular metabolism in photoautotrophic cells. *Plant, Cell & Environment* **35**:202–216.

Schleser GH, Bernhardt K-G, Hurka H. 1989. Climatic adaptability of populations of *Diplotaxis erucoides* D.C. from Sicily, based on leaf morphology, leaf anatomy and δ^{13}C studies. *International Journal of Biometeorology* **33**:109–118.

Schleser GH, Poling R. 1980. δ^{13}C record in forest soil using rapid method for preparing carbon dioxide samples. *The International Journal of Applied Radiation and Isotopes* **31**:769–773.

Schreiber U, Schliwa U, Bilger W. 1986. Continuous recording of photochemical and non-photochemical chlorophyll fluorescence quenching with a new type of modulation fluorometer. Photosynthesis Research **10**:51–62.

Schroeder FG. 1998. *Lehrbuch der Pflanzengeographie*. Wiesbaden: Quelle und Meyer.

Shull H. 1909. *Bursa bursa-pastoris* and *Bursa heegeri*: biotypes and hybrids. *Carnegie Institution of Washington Publication* **112**:3–56.

Shull H. 1911. Defective inheritance-ratios in *Bursa* hybrids. Verhandlungen des naturhistorischen

Vereins in Brünn **49**:156–168.

Shull GH. 1929. Species hybridizations among old andnew species of Shepherd's Purse. *Proceedings of the International Congress of Plant Science* **1**:837–888.

Sicard A, Thamm A, Marona C, Lee YW, Wahl V, Stinchcombe JR, Wright SI, Kappel C, Lenhard M. 2014. Repeated evolutionary changes of leaf morphology caused by mutations to a homeobox gene. *Current Biology* **24**:1880–1886.

Silva B, Roos K, Voss I, Konig N, Rollenbeck R, Scheibe R, Beck E, Bendix J. 2012. Simulating canopy photosynthesis for two competing species of an anthropogenic grassland community in the Andes of southern Ecuador. *Ecological Modelling* **239**:14–26.

Turesson G. 1922a. The species and variety as ecological units. *Hereditas* **3**:323–334.

Turesson G. 1922b. The genotypical response of the plant species to the habitat. *Hereditas* **3**:211–350.

Turesson G. 1930. The selective effect of climate upon the plant species. *Hereditas* **14**:99–152.

Voss I, Sunil B, Scheibe R, Raghavendra AS. 2013. Emerging concept for the role of photorespiration as an important part of abiotic stress response. *Plant Biology* **15**:713–722.

Wright IJ, Dong N, Maire V, Prentice IC, Westoby M, Díaz S, Gallagher RV, Jacobs B, Kooyman R, Law EA, Leishman MR, Niinemets Ü, Reich PB, Sack L, Villar R, Wang H, Wilf P. 2017. Global climatic drivers of leaf size. *Science* **357**:917–921.

Zhou TY, Lu LL, Yang G, Al-Shehbaz IA. 2001. Brassicaceae (Cruciferae). In: Wu ZY, Raven PH, eds. *Flora of China 8*. Beijing, St Louis: Science Press/Missouri Botanical Garden Press, 1–193.

II. Geographical structure of genetic diversity in Shepherd's Purse, *Capsella bursa-pastoris* – a global perspective

Christina Wesse, Herbert Hurka, Erik Welk, Barbara Neuffer

In preparation for submission

AIM

To display and analyze global geographical distribution patterns of isozyme genotypes of the cosmopolitan plant *Capsella bursa-pastoris*, and to understand driving forces of the resulting distribution patterns.

LOCATION

Worldwide

TAXON

Capsella bursa-pastoris (Brassicaceae)

METHODS

We sampled 21,812 *C. bursa-pastoris* individuals randomly taken from natural provenances, covering a broad spectrum of the distribution range. Polyacrylamide gel electrophoresis was performed to assay different isozyme systems. We estimated allele frequencies and recorded genotypes at single loci and at 18 multilocus associations. Geographical structures of alleles and genotypes are shown in maps and tables. Population structure was analyzed with STRUCTURE.

RESULTS

Geographical structure of genetic variation at isozyme level is similar in native and introduced ranges. Population structure analysis revealed two clusters, one distributed predominantly in warm arid to semi-arid climate regions, the other predominantly in more temperate humid to semi-humid climate regions. We observed admixture in both the native and non-native ranges predominantly in regions with intermediate water balance. Middle and Western Europe had the highest genotype diversity followed by eastern Europe. Genotype diversity in the introduced ranges is lower than in the native ranges.

MAIN CONCLUSION

The two clusters detected in *C. bursa-pastoris* point to an early diversification into two lineages or may even suggest multiple origins of the species. The worldwide distribution patterns of genetic variation of *C. bursa-pastoris* can be explained by intra- and intercontinental migration but environmental filtering due to climate pre-adaption seems also involved. We have been able to reconstruct colonization history and invasion routes and identified source areas in the native range. Multiple independent introductions of genotypes from different source regions are obvious. 'Endemic' genotypes might be the outcome of admixture or of *de novo* mutation. We conclude that most colonizing *Capsella* genotypes are pre-adapted and found matching niche conditions in the colonized range parts.

Keywords: adaptation, biogeography of genetic diversity, *Capsella bursa-pastoris*, colonization, isozymes, migration routes, multilocus genotypes, STRUCTURE analysis

Introduction

Range expansion or colonization is *per se* a feature of the evolutionary history of all species and occurs over geological time scales to more recent man-caused dispersal, from intercontinental migration to regional and local range extensions. The 'Genetics of Colonizing Species', edited by Baker and Stebbins in 1965, was the first synthesis on the genetics and evolution of colonizers. One of the topics of the 'Genetics of Colonizing Species' was the evolutionary history of colonizing species, the role of bottlenecks and the genetic diversity of colonizing species, and focussed research on the influence of genetic variation on colonizing success for the years to come. It can be regarded as the foundational document for "invasion genetics" (Barrett, 2015) and the application of genetic techniques to study the introduction and spread of introduced (non-native) species throughout the world. Issues of particular interest to be addressed by genetic surveys are (i) identification of source populations for colonization; (ii) detection of single or multiple introductions; (iii) comparison of population structure between native and introduced populations; (iv) genetic diversity in the non-native range compared to the native range; (v) the detection of bottlenecks and founder events; (vi) consider pre-adaptation vs. post-colonization adaptation enabling invasive spread; (vii) genetic interactions during admixture of multiple source populations; (viii) new mutations in the introduced range. One has to bear in mind the different types of genes, however. Genetic variation at the molecular level is different in quality from that at the phenotypic level. The correlation between molecular genetic marker assays (Mendelian loci) and ecological relevant quantitative genetic variation (polygenic) is generally low.

The use of a diverse array of neutral molecular markers e.g. isozymes, RAPDs, AFLPs, microsatellites and finally DNA sequences and next-generation sequencing, have greatly enhanced the ability to reconstruct the evolutionary history of biological invasions and to assess the magnitude of genetic bottlenecks and founder events (e.g., Barrett, 2015; Cristescu, 2015). There is now evidence from neutral loci that many populations of introduced species have less genetic variation than populations in the native range (Barrett, 2015), although the genetic diversity of introduced, non-native populations seems to be only moderately reduced in comparison to native populations (Bossdorf et al., 2005; Dlugosch & Parker, 2008). However, inferences regarding differentiation between genotypes from the native to the introduced range are prone to sampling errors and are often confounded by non-random geographic sampling, thus missing among-population variation within each range when diversity is geographically structured (Colautti & Lau, 2015).

In the present study, we present genetic data that offer insights into the sources, routes and global spread of one of the most frequent and wide-spread flowering plants on earth *Capsella bursa-pastoris*, Shepherd's Purse.

The genus *Capsella* (Brassicaceae) comprises five species (Chater in Tutin et al., 1964, reduced in Tutin et al., 1993 to four). (The numerous taxa recognised by Almquist, 1907, 1921, and Shull, 1929 are not considered here). All species are of old-world origin. The evolutionary history of the genus *Capsella* has been outlined recently (Hurka et al., 2012). Three species are diploid (2n = 2x = 16) and two are tetraploid (2n = 4x = 32). Two species, the diploid *Capsella rubella* Reuter, and the tetraploid *C. bursa-pastoris* (L.) Medik. are successful intercontinental colonizers. Whereas *Capsella rubella* remained restricted mainly to Mediterranean climatic regions (Paetsch et al., 2010), *C. bursa-pastoris* is a worldwide colonizer. It is one of the most frequent and wide-spread flowering plants on earth (Coquillat, 1951; Zhou et al., 2001; Randall, 2012) avoiding only the hot and wet tropical lowlands. Preferential habitats are cultivated and disturbed soils and ruderal sites. It is an annual to winter annual, predominantly selfing species of enormous seed output and thus high reproductive capacity (Hurka & Neuffer, 1991; 1997). Seeds are incorporated into the soil seed bank (Hurka & Haase, 1982) where they can survive for many years (90 years reported in Kivilaan & Bandurski, 1973). The seeds produce a mucilage thus promoting long distance transport via myxospermy (Neuffer & Linde, 1999). The species displays "fixed heterozygosity" and disomic inheritance despite tetraploidy (Shull, 1929 for morphological characters; Hurka et al., 1989; Hurka & Düring, 1994; Neuffer & Hurka, 1999 for allozymes). The pronounced ecotypic variation in both the native and introduced range is remarkable (e.g., Neuffer & Bartelheim, 1989; Neuffer & Hurka, 1999; Neuffer, 2011; Neuffer et al., 2018). These characteristics may account, at least in part, for the extraordinary colonization success.

Capsella bursa-pastoris originated in Eurasia probably in pre-(last)glacial times (Hurka & Neuffer, 1997; Hurka et al., 2012). In post-Columbian time, it was introduced by European colonists to the New World, Australasia and southern Africa. For certain regions, the colonization history has been traced by molecular markers (e.g. RAPDs in Neuffer, 1996; isozymes in Neuffer & Hurka, 1999; isozymes and RAPDs in Neuffer et al., 1999; isozymes in Neuffer et al., 2011; GBS in Kryvokhyzha et al., 2016).

The largest molecular marker dataset so far available for *Capsella* is based on isozyme analyses. Discrete and co-dominant inheritance, the direct identification of allozymes and allelic variation at single loci, as well as their selection neutrality and low mutation rates make them valuable tools for studying genetic variation at the population level. Estimation of genetic variability by isozymes is conservative since genetic variation is underestimated by isozyme studies. The lower resolution power in comparison to other molecular markers is balanced by the high reliability of results uncovering essential basic features.

Isozymes have been used intensively by us to analyze speciation processes in the genus *Capsella* and to analyze genetic variation at the population level. For certain regions, colonization history has been traced (see above) but a worldwide over-all view is missing. Meanwhile, our isozyme dataset has been substantially enlarged as new datasets are combined with the already published ones. The meta-dataset comprises >20,000 individuals from an array of provenances which cover the whole distribution ranges in the native and the introduced ranges. We evaluated the data in rising complexity, from allozyme frequencies via single locus up to multilocus genotypes. Informative value increases along this sequence. Finally, to assess the structure of allele frequency variation in the allozyme dataset, we used Bayesian clustering to assign multilocus genotypes into clusters.

The objective of the present study is to display and analyze the global genetic variation pattern of *Capsella bursa-pastoris* as expressed in isozyme variability. Is genetic variation spatially structured in the native range and mirrored in the introduced range and, if so, how can this be explained? To what extent is the global variation pattern related to colonization histories and to what extent to adaptation processes?

Methods

Plant Material

Individual seed samples from *Capsella* plants were collected from 1,469 natural provenances from all over the world. Samples were geolocated and assigned to geographical regions: The Iberian Peninsula (IBE): Portugal, Spain. British Isles (BRT): Great Britain, Ireland. Middle and Western Europe (M+WE): Andorra, Austria, Czech Republic, France, Germany, Hungary, Liechtenstein, Netherlands, Poland, Slovakia, Switzerland. Mediterranean areas (MED): Egypt, Greece, Israel, Italy, Morocco, Turkey, former Yugoslavia. Scandinavia (SCN): Denmark, Finland, Iceland, Norway, Sweden. Eastern Europe (EEU): Bulgaria, Estonia and European Russia. Asia (ASIA): Afghanistan, Armenia, Asian Russia, China, Iran, Japan, Kazakhstan, Kyrgyzstan, Mongolia, Nepal, Sri Lanka. California (CAL). North America (NAM): Canada, USA (except California). Middle and South America (M+SA): Argentina, Bolivia, Chile, Costa Rica, Ecuador, Falkland Islands, Mexico, Venezuela. Australasia (AUS): Australia, New Zealand. Africa (AFR): Republic of South Africa.

Seeds were stored in plastic bags at -20 °C until sowing for analyses. Determination of species was performed in the greenhouse with flowering individuals, chromosome counting, flow cytometry and isozyme analysis. Isozymes facilitate the distinction between tetraploid and diploid individuals which occasionally might otherwise be difficult. 27,323 individuals from the whole genus were used in this study, with *n* = 21,812 identified as *C. bursa-pastoris*, *n* = 263 as *C. thracica*, *n* = 109 as *C. orientalis*, *n* = 3,141 as *C. rubella* and *n* = 1,998 as *C. grandiflora*. Herbarium material of many accessions is deposited in the Herbarium of the Osnabrück University OSBU. Plants were grown in the greenhouse of the Department of Botany or the Botanical Garden of the Osnabrück University and rosette leaves of single plants were harvested and stored at -80 °C.

Isozyme analyses

Electrophoresis was performed in a continuous system on vertical polyacrylamide gel slabs (PAGE). Following isozyme systems were assayed: aspartate aminotransferase (AAT; EC 2.6.1.1), glutamate dehydrogenase (GDH; EC 1.4.1.4), and leucine aminopeptidase (LAP; 3.4.11.1). Buffer systems and other experimental details are given in Hurka et al. (1989) for AAT, in Hurka & Düring (1994) for GDH, and in Neuffer & Hurka (1999) for LAP. Isozyme data were either previously published or are presented here for the first time. The genetics of these enzyme systems in Capsella have been deciphered in the above cited literature, and the previous nomenclature of the enzyme loci and their isozymes adopted in the present study with few modifications: Re-evaluation of the isozyme patterns provoked us to include the allele *Lap3-1* into *Lap3-2*, *Lap3-3* into *Lap3-4*, and *Lap3-7* into *Lap3-2*.

Data Evaluation

The complete isozyme data for *C. bursa-pastoris* (i.e. samples without missing data, *n* = 8,076) were evaluated with the program GenAlEx 6 (Peakall and Smouse, 2006; 2012) for *F*-statistics and the Mantel test, which correlates the matrix of pair-wise genetic distances among populations and the matrix of geographic distances (km) among populations (Mantel, 1967). Significance tests were based on 999 permutations. We also quantified population genetic diversity of *C. bursa-pastoris* using an analysis of molecular variance (AMOVA) (Excoffier et al., 1992) and also calculated a measure of genetic variation (SSWP/*n* – 1) by calculating the population-wise AMOVA sums of squares divided by $n - 1$ (Fischer and Matthies, 1998). The SSWP values were sample size-corrected.

We used a Bayesian clustering method to partition population structure in *C. bursa-pastoris*. First, we quantified population structure in the dataset in STRUCTURE v. *2.3.4* (Pritchard et al., 2000). For each analysis, we implemented a model of correlated allele frequencies (Falush et al., 2003) and admixture, and applied the default setting for all other parameters. Ten independent runs for all values of *K* (number of genetic clusters) between 1 and 10 were performed using an MCMC length of 10^6 generations following a burn-in of 10^5 generations. For each *K* value, we used clumpp v. *1.1.2* (Jakobsson & Rosenberg, 2007) to examine consistency across replicate cluster analyses by estimating the highest value of pairwise similarity (H' value) and averaged assignment probabilities for each individual. We applied the *Greedy* algorithm for $K = 1$ to $K = 8$, using 1000 random input orders. The best *K* value was chosen by examining the log probability of the data [ln Pr(X|K)] and plots of ΔK (Evanno et al., 2005) produced by structure harvester (Earl & von Holdt, 2012).

Distribution modelling of genotype clusters

Because *C. bursa-pastoris* is a wide-ranging species, known to be present at various locations around the world, a species distribution modeling (SDM) approach was used to predict and analyze the potential distribution of the two clusters derived from the population genetic clustering approach (STRUCTURE). To this end, we used the software MaxEnt v. *3.3.4*, a machine learning algorithm (Phillips et al., 2006). Unlike presence-absence models, the maxent algorithm is based primarily on presence data as the basis of its predictions, and is therefore especially suitable for the given data, since absence data for genotypes are not available. It has repeatedly been proven to be an effective method for predicting potential species distributions in scarce data situations (e.g. Merow et al., 2013). Elith et al. (2016) found that MaxEnt was one of the best of 16 different methods for modeling the distributions of 226 species in 6 different regions.

The R package ENMeval (Muscarella et al., 2014) was used to maximize predictive ability and avoid overfitting problems that might result from the spatial clustering of sampling localities (see Radosavljevic & Anderson, 2014). With this the most suitable MaxEnt configuration settings are evaluated via a range of regularization multiplier values (0.5 to 2.5, by 0.5) and the combinations of feature class to consider (Linear: L, Quadratic: Q, Polynomic: P). The regularization multiplier imposes a penalty on model complexity and thus results in simpler model predictions. The selection of feature classes determines the potential shape of the response curves (L, Q, LQ, LQP). The given workflow allows for avoiding model overfitting and selecting the 'best' model according to AICc. To determine the accuracy of the resulting models, we used the area under the curve (AUC) of a receiver-operator characteristics curve (ROC). The AUC score is the dominant tool to measure the model performance, mainly due to its independence to threshold choices. The higher the value of AUC (closer to 1), the better is the models ability to handle the inevitable trade-off between under- and overestimation.

From the ENMeval model comparison approach, the best resulting model configuration in terms of complexity (AICc) and accuracy (AUC) was selected for each cluster, respectively. Accordingly, we performed multiple runs, with random 5-fold crossvalidation between test- and trainingdata. Linear, quadratic, and polynomial functions were used as single and combined response options, and the number of background samples was set to 50,000 to enable the worldwide climatic background to be sampled. Finally, the resulting model solutions were compared in terms of complexity (AICc) and accuracy (AUC) and the best resulting models were selected for each cluster, respectively. From this a regularization value of 1.1 was obtained as best option to avoid overfitting.

To derive two distinct groups the cluster affiliation probabilities from the population genetic structure analysis were binarized at the threshold of 0.5 probability for the respective cluster. Climate data at the sampling localities were extracted as part of the MaxEnt distribution modelling with removal of duplicate points per raster gridcell. The iteratively self-optimising MaxEnt algorithm inherently identifies variables that contribute most to increasing predictive success, thus, an *a priori* variable exclusion via cross-correlation or PCA approaches is not mandatory and one of two potentially correlated variables will be downweighted in the further process.

Results

Allele biogeography

Locus *Aat1*: The most common allele *Aat1-1* occurs in high frequencies throughout the world while the other alleles display regional differences (Fig. II.1a). The center of *Aat1-4* distribution in the native range is Europe, and in the introduced ranges NAM and AUS (Fig. II.1a).

Locus *Aat2*: The alleles *Aat2-1* and *Aat2-4* are very common and distributed worldwide. They account for 98 % of the overall frequency. In the native range, *Aat2-1* is most common in M+WE (39 % of the overall frequency) followed by SCN (10 %), IBE (8 %) and MED (9 %). In the non-native range, overall frequency is 9 % in CAL, in NAM and in AUS 5 %. Allele *Aat2-4* is most frequent in M+WE (28 % overall frequency), SCN (13 %) and IBE (12.5 %), and in the introduced ranges in CAL (13 %) and AUS (7.5 %) (Fig. II.1b).

Locus *Aat3*: Most common alleles are *Aat3-5* (48 %), *Aat3-1* (28 %) and *Aat3-3* (17 %) summing up to 93 % overall frequency. In the native range, *Aat3-5* is most common in M+WE (32 % of overall frequency), SCN (11 %) and IBE (11 %), and in the introduced ranges in CAL

(13 %) and AUS (7 %) (Fig. II.1c). Similar to this picture is the overall frequency distribution of allele *Aat3-1* (Fig. II.1c). Frequency distribution of allele *Aat3-3* is different. It is most common in M+WE (66 %) and SCN (6.5 %), and in the non-native range in NAM (7 %) and CAL (4 %) (Fig. II.1c).

Locus *Gdh2*: The two most common alleles, *Gdh2-1* and *Gdh2-2*, are common in the native and introduced ranges, whereas *Gdh2-3* is centered in BRT and in middle and northern Europe. The highest frequency outside the native range is in NAM (Fig. II.1d).

Locus *Lap3*: The most common alleles are *Lap3-2* and *Lap3-5*. Their frequencies add up to ca. 94 % and are distributed all over the world. Allele *Lap3-6* was recorded nearly exclusively from BRT, MED and from AUS (Fig. II.1e).

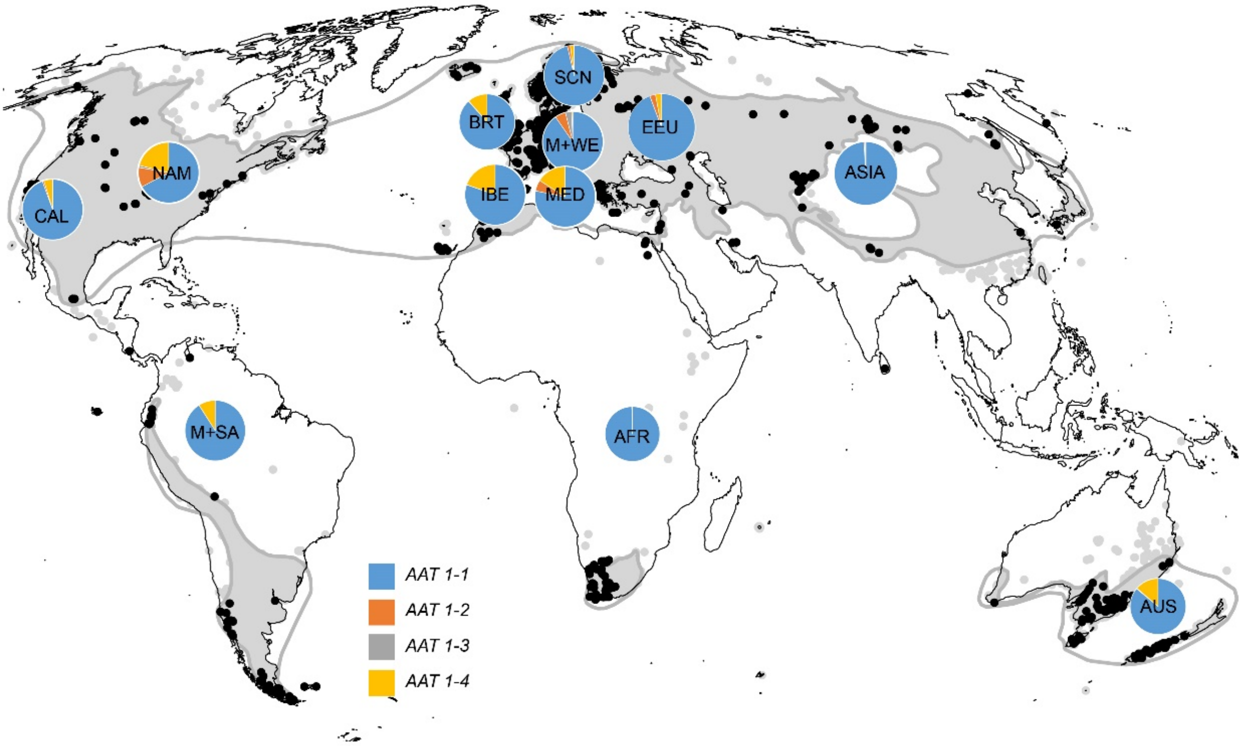

Figure II.1a: Allele frequencies of *Capsella bursa-pastoris* in different regions (*Aat1*). Black dots: sample locations, grey areas: distribution of *C. bursa-pastoris* based on data compiled by EW (CDH, 2018). IBE: Iberian Peninsula; BRT: British Isles; M+WE: Middle and Western Europe; MED: Circum-Mediterranean region; EEU: Eastern Europe; CAL: California; NAM: North America except California; M+SA: Middle and South America; AUS: Australasia; AFR: Africa.

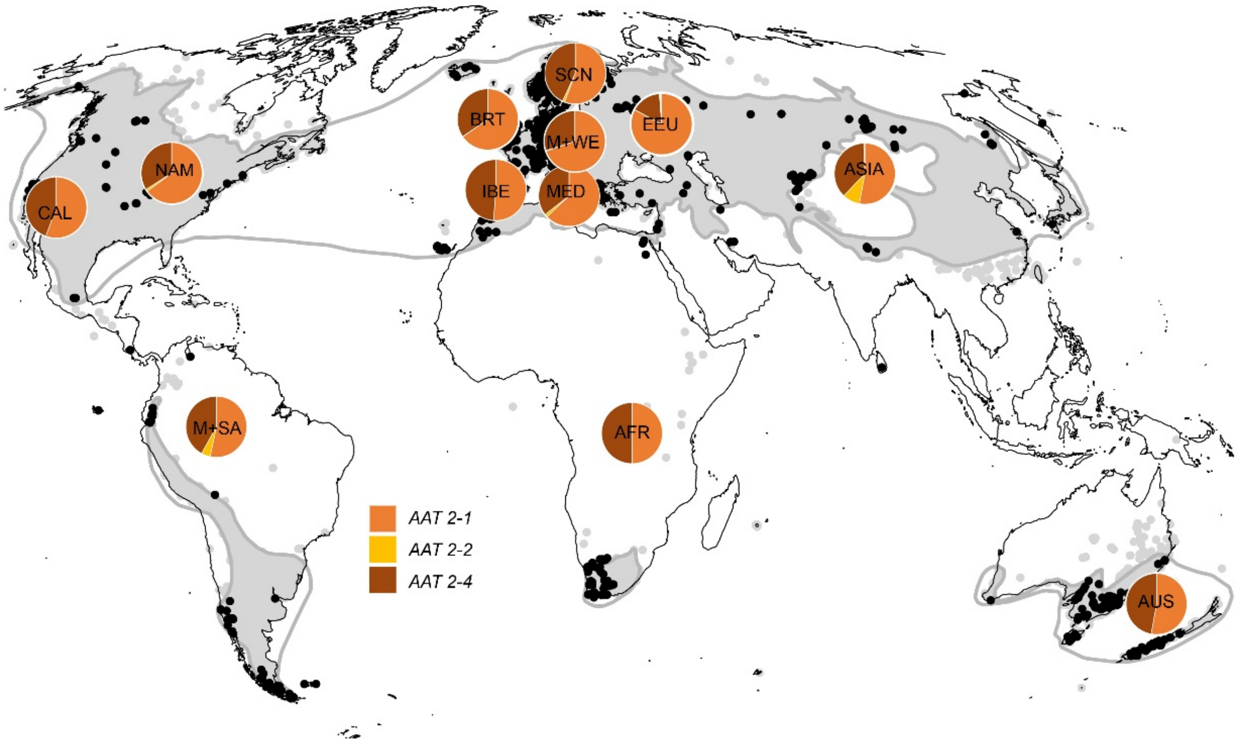

Figure II.1b: Allele frequencies of *Capsella bursa-pastoris* in different regions (*Aat2*).

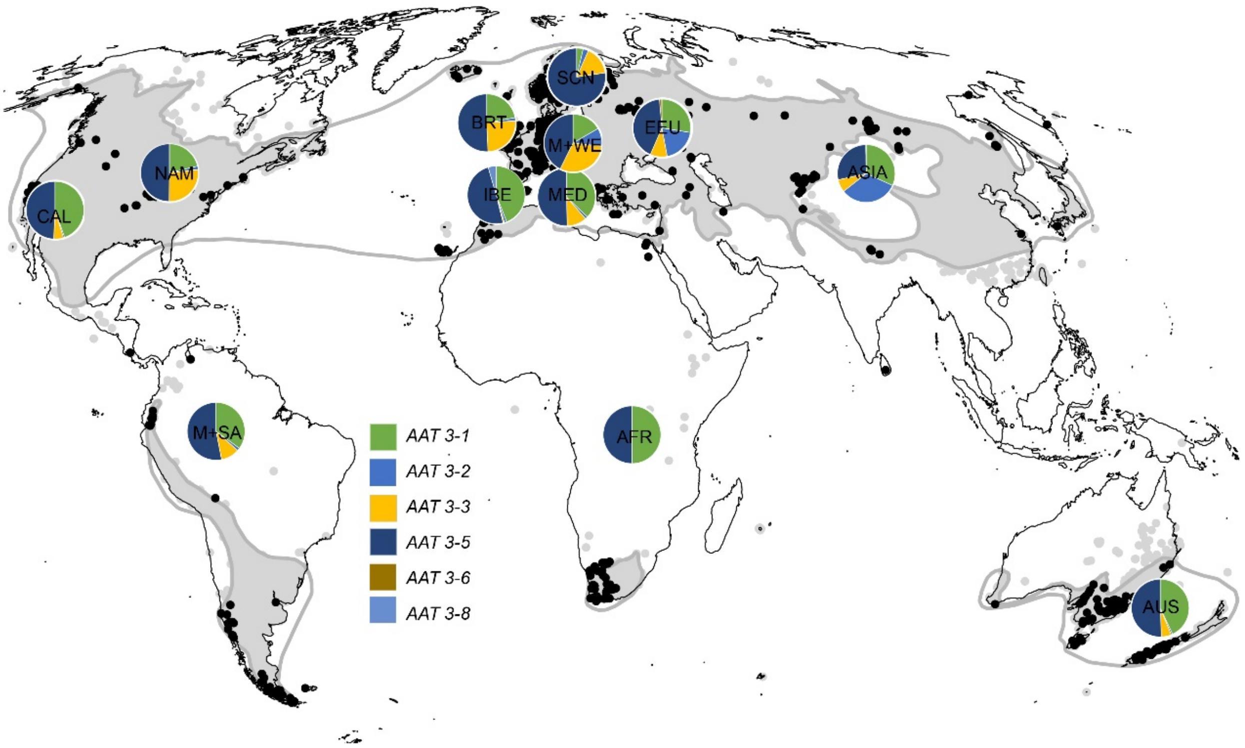

Figure II.1c: Allele frequencies of *Capsella bursa-pastoris* in different regions (*Aat3*). Black dots: sample locations, grey areas: distribution of *C. bursa-pastoris* based on data compiled by EW (CDH, 2018). IBE: Iberian Peninsula; BRT: British Isles; M+WE: Middle and Western Europe; MED: Circum-Mediterranean region; EEU: Eastern Europe; CAL: California; NAM: North America except California; M+SA: Middle and South America; AUS: Australasia; AFR: Africa.

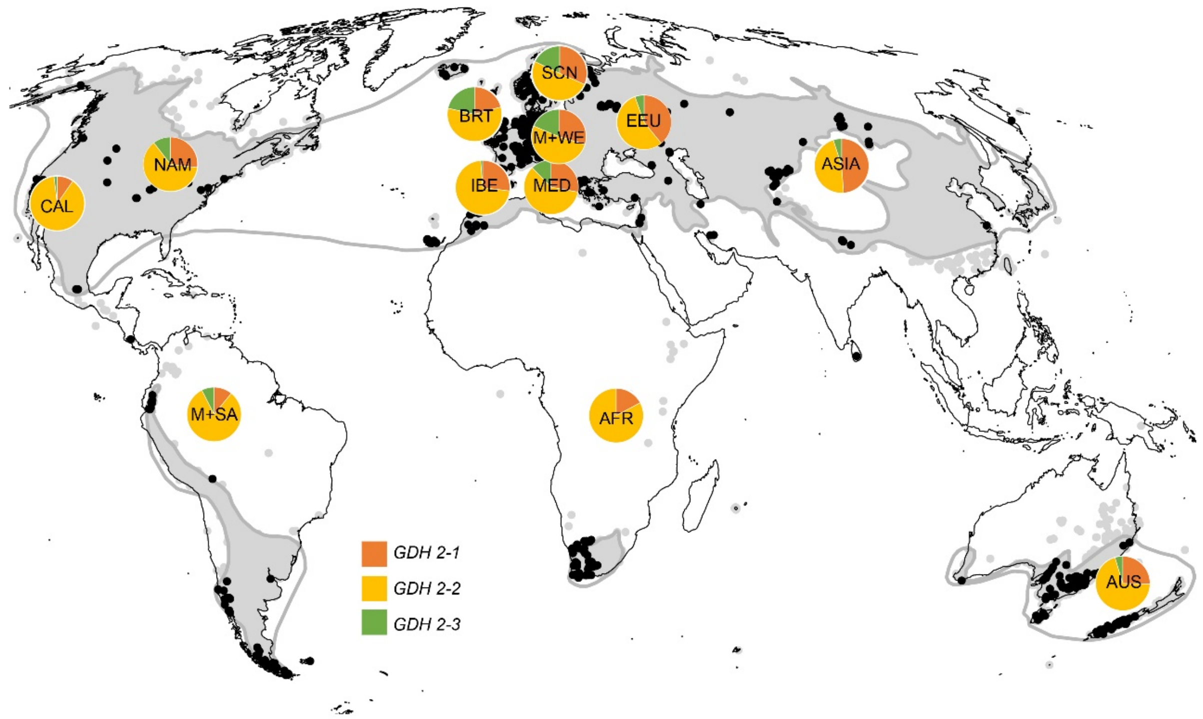

Figure II.1d: Allele frequencies of *Capsella bursa-pastoris* in different regions (*Gdh2*).

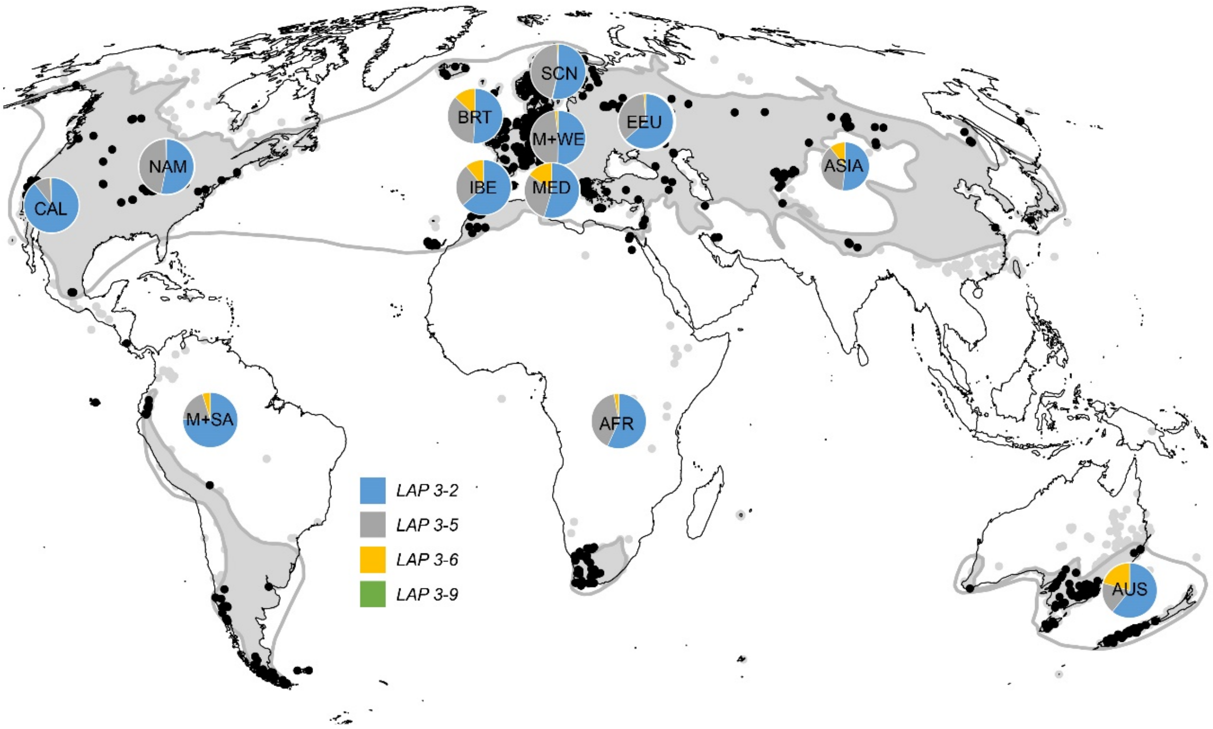

Figure II.1e: Allele frequencies of *Capsella bursa-pastoris* in different regions (*Lap3*). Black dots: sample locations, grey areas: distribution of *C. bursa-pastoris* based on data compiled by EW (CDH, 2018). IBE: Iberian Peninsula; BRT: British Isles; M+WE: Middle and Western Europe; MED: Circum-Mediterranean region; EEU: Eastern Europe; CAL: California; NAM: North America except California; M+SA: Middle and South America; AUS: Australasia; AFR: Africa.

Genotype biogeography

Capsella bursa-pastoris is tetraploid and thus comprises two whole genomes, genome A and genome B. Each of the single loci is doubled in *C. bursa-pastoris* and constitutes a locus pair with four alleles, two from genome A and two from genome B. Inheritance is disomic (see above). Since it is not possible to assign each of the four alleles of a locus pair unambiguously to one of the two loci of a pair, we recorded the presence or absence of the different alleles at each locus pair.

Loci and loci associations

We analyzed genotype frequencies at single loci and at different associations in total and in the geographical regions. Altogether, 1,851 different genotype combinations have been detected. Frequencies of given genotypes differ significantly between regions and are mostly rather low. Only 66 out of the 1,851 recorded genotypes show frequencies of 10 % and higher (hereafter referred to as "frequent genotypes"). It is obvious that frequent genotypes are preferentially shared by certain regions, e.g. IBE and MED with CAL and AUS, M+WE with SCN and EEU, and M+WE with IBE and MED. If we plot regional presence of genotypes irrespective of their frequencies, we observe differences in regional genotype diversity: Transforming the nominal scale into order statistics reveal an interesting rank order of the geographical regions (Fig. II.2). M+WE is the most diverse region (rank order 1) harvesting ca. 55 % of the total sum of genotypes. It is followed by EEU (order 2) with 35 % of the total, and ASIA, IBE, MED and SCN with 28 % - 25 %, approximately half that of M+WE and more or less equal between these regions. Genotype diversity within the introduced range is significantly lower than in the native ranges, displaying only 20 % and less of the total genotypes. AFR, with only 67 different genotypes, occupies the last position (rank order 12). A remarkable exception is BRT from the native range with the penultimate position (Fig. II.2).

The complete multilocus association

Out of all loci associations analyzed, we focus here on complete loci associations with the locus sequence *Aat1*, *Aat2*, *Aat3*, *Gdh1*, *Gdh2*, *Lap3*. A total of 8,076 individuals recorded in the native and non-native ranges displayed this complete multilocus combination (Tab. II.1 a+b). We detected 383 different genotypes at this multilocus, and only 18 of them had frequencies > 1 % out of which only one was frequent (f = 18 %, Tab. II.1 a+b). 5,658 individuals shared these common genotypes, whereas 2,418 individuals displayed rare genotypes. All of the 18 common complete multilocus genotypes were recorded from the native as well as the introduced ranges but with different frequencies between and within the different geographical regions (Tab. II.1 a+b). Of particular

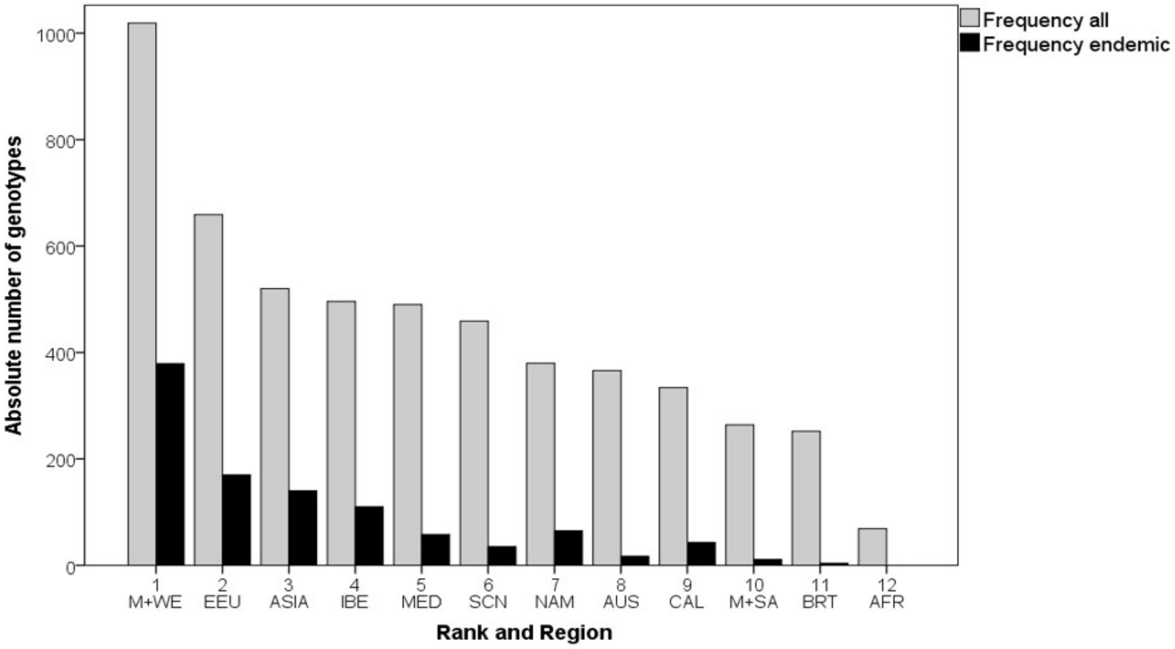

Figure II.2: Rank statistics of genotype diversity of *Capsella bursa-pastoris* within different regions. Order 1: highest diversity, order 12: lowest diversity. IBE: Iberian Peninsula; BRT: British Isles; M+WE: Middle and Western Europe; MED: Circum-Mediterranean region; EEU: Eastern Europe; CAL: California; NAM: North America except California; M+SA: Middle and South America; AUS: Australasia; AFR: Africa.

interest is the so-called Mediterranean Multilocus Genotype (MMG) with the composition *Aat1-1111*, *Aat2-1144*, *Aat3-1155*, *Gdh1-1111*, *Gdh2-2222*, *Lap3-2222* (Neuffer & Hoffrogge, 1999; Neuffer & Hurka, 1999). In the native range, the MMG occurs predominantly in the Iberian Peninsula, and in the introduced ranges with high frequencies in California where it is the most common genotype (Tab. II.1 a+b). It is also rather frequent in Middle and South America and in Australasia (Tab. II.1 a+b) and contributes remarkably to the set of multilocus genotypes in these regions (Tab. II.1 a+b).

'Endemic' genotypes

In each of the geographical regions, some of the recorded genotypes were 'endemic' which means they were not recorded in any other region (Fig. II.2). An exception is AFR where no 'endemic' genotypes were detected. Worth mentioning is also BRT with only four 'endemics' out of the total 250 genotypes (i.e. < 2 %). 'Endemism' is low also in M+SA, SCN and AUS (ca. 4 – 8 %). In the other regions, the percentage of 'endemics' fluctuates between 12 % and 37 %. While the number of genotypes found is positively related to the number of samples taken (Pearson's r = 0.62, p = 0.03), endemism calculated as the ratio of endemic to overall genotypes is unrelated to sampling intensity (Pearson's r = 0.52, p = 0.085).

Table II.1a: Frequencies of complete multilocus genotypes of *Capsella bursa-pastoris* corresponding between regions. The 18 most common genotypes of 383 are shown in detail (the rest summarized in „others“). Frequencies depicted with .000 are <.001. Entries with dash whenever allele not detected. *n*: number of individuals studied. *: Multilocus Mediterranean Genotype (MMG). IBE: Iberian Peninsula. BRT: British Isles. M+WE: Middle and Western Europe. MED: Circum-Mediterranean. SCN: Scandinavia. EEU: Eastern Europe. CAL: California. NAM: North America (except California). M+SA: Middle and South America. AUS: Australasia. AFR: Africa.

		Native							**Introduced**				
Genotype	***F***	**IBE**	**BRT**	**M+WE**	**MED**	**SCN**	**EEU**	**ASIA**	**CAL**	**NAM**	**M+SA**	**AUS**	**AFR**
Total	*n* = 8076	*n* = 1102	*n* = 165	*n* = 1641	*n* = 592	*n* = 468	*n* = 677	*n* = 281	*n* = 1203	*n* = 189	*n* = 423	*n* = 943	*n* = 392
	n = 5658	*n* = 770	*n* = 149	*n* = 1047	*n* = 309	*n* = 317	*n* = 302	*n* = 92	*n* = 993	*n* = 114	*n* = 356	*n* = 818	*n* = 391
18 most common genotypes	.701	.136	.020	.203	.073	.058	.084	.035	.149	.023	.052	.117	.049
111111441155/11112222/2222*	.182	.102	.001	.002	.012	-	.009	.001	.595	.033	.144	.067	.035
111111443355/11111122/2255	.068	.011	.078	.300	.137	.175	.007	.024	.119	.011	.061	.076	-
111111441155/11112222/2255	.051	.149	-	-	.104	-	.046	-	.005	-	.019	.214	.463
111111441155/11111122/2255	.049	.316	.023	.110	.030	.008	.055	.038	.005	.005	.010	.090	.311
111111113355/11112233/2255	.037	-	.003	.878	.003	.014	.034	-	-	.014	.051	.003	-
111111441155/11112233/2255	.036	.007	.007	.703	.106	-	.075	.007	.014	.014	.020	.048	-
111111113355/11111122/2255	.034	.004	.088	.474	.113	.073	.058	.011	.011	.055	.088	.026	-
111111445555/11111122/2255	.034	.004	.015	.096	.033	.577	.033	.173	-	.004	.033	.033	-
111111441155/11112222/2266	.031	.311	.071	.012	.146	-	-	-	.016	-	.020	.386	.039
111111443355/11112233/2255	.028	.004	.103	.429	.103	.004	-	.004	.027	.045	.103	.179	-
111111441155/11111122/2222	.025	.359	.030	.040	.010	-	-	.005	-	-	-	.510	.045
111111441155/11111122/2266	.025	.480	-	.005	.035	.015	.202	-	.071	-	.015	.167	.010
111111115555/11111122/2255	.021	-	.029	.157	-	.122	.634	.035	-	-	-	.023	-
114411441155/11111122/2266	.020	.037	.037	.012	-	-	-	-	-	-	-	.895	.019
114411441155/11112222/2266	.017	.385	.015	.148	.022	-	-	-	-	.030	.007	.393	-
111111111155/11111122/2255	.016	-	.038	.344	.015	.069	.282	.023	.107	.023	.008	.092	-
114411441155/11112222/2255	.014	.354	-	.080	.133	-	-	-	-	.142	.097	.195	-
114411441155/11111122/2255	.012	.796	-	.020	-	.020	.010	-	.020	-	-	.133	-
	n = 2418	*n* = 332	*n* = 16	*n* =594	*n* = 283	*n* = 151	*n* = 375	*n* = 189	*n* = 210	*n* = 75	*n* = 67	*n* = 125	*n* = 1
others	.299	.137	.007	.246	.117	.062	.155	.078	.087	.031	.028	.052	.000

Table II.1b: Frequencies of complete multilocus genotypes of *Capsella bursa-pastoris* corresponding within regions. The 18 most common genotypes of 383 are shown in detail (the rest summarized in „others“). Frequencies depicted with .000 are <.001. Entries with dash whenever allele not detected. *n*: number of individuals studied. *: Multilocus Mediterranean Genotype (MMG). IBE: Iberian Peninsula. BRT: British Isles. M+WE: Middle and Western Europe. MED: Circum-Mediterranean. SCN: Scandinavia. EEU: Eastern Europe. CAL: California. NAM: North America (except California). M+SA: Middle and South America. AUS: Australasia. AFR: Africa.

		Native							**Introduced**				
Genotype	***F***	**IBE**	**BRT**	**M+WE**	**MED**	**SCN**	**EEU**	**ASIA**	**CAL**	**NAM**	**M+SA**	**AUS**	**AFR**
Total	*n* = 8076	*n* = 1102	*n* = 165	*n* = 1641	*n* = 592	*n* = 468	*n* = 677	*n* = 281	*n* = 1203	*n* = 189	*n* = 423	*n* = 943	*n* = 392
	n = 5658	*n* = 770	*n* = 149	*n* = 1047	*n* = 309	*n* = 317	*n* = 302	*n* = 92	*n* = 993	*n* = 114	*n* = 356	*n* = 818	*n* = 391
18 most common genotypes	.701	.699	.903	.638	.522	.667	.446	.327	.825	.603	.842	.867	.998
111111441155/11112222/2222*	.182	.136	.006	.002	.029	-	.019	.004	.728	.259	.501	.105	.130
111111443355/11111122/2255	.068	.005	.261	.101	.128	.207	.006	.046	.055	.032	.080	.045	-
111111441155/11112222/2255	.051	.056	-	-	.073	-	.028	-	.002	-	.019	.094	.490
111111441155/11111122/2255	.049	.114	.055	.027	.020	.006	.032	.053	.002	.011	.009	.038	.316
111111113355/11112233/2255	.037	-	.006	.158	.002	.009	.015	-	-	.021	.035	.001	-
111111441155/11112233/2255	.036	.002	.012	.126	.052	-	.032	.007	.003	.021	.014	.015	-
111111113355/11111122/2255	.034	.001	.145	.079	.052	.043	.024	.011	.002	.079	.057	.007	-
111111445555/11111122/2255	.034	.001	.024	.016	.015	.335	.013	.167	-	.005	.021	.010	-
111111441155/11112222/2266	.031	.072	.109	.002	.063	-	-	-	.003	-	.012	.104	.026
111111443355/11112233/2255	.028	.001	.139	.059	.039	.002	-	.004	.005	.053	.054	.042	-
111111441155/11111122/2222	.025	.064	.036	.005	.003	-	-	.004	-	-	-	.107	.023
111111441155/11111122/2266	.025	.086	-	.001	.012	.006	.059	-	.012	-	.007	.035	.005
111111115555/11111122/2255	.021	-	.030	.016	-	.045	.161	.021	-	-	-	.004	-
114411441155/11111122/2266	.020	.005	.036	.001	-	-	-	-	-	-	-	.154	.008
114411441155/11112222/2266	.017	.047	.012	.012	.005	-	-	-	-	.021	.002	.056	-
111111111155/11111122/2255	.016	-	.030	.027	.003	.019	.055	.011	.012	.016	.002	.013	-
114411441155/11112222/2255	.014	.036	-	.005	.025	-	-	-	-	.085	.026	.023	-
114411441155/11111122/2255	.012	.071	-	.001	-	.004	.001	-	.002	-	-	.014	-
Others	*n* = 2418	*n* = 332	*n* = 16	*n* =594	*n* = 283	*n* = 151	*n* = 375	*n* = 189	*n* = 210	*n* = 75	*n* = 67	*n* = 125	*n* = 1
	.299	.301	.097	.362	.478	.323	.554	.673	.175	.397	.158	.133	.003

Population structure analysis

The pairwise calculated fixation indices (F_{ST}), a measure of population differentiation, revealed population differentiation between regions (Tab. II.2). The $\mathrm{F_{ST}}$ values were highest between ASIA and CAL (0.137) and AFR and NAM (0.132), and lowest between AUS and IBE (0.014) and CAL and IBE (0.016) (Tab. II.2).

The average number of different alleles at a locus varied among regions from 1.4 to 3.6 (mean across regions = 2.6), and the percentage of polymorphic loci per population ranged from 30 to 100, mean 81.7 (Table II.3). Observed heterozygosity H_o was zero in AFR or near zero (0.001) in BRT, M+SA and AUS (0.003 overall loci, Table II.4), and expected heterozygosity H_e overall loci was 0.23 (Table II.4). F_{is}, the degree of inbreeding within populations, was high (0.980 across populations varying from 0.947 to 0.99, Table II.4) as was the overall inbreeding coefficient F_{it} (ranging from 0.95 to 0.998, overall 0.982, Table II.4). The degree of population divergence F_{ST} varied from 0.088 to 0.186, and was 0.134 across loci (Table II.4).

A Mantel test calculated on the basis of pair-wise genetic and geographic distances revealed a regression coefficient of -0.019 (p = 0.001) for the total dataset. Based on this result, we can reject the null hypothesis that spatial and genetic distances are unrelated. Therefore, genetic distances increased with geographic distances and isolation by distance could be assumed, even if the slope is very low. Using reduced datasets for each region, correlation of genetic and geographic distances could also be observed. The regression coefficient was highest in SCN (0.484, p = 0.001) and EEU (0.241, p = 0.001), and lowest in IBE (0.042, p = 0.01) and BRT (0.069, p = 0.016).

Population admixture analysis revealed two clusters, and also identified a high number of mixed populations (Fig. II.4a). The bipartitioning of *C. bursa-pastoris* is confirmed via NMDS (Fig. II.6). The MMG is the determinant factor for this clustering, 86.2 % of the MMG's Aat component, 96.9 % of its Gdh component, and 92.4 % of the Lap component are assigned to cluster 2. In the native Eurasian range, cluster 1 (blue) is predominantly distributed in cold and temperate climate type regions, and cluster 2 (orange) in hot to warm and dry climate regions (Fig. II.4b). Bayesian analysis of population structure revealed genetic membership of populations of arid-semi-arid versus semi-humid-humid origin to different genetic clusters and a higher admixture of populations of regions with intermediate water balance to both clusters (Fig. II.5).

For the AMOVA, we used the the regions, the cluster affiliation as derived from population admixture analysis, and native and introduced environment as the grouping variables (Tab. II.5). Genetic variation among groups of populations was highest when partitioning the samples into two groups according to their cluster affiliation (27 %) and lowest according to whether samples were from native or introduced regions (11 %, Tab. II.5). Using the regions as group variable, sample size-corrected SSWP/n - 1 values differed significantly to some extend and detected ASIA as being

Table II.2: Pairwise population F_{ST} values of *C. bursa-pastoris* between regions. AFR: Africa. AUS: Australasia. BRT: British Isles. CAL: California. MED: Circum-Mediterranean. EEU: Eastern Europe. IBE: Iberian Peninsula. M+SA: Middle and South America. M+WE: Middle and Western Europe. NAM: North America (except California). SCN: Scandinavia.

	AFR	ASIA	AUS	BRT	CAL	MED	EEU	IBE	M+SA	M+WE	NAM	SCN
AFR	0.000											
ASIA	0.140	0.000										
AUS	0.063	0.107	0.000									
BRT	0.119	0.062	0.070	0.000								
CAL	0.060	0.137	0.037	0.091	0.000							
MED	0.068	0.058	0.038	0.021	0.063	0.000						
EEU	0.108	0.036	0.083	0.053	0.101	0.045	0.000					
IBE	0.044	0.106	0.014	0.082	0.040	0.054	0.090	0.000				
M+SA	0.067	0.109	0.045	0.065	0.016	0.039	0.076	0.048	0.000			
M+WE	0.129	0.072	0.108	0.034	0.124	0.033	0.052	0.119	0.082	0.000		
NAM	0.132	0.104	0.053	0.046	0.066	0.027	0.057	0.064	0.053	0.045	0.000	
SCN	0.114	0.039	0.097	0.037	0.127	0.051	0.049	0.096	0.099	0.072	0.076	0.000

Table II.3: Measurements of genetic variation within *C. bursa-pastoris* regions. *n*: number of individuals per region; N_a: average number of different alleles at a locus; s N_a: standard error of N_a; *P*: percentage of polymorphic loci; N_e: effective number of alleles; s N_e : standard error of N_e; H_o: observed heterozygosity; s H_o: standard error of observed heterozygosity; H_e: expected heterozygosity; s H_e : standard error of H_e.

Region	*n*	N_a	s N_a	*P*	N_e	s N_e	H_o	s H_o	H_e	s H_e
IBE	1102	3.0	0.30	100	1.428	0.193	0.002	0.001	0.213	0.073
BRT	165	2.2	0.33	70	1.391	0.134	0.001	0.001	0.224	0.068
M+WEU	1641	3.6	0.37	90	1.572	0.185	0.004	0.002	0.290	0.073
CIR	592	2.9	0.28	90	1.491	0.139	0.004	0.003	0.278	0.064
SCN	468	2.3	0.26	90	1.378	0.192	0.003	0.002	0.202	0.060
EEU	677	3.2	0.44	100	1.550	0.166	0.005	0.003	0.291	0.070
ASIA	281	3.0	0.37	80	1.663	0.238	0.006	0.004	0.301	0.081
CAL	1203	2.4	0.16	100	1.222	0.070	0.006	0.002	0.158	0.045
NAM	189	2.3	0.34	70	1.574	0.160	0.006	0.003	0.298	0.075
M+SA	423	2.1	0.23	80	1.362	0.131	0.001	0.001	0.212	0.064
AUS	943	2.4	0.31	80	1.434	0.187	0.001	0.000	0.219	0.074
AFR	392	1.4	0.22	30	1.134	0.092	0.000	0.000	0.080	0.053
Total	8076	2.6	0.1	81.7	1.433	0.047	0.003	0.001	0.232	0.019

the most diverse population, followed by EEU and M+WE (Fig. II.3). Among the least diverse populations were CAL and AFR (Fig. II.3).

Furthermore, in SDM, the feature combination of LQP and a regularization value of 1.1 was obtained as best configuration option to avoid overfitting. The distribution models for the two cluster solution proposed by the different classification approaches resulted in a quite distinct geographical pattern (Fig. II.7), overlapping only in the high-oceanic regions of Western Europe, Chile/Argentina and Tasmania/New Zealand. Cluster 1 (CL-01) tends to have a more humid-temperate distribution, while Cluster 2 (CL-02) is situated in semiarid-mediterranoid regions, yet when modelled based on macroclimatic data, this pattern gets even more distinct.

Table II.4: Measurements of genetic variation and *F*-statistics. *n*: number of alleles, N_e: effective number of alleles; H_o: observed heterozygosity; H_e: expected heterozygosity; F_{is}: inbreeding coefficient; F_{it}: overall inbreeding coefficient; F_{ST}: degree of population divergence.

Locus	*n*	N_e	H_o	H_e	Fixation index	F_{is}	F_{it}	F_{ST}
AAT1	2.209	1.213	0.003	0.131	0.976	0.947	0.95	0.088
AAT2	2.292	1.286	0.001	0.169	0.993	0.99	0.991	0.121
AAT3	3.292	1.667	0.007	0.302	0.978	0.975	0.979	0.186
GDH2	2.417	1.590	0.001	0.336	0.998	0.998	0.998	0.154
LAP3	2.625	1.410	0.004	0.213	0.980	0.989	0.992	0.119
Mean	2.567	1.433	0.003	0.230	0.986	0.980	0.982	0.134

Table II.5: Results of the analyses of molecular variance (AMOVA) of *C. bursa-pastoris*. *Df*: degrees of freedom; *SS*: sums of squares; *%*: percentage of variance. All *p*-values were 0.01.

Source of variation	d.f.	SS	Variance	%
Regional analysis				
Among populations	11	4297.875	0.301	21 %
Within populations	16140	18616.527	1.153	79 %
Total	16151	22914.402	1.454	100 %
Cluster affiliation				
Among populations	1	3496.531	0.438	27 %
Within populations	16150	19418.699	1.202	73 %
Total	16151	22915.230	1.640	100 %
Native vs. introduced				
Among populations	1	1258.671	0.164	11 %
Within populations	16150	21655.731	1.341	89 %
Total	16151	22914.402	1.505	100 %

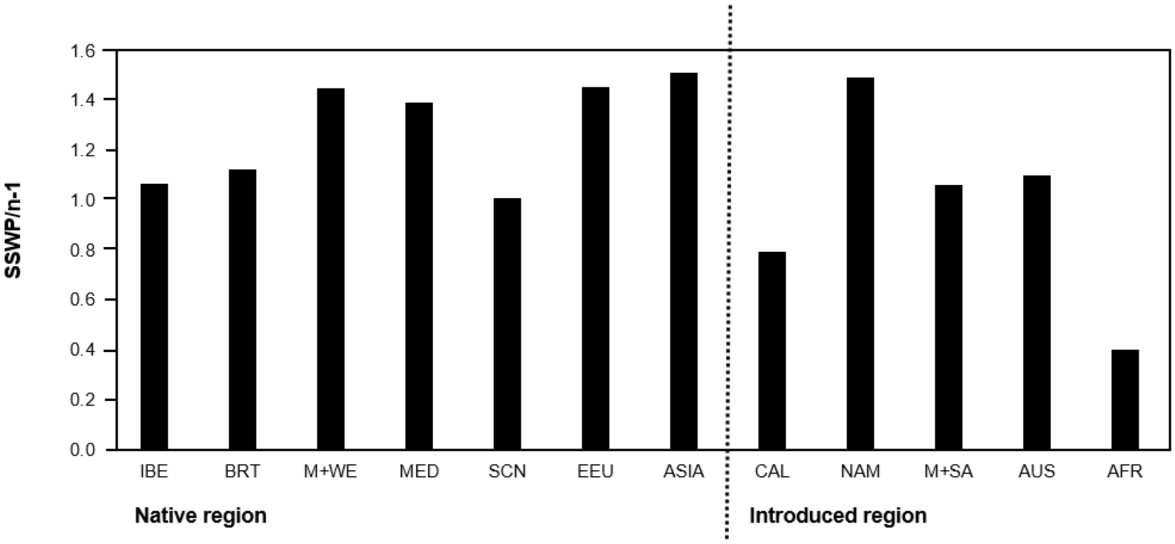

Figure II.3: SSWP/*n* - 1 diversity values of *Capsella bursa-pastoris* populations within the native and introduced regions. AMOVA tested for differences among groups. IBE: Iberian Peninsula; BRT: British Isles; M+WE: Middle and Western Europe; MED: Circum-Mediterranean region; EEU: Eastern Europe; CAL: California; NAM: North America except California; M+SA: Middle and South America; AUS: Australasia; AFR: Africa.

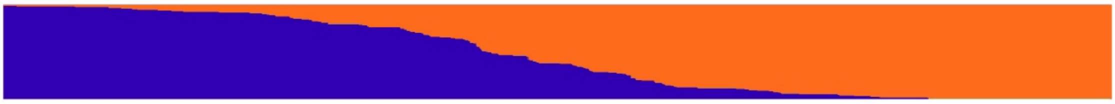

Figure II.4a: Population structure analysis of *Capsella bursa-pastoris.* The two clusters and their admixture. Blue: cluster 1; Orange: cluster 2.

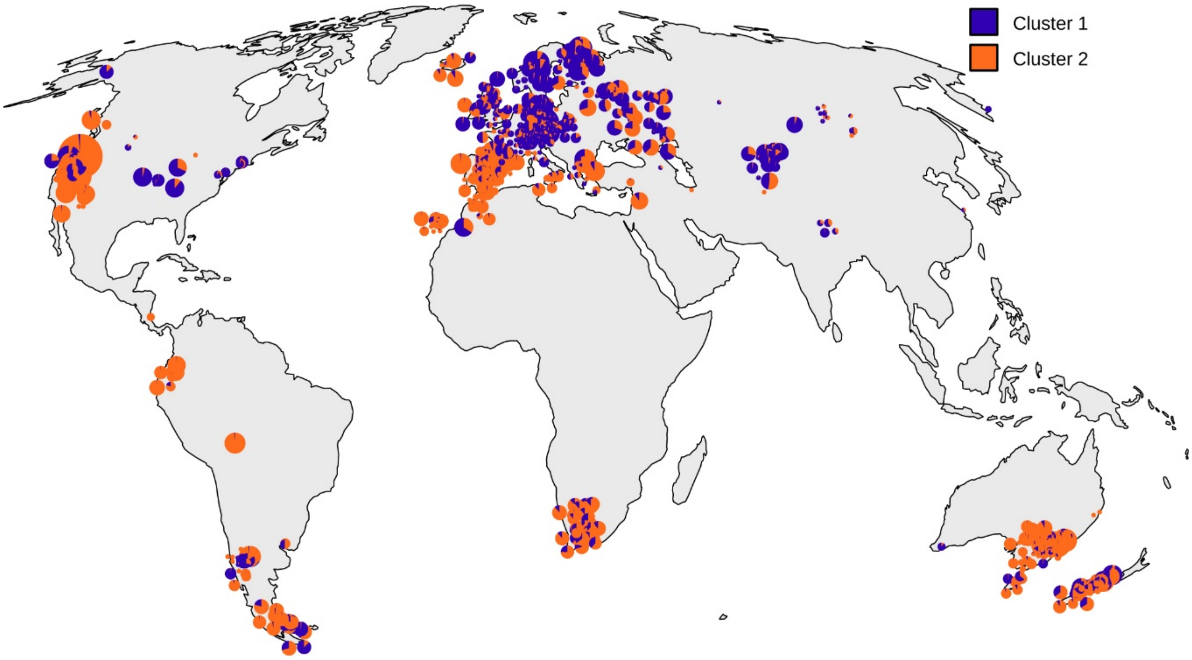

Figure II.4b: Population structure analysis of *Capsella bursa-pastoris.* Cluster affiliation of analyzed populations. Pie chart diameters refer to number of sampled individuals. Blue: cluster 1; Orange: cluster 2.

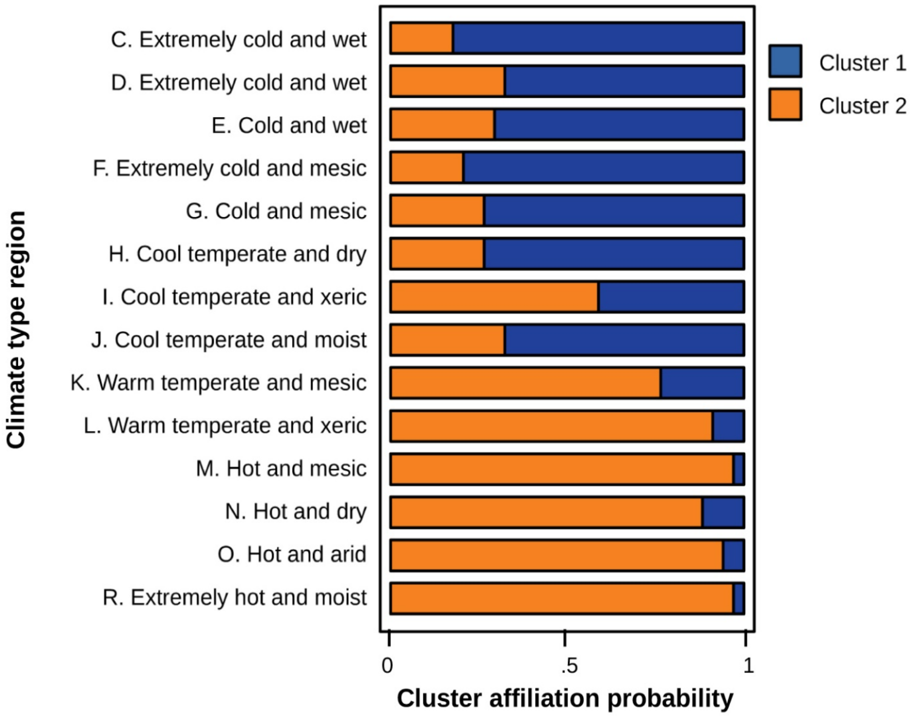

Figure II.5: Cluster affiliation of *Capsella bursa-pastoris* assigned to GEnS (Metzger et al., 2013) climate type regions. Blue: cluster 1, Orange: cluster 2.

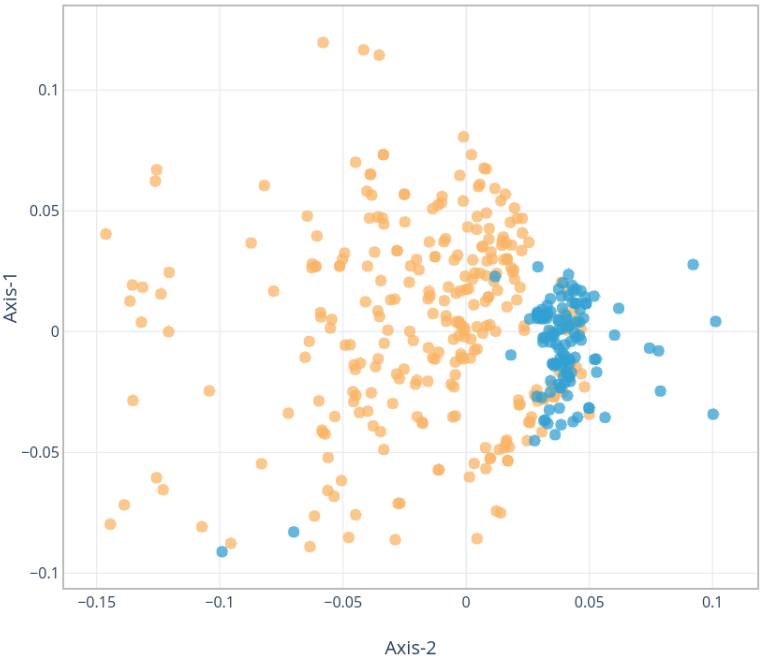

Figure II.6: NMDS plot confirming the division into two populations within *C. bursa-pastoris*.

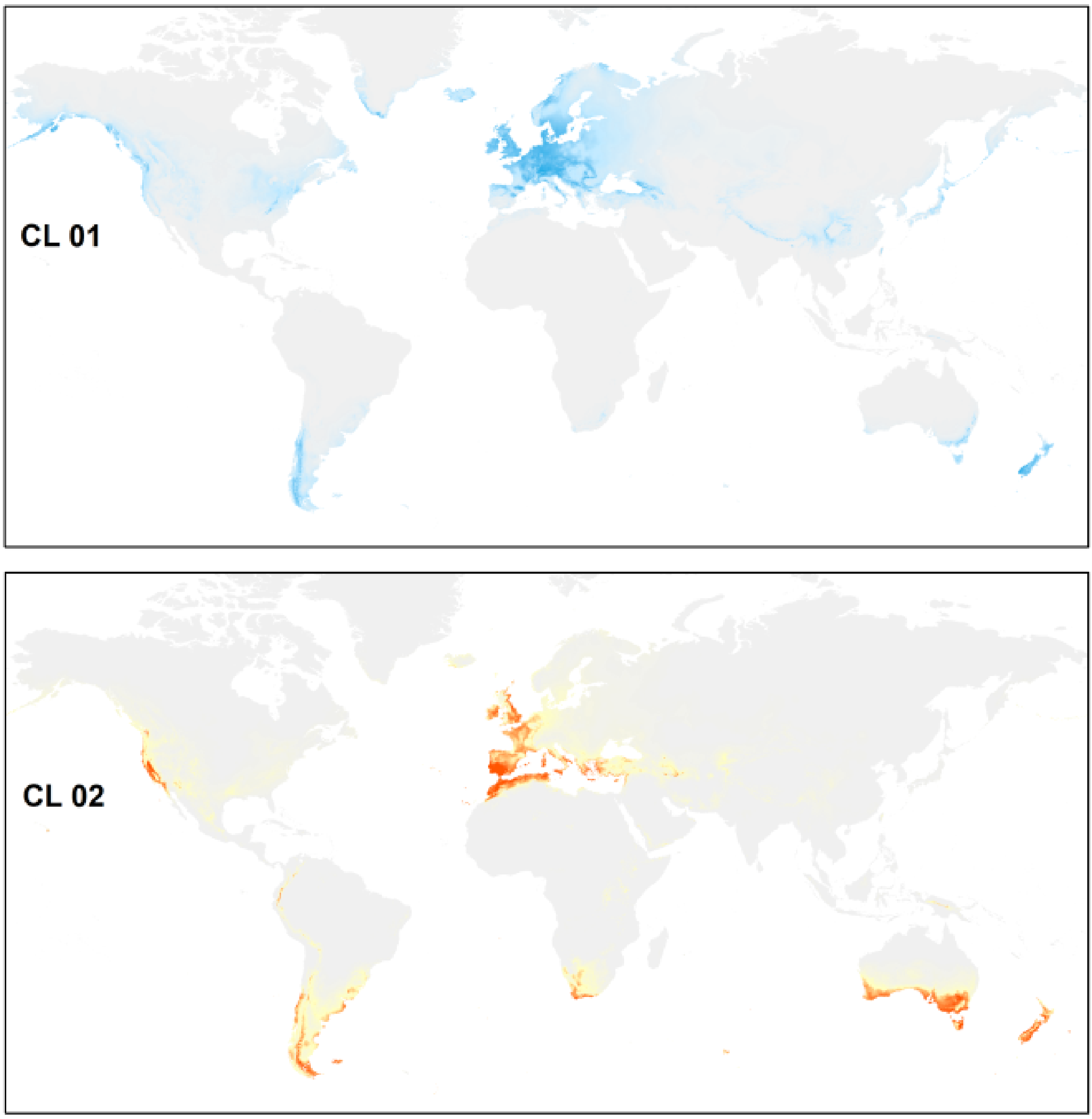

Figure II.7: Distribution models of the macroclimatic preference of the two STRUCTURE clusters projected onto today's climate.

Discussion

For approximately 500 years, vascular plant species have been migrating in large numbers and at high rates between and within continents, mainly caused by human activities. Introduction dynamics and patterns of migration are mostly known in general terms only. Detailed knowledge of colonization history and migration patterns can be detected by analysis of historical records, molecular evidence and statistical evaluation, as will be shown for *Capsella bursa-pastoris*. First, we will shortly outline the history of worldwide weed dispersal by European colonists. We will analyze the global genetic diversity patterns of *C. bursa-pastoris* and discuss the invasion process in terms of colonization history and adaptation.

Weed introduction by European settlers

The history of weed introduction into the New World, South Africa and Australasia is incidental to European colonization activities.

Middle and South America: During the 16th century, the Spanish Crown conquered Middle America and large parts of South America. Later, Portugal and the Dutch Crown occupied large parts of eastern South America. The main immigration to Patagonia was not before 1840, mainly by immigrants from the British Isles, Scandinavia and from southern and eastern Europe (Neuffer et al., 1999). Already by 1600, the weed flora of Mexico was mainly Eurasian with Mediterranean plants predominating (Crosby, 1986). The first *Capsella* herbarium record from Patagonia is from 1877 (Neuffer et al., 1999).

North America: European weeds seem to have already established themselves by the first half of the 17th century, including Shepherd's Purse (Crosby, 1986). California remained a remote region until the end of the 18th century, when the Spanish Crown founded missions along the Pacific coast. This resulted in the introduction of a Mexican weed flora (Mediterranean). With the great hordes of gold seekers, weeds from temperate North American and European regions were brought to California (Neuffer & Hurka, 1999). By 1860, at least 90 alien weeds were naturalised in California (Robbins, 1940).

South Africa: In 1652, The Dutch East India Company established a permanent settlement at the Cape, which was the beginning of Cape Town and the Cape Colony. About 1800, the Cape Colony became a British colony. Little is known about the introduction history of European weeds but Thunberg during his stay in the Cape Colony from 1772 – 1775, reported *Capsella bursa-pastoris* (Marais, 1970). In 1860, Sonder in the Flora Capensis (Harvey & Sonder, 1860), listed *C. bursa-pastoris* as a common weed throughout the colony, introduced from Europe.

Australasia: From around 1800, people from the British Isles began to found farms in southeastern Australia (Lamping, 1985). In the 19th century, people from Mediterranean countries immigrated to Australia and some of them settled on farms. The 19th century is probably the time when *Capsella* populations established themselves in Australia. The first herbarium records from South Australia are from 1847 (Kloot, 1983). In 1841, New Zealand became a British crown colony and, since then, colonization activities have been intensified, especially sheep farming in the Southern Island. However, European weeds were widespread already before 1840 (Crosby, 1986).

Biogeography of genetic diversity

Genetic diversity within *Capsella bursa-pastoris* is clearly geographically structured and the *F*-statistics (Tab. II.4) depict a high degree of selfing. The observed heterozygosity H_0 was near zero, and the inbreeding coefficients F_{is} and F_{it} amounted to almost 1.0 indicated a global deficit of heterozygotes (Tab. II.4). The outcrossing rates, estimated by the *F*-values (fixation index), $t = (1 - F)/(1 + F)$ (Brown and Weir, 1983), vary between 0 – 1 %. Percentage of polymorphic loci (= 82), average number of different alleles per locus (= 2.6), and effective number of alleles per locus (N_e = 1.4) lie within the range of typical annual selfers (Hamrick & Godt, 1990), but the expected heterozygosity (genetic diversity index; H_e = 0.2) indicates a mixed mating system (Hamrick & Godt, 1990) (Tab. II.3). Similar values were achieved in previous allozyme analyses with *C. bursa-pastoris* (Neuffer et al., 2011). Polymorphisms within populations expressed by the percentage of polymorphic loci varied between populations from 30 up to 100 (Tab. II. 3), and the degree of population divergence is rather low (F_{ST} = 0.134, Tab. II.4).

Phylogeographic structure was evident as shown by the isolation by distance. To exclude possible sample errors, we concentrated on the more common alleles to illustrate allele frequencies (Fig.II.1). They are distributed worldwide indicating that all of the common alleles have been introduced from the native into the non-native regions (Fig. II.1). Allozyme diversity in the native range is more pronounced in western Eurasia than in eastern Eurasia (Fig. II.1). Some of the allozymes are distributed more or less evenly throughout the world (Fig. II.1). However, the frequency of the majority of the common alleles varies conspicuously between the geographical regions (Fig. II.1, Fig. II.3, and Results). Source populations are often located in M+WE, MED, IBE, and BRT (Fig. II.1). Remarkably, we found similar relative frequencies of alleles between M+WE and MED (e.g., *Aat1-4*) and CAL and AUS, between IBE and CAL (e.g., *Aat2-4*), GBI and NAM and AUS (e.g., *Gdh2-3*). F_{ST} values indicate commutation rather between BRT and M+WE and between MED and M+WE than between CAL and M+WE and between CAL and SCN.

Rather rare alleles can provide valuable information also. The overall frequency of allele *Aat1-3* amounts to only 1.3 % but the occurrence of the allele is concentrated in M+WE (93 %) in the native range and 4 % in NAM and 2 % in AUS in the introduced range. It is either missing in all other regions or is present with an overall frequency of less than 1 % (Fig. II.1).

It appears that allozyme frequencies reflect, to some extent, the history of distribution areas. This becomes more obvious when comparing the frequency of isozyme genotypes instead of alleles alone as has been outlined in the results. Genotypes frequent in the native Mediterranean regions IBE and MED are preferentially shared with the colonized ranges CAL and AUS. Frequent genotypes (frequency > 10 %) in the temperate native range (M+WE, SCN, EEU) are shared with native Mediterranean regions and with both, temperate and Mediterranean regions in the introduced ranges. These distribution patterns argue for intercontinental introduction routes from native Mediterranean and temperate regions into the colonized continents New World, Africa and Australasia which is in agreement with the respecting colonizing history (see above).

The geographical distribution of genotype diversity within the native Eurasian range is surprising. Genotype diversity was highest in M+WE and EEU but very low in BRT (only 25 % of

M+WE) and more-or-less half of that of M+WE and ASIA, IBE, MED and SCN (Fig. II.2). Similar results were apparent when observing the molecular variance (Fig. II.3). This pattern can be explained by intracontinental migration routes, assuming two centres of initial diversity, namely nemoral Asia and the Mediterranean region. Migrations from east to west and from south to north, probably in post-glacial times, overlapped in M+WE and EEU and thus enriched genotype diversity in these regions. The British Isles, because of their geographically isolated position, only received part of the diversity. This scenario is supported by the population structure analyses which showed admixture of the two clusters in continental Europe (Fig. II.4b).

Two lineages within *Capsella bursa-pastoris*

It appears that *Capsella bursa-pastoris* is split into two lineages, or clusters, one occurring in mediterranean climate regions and the other occurring in temperate climate regions. Crossing experiments between populations with typical Mediterranean and Temperate genotypes indicated, that success rate of crossing is restricted and even failed in the case where the mother plant was of the Eurasian type (Linde, 1999, unpublished), leading to the assumption that there is some incompatibility between them. First insights into the mechanism and establishment of gene-flow barriers between two diploid *Capsella* species have recently been provided by Sicard et al. (2015) and involve differences in petal size and flower opening, both with a complex genetic basis. One may hypothesise similar mechanisms at work between the two *C. bursa-pastoris* lineages.

Shull (1929) described a new species, *Bursa* (= *Capsella*) occidentalis, which, however, was never recognised let alone accepted. Plants belonging to this taxon are early flowering and display some leaf characteristics varying from other *Capsella* provenances. Shull recorded it from Arizona, California, Hawaii, Peru, Chile, Argentina and Uruguay, and noted, "how closely the range of this species […] agrees with the region occupied by the Spanish settlers in America. It seems probable that there is a causal relation between these two distributions." Apparently, Shull's *C. occidentalis* belongs to the Mediterranean lineage shown here.

Arabidopsis thaliana also displays several intercrossing lineages (Durvasula et al., 2017), but introduced North American plants mostly belong to a single haplogroup, which could be due to some adaptive advantage, or be the result of being derived from one of the first arrivals (Exposito-Alonso et al., 2018).

The two lineages within *C. bursa-pastoris* might point to multiple origins of the polyploid *C. bursa-pastoris*. It has already been argued that *C. bursa-pastoris* originated in the Eurasian steppe belt (Hurka et al., 2012) whereas Slotte et al. (2006) and Douglas et al. (2015) discuss an east Mediterranean origin. However, the two lineages might also be the result of an early diversification after the origin of *C. bursa-pastoris*, which we think is more parsimonious.

Colonization and adaptation

Is the colonizing success of *C. bursa-pastoris* based on the introduction of pre-adapted genotypes, or on selection for adaptive genetic variation after the introduction? All common multilocus genotypes detected in Europe were also recorded in the introduced continents (Tab. II.2). The variable European *Capsella* gene pool was nearly completely introduced into the other continents. This provides evidence for multiple introductions instead of rearrangements of a single or few introduced genotypes in the newly colonized regions as was argued for Avena barbata (Pérez de la Vega et al., 1991; Allard et al., 1993). However, the Temperate and the Mediterranean lineage are affiliated with climate parameters and may reflect adaptive value despite the fact that isozymes are selection neutral markers. The invasion of Mediterranean ecosystems by the Mediterranean *Capsella*-lineage (Fig. II.4b) is strong evidence for canalization of the invasion process by natural selection. California is a good example (Neuffer & Hurka, 1999). Specific features of Mediterranean-climate ecosystems which allow some plants and not others to pass through the invasion stage filters (introduction, colonization and naturalization) include disturbance and the interaction between soil moisture levels and temperature (Groves, 1986). It has been shown in previous *Capsella* studies that variation in isozymes is correlated with detectable ecological important life history traits, such as flowering time and growth form parameters (Neuffer & Hoffrogge, 1999; Neuffer & Hurka, 1999). This correlation can be explained, at least partly, by linkage of isozyme loci to life history traits. Linde et al. (2001) found three major QTL controlling flowering time differences among ecotypes, which are linked to isozyme loci. These linkage groups correspond to single chromosomes. In addition, due to the predominantly selfing breeding system of *C. bursa-pastoris* (highly selfing but outcrossing rates up to 12 % have been reported; Shull, 1929; Hurka et al., 1989), gene combinations will stay together even if they are not located on the same chromosome.

Multiple introductions are very common features of successful invasions (Bossdorf et al., 2005, Dlugosch & Parker, 2008). While some successful colonizers arrive well-suited to new environments, the success of others appears to depend on rapid local adaption (Bock et al., 2015). Populations adapt to novel environments in two ways: selection on pre-existing standing variation, and selection on new, *de novo* mutations. One source of standing variation in the introduced range is admixture, the mixing of historically isolated gene pools (Dlugosch et al., 2015). It is the result of multiple introductions and introgression among diverse genotypes from structured populations in the native range thereby generating heterozygosity (Keller, 2014). In our *Capsella* study, admixture between the Temperate and Mediterranean lineages has been demonstrated (Fig. II.4a), but whether this is a significant source of new standing variation in the introduced range is questionable. Given the high probability of spatial population admixture due to abundantly repeated re- and cross-introductions, this seemingly resilient and relatively stable pattern points to a certain environmental filtering/selection acting at establishment and survival of introduced genotypes. Taken together with the genetical data obtained, a split into two "main clades" seems probable also from a macro-ecological point of view. The summer-dry, warm and partly semiarid climate niche of Cluster 2 (CL-02; fig. II.7) might have supported the winter-annual lifecycle of *C. bursa-pastoris* earlier, before agriculture provided suitable habitats in the geographical range of Cluster 1 (CL-01; fig. II.7). Nevertheless, in nearly all regions, we recorded 'endemic' genotypes (Fig. II.2). They may be the outcome of admixture or may be *de novo* mutations, but this cannot yet be determined. Since the degree of endemism calculated as the ratio of endemic to overall genotypes is unrelated to sampling intensity (Pearson's r = 0.52, p = 0.085), sampling bias seems unlikely. The generally relative low number of (frequent) genotype endemism in regions of post-Columbian colonization points to a limited importance of new genotypes.

Little work has been done with respect to the role of *de novo* mutation in "invasion genetics". In *Arabidopsis thaliana*, *de novo* mutations in a colonizing lineage in North America were detected, but their adaptive value remains open (Exposito-Alonso et al., 2018). It seems that natural selection in invaders relies mainly on standing variation (Bock et al., 2015).

Conclusion

In *Capsella bursa-pastoris*, allozymes and isozyme genotypes are not randomly distributed, neither in the Eurasian source continent, nor in the introduced regions. Genetic variation at isozyme level is clearly geographically structured and is split into two lineages, one distributed predominantly in Mediterranean climate regions, the other predominantly in temperate climate regions. The distribution pattern of these lineages in native Eurasia can be explained by the evolutionary history of *C. bursa-pastoris* and intracontinental migration in pre-historic times, whereas intercontinental migration in historic times explains the geographical patterns in the introduced ranges. However, environmental filtering due to climate pre-adaptation seems also to be involved. The global biogeography of genetic variation of *C. bursa-pastoris* mirrors the colonization histories and is in accordance with the history of weed introduction into the continents. We have been able to reconstruct invasion routes and to identify source areas. Multiple independent introductions of genotypes from different sources and climate regions are obvious. We can conclude that most colonizing *Capsella* genotypes were pre-adapted and found their respective matching niches in the colonized ranges.

It would be highly interesting to see whether the global genetic variation pattern at the isozyme level can be corroborated or even improved in resolution by employing other molecular markers.

Acknowledgments

We are grateful to the Botanical Garden of the Osnabrück University for growing plants, and we thank Claudia Gieshoidt, Rudolf Hungerland-Grupe and Mechthild Wickenbrock for conducting isozyme analyses, and Claudia Gieshoidt also for data input and database management. Many thanks to all colleagues who helped collecting seed material. We thank Lucille Schmieding for correcting the English text. Financial support by the German Research Foundation DFG and the German Academic Exchange Service DAAD is greatly acknowledged.

Literature Cited

Allard, R. W., Garcia P., Saenz-de-Miera, L. E., Pérez de la Vega, M. (1993). Evolution of multilocus genetic structure in *Avena hirtula* and *Avena barbata*. *Genetics*, *135*(4), 1125-1139.

Almquist, E. (1907). Studien über die *Capsella bursa-pastoris* (L.). *Acta Horti Bergiani*, *4*(6), 1–92.

Almquist, E. (1921). Studien über die *Capsella bursa-pastoris* II. *Acta Horti Bergiani*, *7*(2), 41–95.

Baker, H. G., Stebbins, G. L. (Eds.), (1965). *The Genetics of Colonizing Species*. New York: Academic Press.

Barrett, S. C. H. (2015). Foundation of invasive genetics: The Baker and Stebbins legacy. *Molecular Ecology*, *24*(9), 1927-1941.

Barrett, S. C. H., Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology and Evolution*, *23*(1), 38-44.

Bock, D. G., Caseys, C., Cousens, R. D., Hahn, M. A.,Heredia S. M., … Rieseberg, L. H. (2015). What we still don't know about invasion genetics. *Molecular Ecology*, *24*(9), 2277-2297.

Bossdorf, O., Auge, H., Lafuma, L., Rogers, W. E., Siemann, E., Prati, D. (2005). Phenotypic and genetic differentiation between native and introduced plant populations. *Oecologia*, *144*(1), 1-11.

Brown, A.H.D., Weir, B.S. (1983). Measuring genetic variability in plant populations. In: Tanksley, S.D., Ortin, T.J. (Eds.), *Isozymes in Plant Genetics and Breeding, Part A*. Elsevier, Amsterdam, pp. 219–239.

CDH (2018). Chorology Database Halle, https://www.botanik.uni-halle.de/chorologie/?lang=en

Cristescu, M. E. (2015). Genetic reconstructions of invasion history. *Molecular Ecology*, *24*(9), 2212-2225.

Colautti, R. I., Lau, J. A. (2015). Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. *Molecular Ecology*, *24*(9), 1999-2017.

Coquillat, M. (1951). Sur les plantes les plus communes a la surface du globe. *Bulletin Mensuel de la Société Linnéenne de Lyon*, *20*(20), 165–170.

Crosby, A. W. (1986). *Ecological imperialism. The biological expansion of Europe, 900-1900*. Cambridge: Cambridge Univ. Press.

Dlugosch, K. M., Parker, I. M. (2008). Founding events in species invasions: Genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology*, *17*(1), 431-449.

Dlugosch, K. M., Anderson, S. R., Braasch, J., Cang, A., Gillette, H. D. (2015). The devil in the details: genetic variation in introduced populations and its contribution to invasiveness. *Molecular Ecology*, *24*(1), 2095-2111.

Durvasula, A., Fulgione A., Gutaker, R. M., Alacakaptan, S. I., Flood, P. J., Neto, C., … & Hancock, A.M. (2017). African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America*, *114*(20), 5213-5218.

Douglas, G., Gos, G., Steige, K., Salcedo, A., Holm, K., Ågren, J. A., … & Wright, S. (2015). Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris. Proceedings of the National Academy of Sciences of the United States of America*, *112*(9), 2806–2811.

Earl, D. A. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359-361.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., ... & Li, J. (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography, *29*(2), 129-151.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, *14*(8), 2611-2620.

Excoffier, L., Smouse, P. E., Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, *131*(2),479–491.

Exposito-Alonso, M., Becker, C., Schuenemann, V. J., Reiter, E., Setzer, C., Slovak, R., ... & Busch, W. (2018). The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genetics*, *14*(2), e1007155.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allelefrequencies. *Genetics*, *164*(4), 1567-1587.

Fischer, M., Matthies, D. (1998). RAPD variation in relation to population size and plant fitness in the rare *Gentianella germanica* (Gentianaceae). *American Journal of Botany*, *85*(6), 811–819.

Gao, H., Williamson, S., & Bustamante, C. D. (2007). An MCMC approach for joint inference of population structure and inbreeding rates from multi-locus genotype data. *Genetics*, *176*(3), 1635-1651.

Groves, R. H. (1986). Invasion of mediterranean ecosystems by weeds. In Hopkins, A. J. M. & Lamont, B.B. (Eds.), *Resilience in Mediterranean Ecosystems* (pp. 129-145). Springer, Dordrecht.

Hamrick, J. L., Godt, M. J. W., (1990). Allozyme diversity in plant species. In: Brown, H.D., Clegg, M.T., Kahler, A.L., Weir, B.S. (Eds.), *Plant Population Genetics, Breeding, and Genetic Resources*. Sinauer, Sunderland, Mass., pp. 43–63.

Harvey, W. H. & Sonder O. W. (1860). *Flora Capensis*. Vol. 1. Cambridge University Press, Cambridge.

Hurka, H., Düring, S. (1994). Genetic control of plastidic L-glutamate dehydrogenase isozymes in the genus *Capsella* (Brassicaceae). *Heredity*, *72*(2), 126-131.

Hurka, H., Freundner, S., Brown, A. H. D., & Plantholt, U. (1989). Aspartate aminotransferase isozymes in the genus *Capsella* (Brassicaceae): Subcellular location, gene duplication, and polymorphism. Biochemical Genetics, *27*(1–2), 77–90.

Hurka, H., Friesen, N., German, D., Franzke, A., Neuffer, B. (2012). "Missing link" species *Capsella orientalis* and *Capsella thracica* elucidate evolution of model plant genus *Capsella* (Brassicaceae). *Molecular Ecology*, *21*(5), 1223–1238.

Hurka, H., & Haase, R. (1982). Seed ecology of Capsella bursa-pastoris (Cruciferae): dispersal mechanisms and the soil seed bank. *Flora*, *172*(1), 35–46.

Hurka, H., & Neuffer, B. (1997). Evolutionary processes in the genus *Capsella* (Brassicaceae). *Plant Systematics and Evolution*, *206*(1-4), 295–316.

Hurka, H., & Neuffer, B. (1991). Colonizing success in plants: genetic variation and phenotypic plasticity in life history traits in *Capsella bursa-pastoris*. In Esser, G. & Overdieck, D. (Eds.), *Modern Ecology: Basic and Applied Aspects* (pp. 77–96). Amsterdam: Elsevier.

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801-1806.

Keller, S. R., Fields, P. D., Berardi, A. E., Taylor, D. R. (2014). Recent admixture generates heterozygosity-fitness correlations during range expansion of an invading species. *Journal of Evolutionary Biology*, 27(3), 616-627.

Kivilaan, A., & Bandurski, R. S. (1973). The Ninety-Year Period for Dr. Beal's Seed Viability Experiment. *American Journal of Botany*, *60*(2), 140–145.

Kloot, P. M. (1983). Early records of alien plants naturalized in South Australia. *Journal of the Adelaide Botanic Gardens*, 6, 93-131.

Kryvokhyzha, D., Holm, K., Chen, J., Cornille, A., Glémin, S., Wright, S., … Lascoux, M. (2016). The influence of population structure on gene expression and flowering time variation in the ubiquitous weed *Capsella bursa-pastoris* (Brassicaceae). *Molecular Ecology*, *25*(5), 1106-1121.

Lamping, H. (1985). *Australien*. Stuttgart: Klett.

Linde, M., Diel, S., Neuffer, B. (2001). Flowering ecotypes of *Capsella bursa-pastoris* (L.) Medik. (Brassicaceae) analysed by a cosegregation of phenotypic characters (QTL) and molecular markers. *Annals of Botany*, *87*(1), 91-99.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1), 209–220.

Marais, W. (1970). Cruciferae. In Codd, L. E., De Winter, B., Killick, D. J. B., Rycroft, H. B., eds., *Flora of Southern Africa*. Vol. 13. Government Printer, Pretoria.

Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058-1069.

Metzger, M. J., Bunce, R. G., Jongman, R. H., Sayre, R., Trabucco, A., & Zomer, R. (2013). A high resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography*, 22(5), 630-638.

Muscarella, R., Galante, P. J., Soley Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENM eval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution*, *5*(11), 1198-1205.

Neuffer, B., & Bartelheim, S. (1989). Gen-ecology of *Capsella bursa-pastoris* from an altitudinal transsect in the Alps. *Oecologia*, *81*(4), 521–527.

Neuffer, B., & Hurka, H. (1999). Colonization history and introduction dynamics of *Capsella bursa-pastoris* (Brassicaceae) in North America: Isozymes and quantitative traits. *Molecular Ecology, 8*(10), 1667–1681.

Neuffer, B., Linde, M. (1999). *Capsella bursa-pastoris* - Colonization and adaptation: A globetrotter conquers the world. In: van Raamsdonk, L. W. D., den Nijs, J. C. M. (eds.) *Plant Evolution in Man-Made Habitats*. (pp. 49-72). Amsterdam: Hugo de Vries Laboratory.

Neuffer, B. (1996). RAPD analyses in colonial and ancestral populations of *Capsella bursa-pastoris* (L.) Med. (Brassicaceae). *Biochemical Systematics and Ecology*, *24*(5), 393–403.

Neuffer, B. (2011). Native range variation in C*apsella bursa-pastoris* (Brassicaceae) along a 2500km latitudinal transect. *Flora*, *206*(2), 107–119.

Neuffer, B., Bernhardt, K. G., Hurka, H., & Kropf, M. (2011). Monitoring population and gene pool dynamics of the annual species *Capsella bursa-pastoris* (Brassicaceae): A review of relevant species traits and the initiation of a long-term genetic monitoring programme. *Biodiversity and Conservation*, *20*(2), 309–323.

Neuffer, B., Hirschle, S., & Jäger, S. (1999). The Colonizing History of *Capsella* in Patagonia (South America) – Molecular and Adaptive Significance. *Folia Geobotanica*, *34*(4), 435-450.

Neuffer, B., & Hoffrogge, R. (1999). Ecotypic and allozyme variation of *Capsella bursa-pastoris* and C. rubella (Brassicaceae) along latitude and altitude gradients on the Iberian Peninsula. *Anales Del Jardin Botanico de Madrid*, *57*(2), 299–315.

Neuffer, B., Wesse C., Voss, I., & Scheibe R. (2018). The role of ecotypic variation in driving worldwide colonization by a cosmopolitan plant. *Annals of Botany Plants* 10: ply005.

Paetsch, M., Mayland-Quellhorst, S., Hurka, H., & Neuffer, B. (2010). Evolution of the Mating System in the Genus *Capsella* (Brassicaceae). In Glaubrecht, M. (Ed.), E*volution in Action* (pp. 77-100). Berlin Heidelberg: Springer.

Peakall, R. & Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology* Notes, *6*(1), 288-295.

Peakall, R. & Smouse, P. E. (2012). GenAlEx6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, *28*(19), 2537-2539.

Pérez de la Vega, M., Garcia, P., & Allard, W. (1991). Multilocus genetic structure of ancestral spanish and colonial Californian populations of *Avena barbata*. *Proceedings of the National Academy of Sciences of the United States of America*, 88, 1202-1206.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling, 190*(3-4), 231-259.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629-643.

Randall, R. P. (2012). A global compendium of weeds. *Department of Agriculture and Food Western Australia*, No.Ed.2 pp.1124 pp.

Robbins, W.W. (1940). Alien plants growing without cultivation in California. Agr*icultural Experimental Station Bulletin no. 6*, California.

Shull, G. H. (1929). Species hybridizations among old and new species of Shepherd's Purse. *Proceedings of the International congress of Plant Sciences*, *1*, 837–888.

Sicard, A., Kappel, C., Josephs, E. B., Lee, Y. W., Marona, C., Stinchcombe, J. R., … Lenhard, M. (2015). Divergent sorting of a balanced ancestral polymorphism underlies the establishment of gene-flow barriers in *Capsella*. *Nature Communications*, *6*, 7960.

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. *American Journal of Botany*, *93*(11), 1714–1724.

Tutin, T. G., Burges, N. A., Chater, A.O., Edmondson, J. R., Heywood, V. H., Moore, D. M., … Webb, D.A. (Eds.) (1993). *Flora Europaea vol. 1*, 2nd edit. Cambridge: Cambridge University Press.

Tutin, T. G., Heywood, V. H., Burges, N. A., Valentine, D. H., Walters, S. M., Webb, D. A. (Eds.) (1964). *Flora Europaea, vol. 1.* First edition. Cambridge: Cambridge University Press.

Zhou, T., Lu, L., Yang, G., & Al-Shehbaz, I. A. (2001). Brassicaceae. *Flora of China*, 8, 1–193.

III. Taking the long way around – Worldwide geographical structure of the cosmopolitan weed *Capsella bursa-pastoris* (Brassicaceae)

Christina Wesse, Daniel Koenig, Barbara Neuffer, Detlef Weigel

In preparation for submission

Abstract

The very common annual weed *C. bursa-pastoris* originated in the steppes of Eurasia and is now widespread in both cold and mesic and hot and arid areas almost all over the world. To display and analyze global geographical distribution patterns of genotypes of this cosmopolitan plant, we used a combination of phenotyping and RADseq data from 1,273 individuals from 384 different collection sites from every continent except Antarctica. Population structure analysis with ADMIXTURE revealed two clusters, one predominantly occuring in warm climate regions and the other in more temperate regions. The obtained clusters not only correlated significantly with the climate of the source areas, but also with the onset of flowering and the genome size of the individuals, indicating a high degree of adaptation of the plants. We argue that the two clusters point to an early diversification into the two lineages and may even suggest multiple origins of the species.

Introduction

Many species invasions are the direct or indirect consequence of human activities, and the economic impact of invasive species cause costs ranging from millions to billions of dollars per year (Sakai et al., 2001). Exotic plants for example have been imported intentionally for medical purposes or ornamentation, but also accidental as by-catch in crop seeds or adhesion to domesticated animals (Sakai et al., 2001). Invasive species also play a remarkable role with regard to global change, because changing environments offer new possibilities for species to spread (Vitousek et al., 1996). A successful establishment of a species into a new habitat involves phenotypic plasticity and the potential for genetic changes through drift or selection (Sakai et al., 2001). While the term “Invasion” specifies a more aggressive form, “Colonization” in the broader sense describes the process by which species successfully immigrate to new areas *per se*. Range expansion as such is a feature of the evolutionary history of all species, whether intercontinental or on a more local scale. John Josselyn reported two dozens newly introduced European weeds in Massachusetts Bay only about 50 years after immigration of Europeans (Mack & Lonsdale, 2001), including dandelion (*Taraxacum officinale*) and the broadleaf plantain (*Plantago major*) – the latter is also known as “the white man‘s footprint“, because it sprawled wherever Europeans put a foot on (Mack & Lonsdale, 2001; Cronon, 1983). Until today, many other species like absinthe (*Artemisia absinthium*), wild teasel (*Dispsacus fullonum*) and *Atropa belladonna* followed and are now naturalized in the United States (Mack & Lonsdale, 2001).

The study of population biology and genetic diversity provides insights to the potential for colonization and can detect geographic patterns of invasion and range expansion. The ‘Genetics of Colonizing Species’, edited by Baker and Stebbins in 1965, can be regarded as the foundational document for “invasion genetics” (Barrett, 2015) and addressed the issues of particular interest in genetic surveys: (i) identification of source populations; (ii) single or multiple introductions; (iii) population structure between native and introduced populations; (iv) genetic diversity in the non-native range compared to the native range; (v) establishment of bottlenecks and founder events; (vi) pre-adaptation vs. post-colonization adaptation to invasive spread; (vii) genetic interactions during admixture of multiple source populations; (viii) new mutations in the introduced range.

An outstanding example for a successful colonizer is the Shepherd’s Purse (*Capsella bursa-pastoris* (L.) Medik.), a member of the mustard family (Brassicaceae). This common weed is tetraploid and predominantly inbreeding. *C. bursa-pastoris* has been widely distributed throughout whole Eurasia and around the Mediterranean Sea in prehistoric times by early agricultural activities of humans. When Europeans colonized other continents from the beginning of the 16th century, the Shepherd‘s Purse was introduced to new habitats alongside other weeds as neophytes (Mooney et al., 2005). First acknowledgments of occurrence of the Shepherd‘s Purse in North America date back to the 17th century when Josselyn visited the east coast (Crosby, 1986). Dispersal of the species presumably increased by cause of anthropogenic colonization of North West America during the gold rush in the 1840s (Hornbeck, 1983; Neuffer, 1996; Neuffer & Linde, 1999). These unintentional transports allowed *C. bursa-pastoris* to also reach South America, Australia, South Africa and nearly every other possible locality, avoiding merely the very hot and humid tropics and arctic climates (Neuffer & Hurka, 1999; Neuffer et al., 1999; Neuffer et al., 2011; Kryvokhyzha et al., 2016). This enormous expansion could be established with extraordinary ecotypic differentiation (e.g. Neuffer & Bartelheim, 1989; Neuffer, 2011), the predominantly selfing mating system, the production of thousands of seeds spawned per individual (Hurka & Neuffer, 1991), the ability to survive in a soil seed bank for many years (Hurka & Haase, 1982), and the power for long distance dispersal via myxospermy (Neuffer & Linde, 1999). All these factors made the Shepherd’s Purse one of the most wide spread flowering plant species on earth today (Coquillat, 1951; Zhou et al., 2001; Randall, 2012). The colonization history of this plant has been traced in parts by molecular markers in previous studies (e.g. RAPDs in Neuffer, 1996; isozymes in Neuffer & Hurka, 1999; isozymes and RAPDs in Neuffer et al., 1999; isozymes in Neuffer et al., 2011; GBS in Kryvokhyzha et al., 2016).

The consequences of a predominantly self-pollinating plant are local differentiation due to founder effects and restricted gene flow between distant populations. Hence, it is an interesting evolutionary question whether populations from newly colonized continents differ from the source continent. Genome-wide marker analyses such as restriction site-associated DNA sequencing (RADseq) are useful to perform population genetic studies akin to analyses like restriction fragment length polymorphisms (RFLPs) and amplified fragment length polymorphisms (AFLPs) by reducing the complexity of the genome by the use of restriction enzymes (Davey & Blaxter, 2010). RADseq surpasses these methods by identifying several thousands of genetic markers from a group of individuals simultaneously (Davey & Blaxter, 2010). Therefore, RADseq makes possible population genetics studies of unprecedented depth and complexity and allows the exploration of evolutionary history, range expansion and invasion patterns of colonizing species.

To display the genetic diversity of *C. bursa-pastoris* has been subject of many studies before (e.g. Neuffer & Hurka, 1999, Ceplitis et al., 2005, Slotte et al., 2008; Cornille et al., 2016; Wesse et al., 2019). However, we report here a more extensive sampling from sites from every continent except Antarctica. We show here that a large number of loci and a wide global sampling area reveal finer-scale population structure of *C. bursa-pastoris* than has previously been detected. The aims of our study are to (i) describe the spatial distribution of *C. bursa-pastoris* using a large number of genomewide SNPs, (ii) to what extent the observed population structure is due to colonization history and environmental patterns, (iii) show local adaptations both in the native and non-native range of the species, and (iv) reconstruct migration patterns. To answer these questions, we use a combination of phenotyping and SNP sequencing data from 1,273 individuals from 384 different collection sites.

Materials and Methods

Plant material

The seeds come from parental plants from populations from a variety of locations from all over the world. The seeds were randomly taken from natural provenances and collected over a period of three decades from 1982 to 2016. A list of all samples with geographical coordinates and other sampling site information can be found in the supplement. As the germination rate is significantly reduced only five years after collection (Neuffer & Hurka, 1988) the seeds have been stored in special plastic bags in -20 °C until until usage for the experiments. All seed vouchers are stored at the Botanical Garden of the Osnabrück University. Herbarium material of many accessions is deposited in the Herbarium of the Osnabrück University OSBU.

The seed weight was measured before sowing in mg per 50 seeds. Sowing happened in sowing substrate (1:2 TKS®1 + gravel sand, sieved) in the greenhouse. The seeds were then preventatively treated with fungicide Previcur® (0.25 %) and covered with a transparent plastic hood. The germination percentage was recorded. In some cases, germination was induced with gibberellic acid whenever families did not germinate at first at all. When possible, shortly before reaching the 4-leaves-stage, six seedlings were transplanted for each family, of which five were used for the common garden experiment for phenotyping and one was kept in the green house for DNA extraction and flow cytometry. Whenever germination rate was too low to obtain a sixth individual, tissue for aforementioned analyses was taken directly from field specimens. In some rare cases, the amount of siblings (i.e. number of replicates) was less than five as a consequence of low germination rate. Individuals were planted on the experimental field of the Botanical Garden of the Osnabrück University (N52° 16'56.21", E8° 1'46.30") in randomized order. At this point, most specimens had already developed a firm rosette. Planting was performed in a randomized controlled arrangement. Flowering time was recorded as the exact day after sowing when the first white petals showed at the bud (buds were observed daily) and was averaged between replicates. We also recorded the number of basal inflorescences and the height of the heighest inflorenscence. Leaf shapes were determined on fully differentiated adult leaves according to the classification of Shull (1909), primarily discriminating the dissection of leaves, which ranges from entire leaves to very deeply dissected ones: *simplex*, *tenuis*, *rhomboidea*, and *heteris*.

Estimation of genome size

For estimation of nuclear DNA content, flow cytometry (FCM) was carried out relatively to the garden parsley *Petroselinum crispum* as an internal reference standard. Genome sizes were estimated with the CyStain® UV Precise P reagents (Sysmex Partec GmbH, Germany) following the protocols described in Doležel et al., 2007 with few modifications: 1 cm² of young fresh leaf tissue was co-chopped with 1 cm² of tissue from the standard in the presence of 0.4 ml cold CyStain® Nuclei Extraction Buffer manually in a glass petri dish with a razor blade. The cell suspension was mixed briefly with a vortexer and then stained with 1.6 ml CyStain® Staining Buffer containing 4’,6-diamidino-2-phenylindole (DAPI) as dye. The sample was then filtered through a 50 µm CellTrics® filter. FCM analysis was performed with the CyFlow® Ploidy Analyser (Sysmex Partec GmbH, Germany) with following settings: GAIN: 540 V, velocity: 0.4 µl/s, 365 nm UV-LED, 532 nm excitation, 532 nm emission. Samples were run until ca. 50 ml of flow-through. The nuclear DNA content (C-values) were calculated from gated fluorescence histograms: (G1 peak of *C. bursa-pastoris* / G1 peak of *P. crispum* standard) x 2C DNA content of *P. crispum* (4.46 pg; Yokoya et al., 2000). None of the peaks of any of the samples overlapped with the standard. In most cases, FCM measurements were replicated three times per individual and averaged.

RAD sequencing and SNP call

Extraction of genomic DNA from *C. bursa-pastoris* rosette leaves was performed with the CTAB method and DNA quality and concentration were determined on a 1 % agarose gel and with the Tecan Fluorescence Microplate Reader (Tecan Group AG, Swiss). Libraries were prepared after normalization of the DNA extracts and following the *KpnI* RAD protocol: DNA was digested with the restriction endonuclease FastDigest *KpnI* and 10X FastDigest Buffer (Thermo Fisher Scientific, United States) in a thermal cycler at 37 °C for 30 min. Nucleotide multiplex identifier were ligated to the samples using T4 Ligase (5 U/µl), PEG 4000 and 10X ligase buffer (Thermo Fisher Scientific, United States) at room temperature for 30 min. DNA was fragmented by sonication with a Focused-ultrasonicator (Covaris, United States) to target 500 bp fragments. End-repair of sheared DNA fragments, A-tailing and adapter ligation were performed with NEBNext DNA Sample Prep MMS1 (New England BioLabs, United States) using Universal adapters G-34024 and G-34025. The single-end DNA libraries were amplified by PCR for 14 cycles with PCR primers G-26878 and G-33106. After each step during library preparation the samples were cleaned-up with AMPure® XP SPRI® beads (Beckman Coulter, United States) and 80 % ethanol. Libraries were sequenced on Illumina HiSeq Analyzer in a total throughput of six lanes.

After demultiplexing, all reads were trimmed with *Trimmomatic* version *0.36* (Bolger et al., 2014) with a sliding window to eliminate bad quality reads, removal of the illumina adaptor sequences and verified for a minimum length of 75 bases. The trimmed reads were then mapped either to a *C. bursa-pastoris* reference genome or a pseudoreference generated *in silico* from concatenated genomes of *C. orientalis* and *C. rubella* using *BWA* version *0.7.12* (Li and Durbin, 2009) with default parameters but ignoring indels. Sorting of the resulting alignment files was performed with *SAMtools* version *1.4.1* (Li et al., 2009). The program *freebayes* (Garrison and Marth, 2012) was used for initial SNP calling using the freebayes-parallel script. A total number of 709,542 raw SNPs were called. The resulting vcf was filtered with *VCFtools* version *0.1.13* (Danecek et al., 2011) for minimum quality of 30 and maximal fraction of missing data 70 %. Samples with reads < 50,000 were removed from the vcf, as also samples which seemed to be diploid or triploid according to preferred mapping on either one of the diploid genomes within the aforementioned artificially created pseudoreference. Samples with unusually high heterozygosity were also removed. Finally, the dataset was filtered for MAF = 0.05. This resulted in a final vcf-file containing 1,273 different sequenced *C. bursa-pastoris* individuals with 13,006 high quality SNPs.

Population structure analyses

The bioclimatic variables were derived from WORLDCLIM (http://www.worldclim.org/) via the R-package „raster" with a spatial resolution of 2.5 minutes on March 23th 2018. Population structure was analyzed using ADMIXTURE *1.3.0* (Alexander et al., 2009). Principal component analysis was performed with the R-package "SNPrelate".

Results

Population structure

Figure III.1 shows the worldwide genetic population structure of *C. bursa-pastoris* as derived from ADMIXTURE analysis. We found two major clusters, with one cluster mainly located in most parts of North America outside California, Middle and Eastern Europe and across the Asian continent (blue), whereas the other cluster is to be found in California, parts of South America, Mediterranean Europe, Africa, India and Australia (red) (Fig. III.1a). We tested ADMIXTURE from $K = 2$ to $K = 10$, but since cross-validation (CV) errors decreased more slowly after $K = 2$, we chose this as most relevant number of clusters to describe the population structure (Fig. 1b). For $K = 2$, the analysis revealed two separated clusters with little admixture (Fig. III.1c). The red cluster occurs predominantly in warm habitats whereas the blue cluster prefers warmer climate (Fig. III.1d).

Since the ADMIXTURE algorithm tends to underestimate the true value of *K*, it is recommended to compare the results of multiple *K* outcome (Lawson et al., 2018). In Figure III.2, the results of the ADMIXTURE analysis are shown for $K = 2$, $K = 3$ and $K = 4$ in comparison. The figure shows the admixture subdivided in single continents besides the sampling locations and the PCA plots with the equivalent coloring (Fig. III.2).

With $K = 2$ (Fig. III.2a – c), the North American continent comprises both clusters with very little admixture, with one cluster predominantly occurring in lowland California (red) and the other one mainly in highland California and the rest of the USA (blue). In South America one cluster is prevalent (red) with very few exceptions (blue). The blue cluster is almost entirely located in south Chile and south Argentina with one exception in north Venezuela. Both clusters do also occur in Europe, with one predominantly in Middle and Eastern Europe and Scandinavia (blue) and the other one in the mediterranean areas like the Iberian Peninsula (red). The British Isles comprises of individuals from both clusters. The African populations show great homogeneity with only one of the clusters occurring (red) with very little admixted samples. The Asian continent houses mainly one cluster in central Asia, Russia and Japan (blue), and very few of the other one occurring in south India, east Russia and the country square between Russia, Kazakhstan, Mongolia and China. Australia, comparable with Africa, has only one cluster (red) with insignificant amount of admixture.

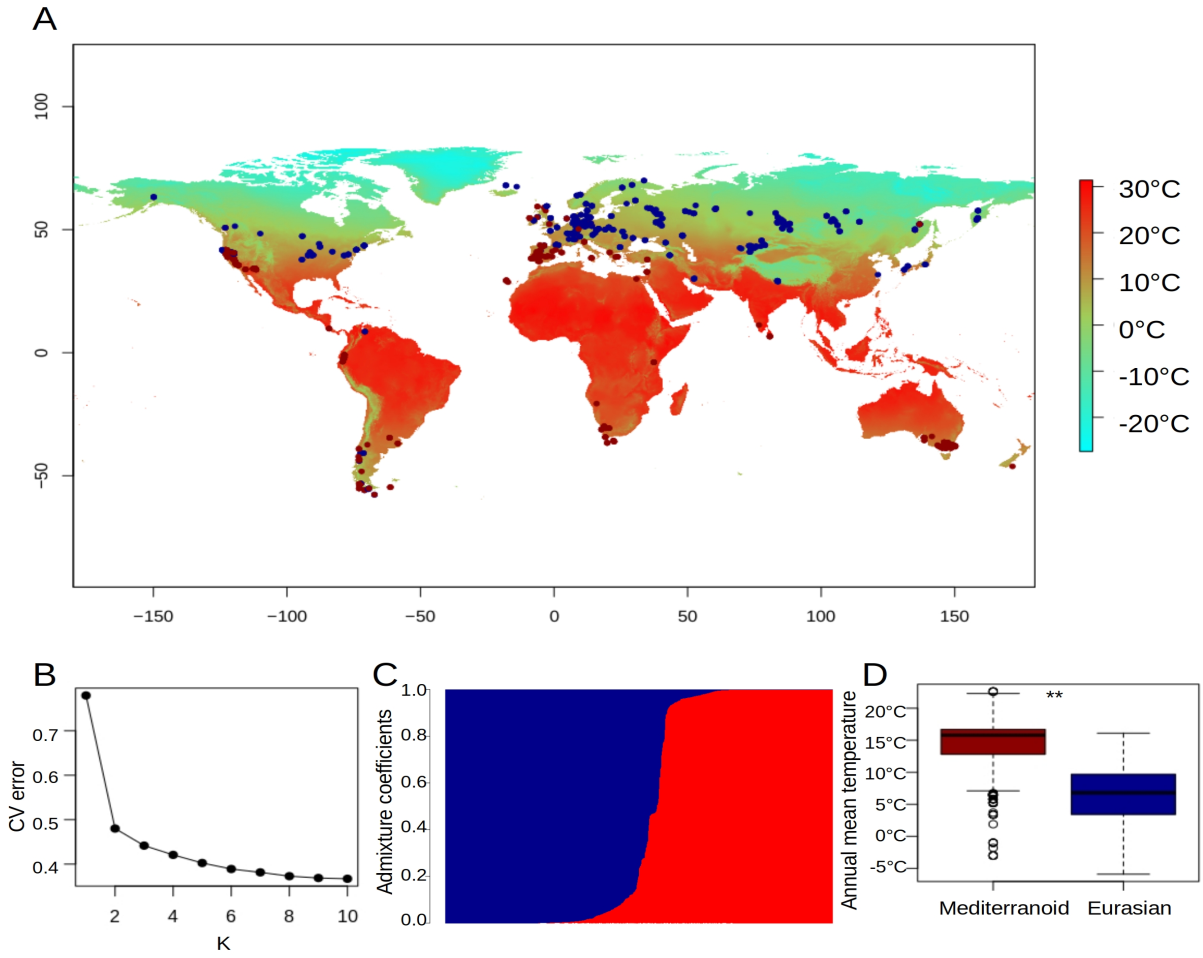

Figure III.1: Worldwide population structure of *Capsella bursa-pastoris*.
A: Sampling sites with cluster affiliation as derived from ADMIXTURE analysis. Map colors refer to hot (red) and cold (aquamarine) climates as derived from Worldclim data for annual mean temperature (BIO1). B: *K*-estimation from CV errors. C: Population structure analysis showing two distinct clusters with little admixture. D: Means of annual mean temperature of the source habitat. **: $p < 0.003$

With $K = 3$ (Fig. III.2d – f), the majority of the former red cluster from $K = 2$ (Fig. III.2a – c) divides into 2 subclusters, with the one subcluster (orange) mostly occurring in lowland California, Middle- and South America and parts of south Africa, and the other subcluster (yellow) mostly in southern Europe, parts of the British Isles and Australia. The blue cluster resembles the blue cluster from $K = 2$ in most parts.

With increasing number to $K = 4$ (Fig. III.2g – i), the majority of the former blue cluster (Fig. III.2a – c) subdivides in two, with one part predominantly occurring in North America outside lowland California, Middle and Eastern Europe, Scandinavia and north and central Asia (light blue) and the other one mainly in Turkey, Kazakhstan, Kyrgyzstan, Nepal, Eastern Russia and Japan (dark blue). The apportionment from the former red cluster from K = 2 (Fig. III.2a – c) follows the pattern from K = 3 (Fig. III.2d – f) in most parts.

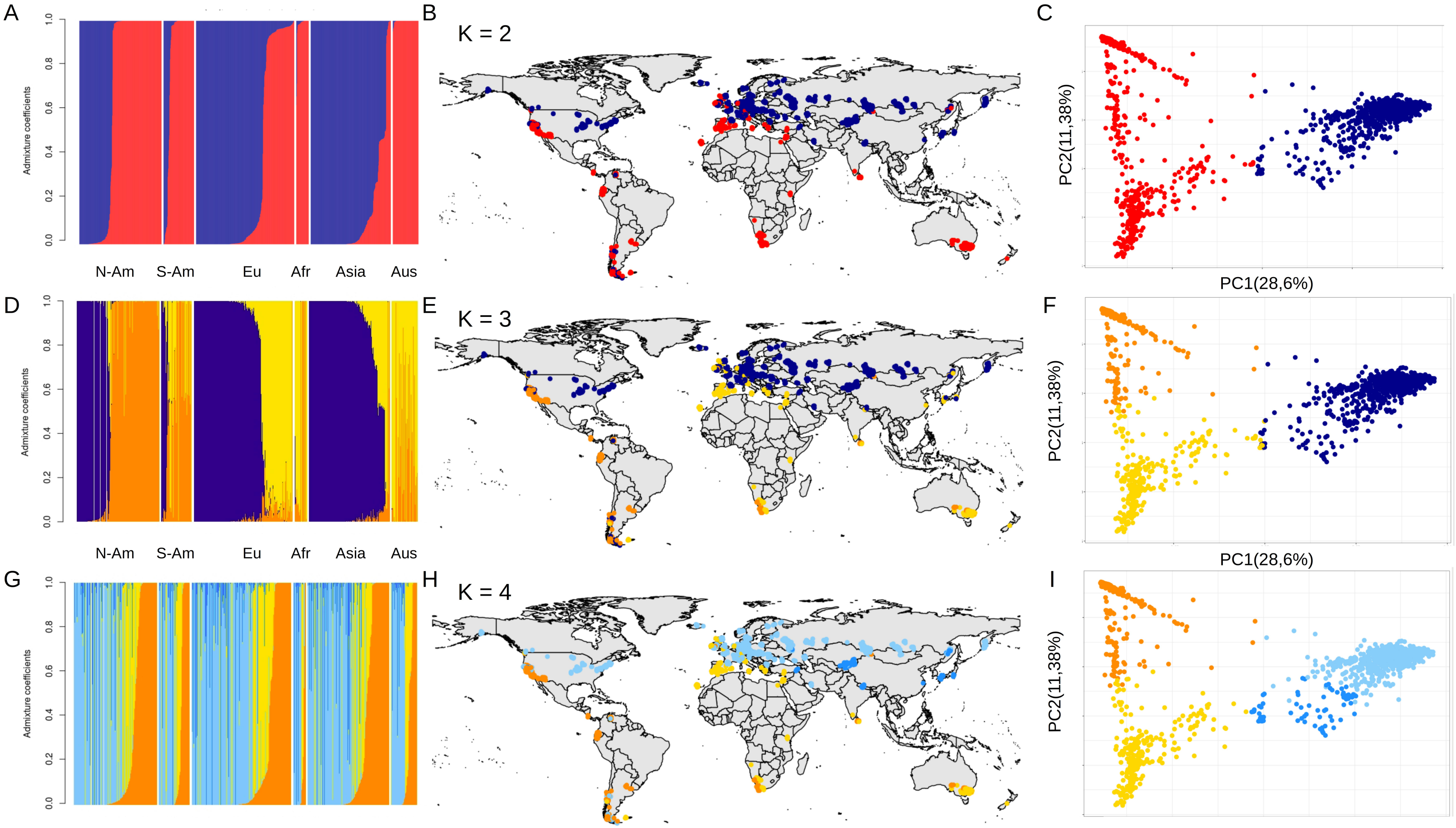

Figure III.2: Comparison of population structure analysis of *C. bursa-pastoris*. Colors refer to ADMIXTURE cluster affiliations. A-C: *K* = 2. D-F: *K* = 3. G-I: *K* = 4. A, D, G: Population structure analysis as derived from ADMIXTURE. B, E, H: Sampling sites. C, F, I: Principal component analysis.

Assuming that K = 2 is the optimal number of clusters (Fig. III.1b), the following analyses will focus on two obtained clusters. According to the native range of the species set in Europe and Asia, the two *C. bursa-pastoris* clusters will be called “Mediterranoid“ (warm climate, red) and “Eurasian“ (cold climate, blue) from this point onwards.

Cluster Adaptation

Plotted on a map, the two obtained clusters from the K = 2 ADMIXTURE analysis seem to be adapted to different climate zones: cold (blue) and warm (red) climate (Fig. III.1a, III.1d; other adaptions to a variety of climate factors are shown in appendix *5.2.13 Cluster climate adaption*, Fig. S.14). Furthermore, the two clusters show highly significant differences in means of some of the recorded phenotypes (Fig. III.3, Tab. III.1, Fig. III.5): Both clusters differ obviously in flowering time (Fig. III.3c) and genome size (Fig. III.3f). The differences in germination (Fig. III.3b) and the number of branches (Fig. III.3e) are not obvious by eye but still significant (Tab. III.1). The clusters are statistically equal relating to seed weight (Fig. III.3a) and plant height (Fig. III.3d).

The genome of *C. bursa-pastoris* is relatively small (mean 2C = 0.91 pg), and our measured values approximate the 2C-value of 0.8 pg reported by Lysak et al. (2009) for the same species. However, the recorded values for the genome size varied enormously, and according to the data presented here, there is a strong correlation between onset of flowering and genome size (pearson *cor* = 0.551, p = 2.2e-16): Specimens from warmer climates (i.e. Mediterranoid cluster, red) showed early flowering and small genome sizes, whereas individuals from colder regions (i.e. Eurasian cluster, blue) showed late flowering and bigger genome sizes (Fig. III.4).

Table III.1: Comparison of means of phenotypes of the two clusters.
SW: seed weight. GER: germination percentage. FLW: flowering day after sowing. HGT: plant height. BR: number of branches. GS: genome size. **: p < 26.79e-6. *: 26.79e-6 < p < 0.05. n.s.: p > 0.05.

	SW	GER	FLW	HGT	BR	GS
All samples	5.015 mg	45.07 %	58.8 days	59.91 cm	7.338	0.9137 pg
Mediterranoid cluster	5.039 mg	48.06 %	53.16 days	60.21 cm	7.677	0.8851 pg
Eurasian cluster	4.999 mg	42.99 %	62.88 days	59.22 cm	6.572	0.936 pg
Wilcoxon rank sum test *p*	n.s.	*	**	n.s.	*	**

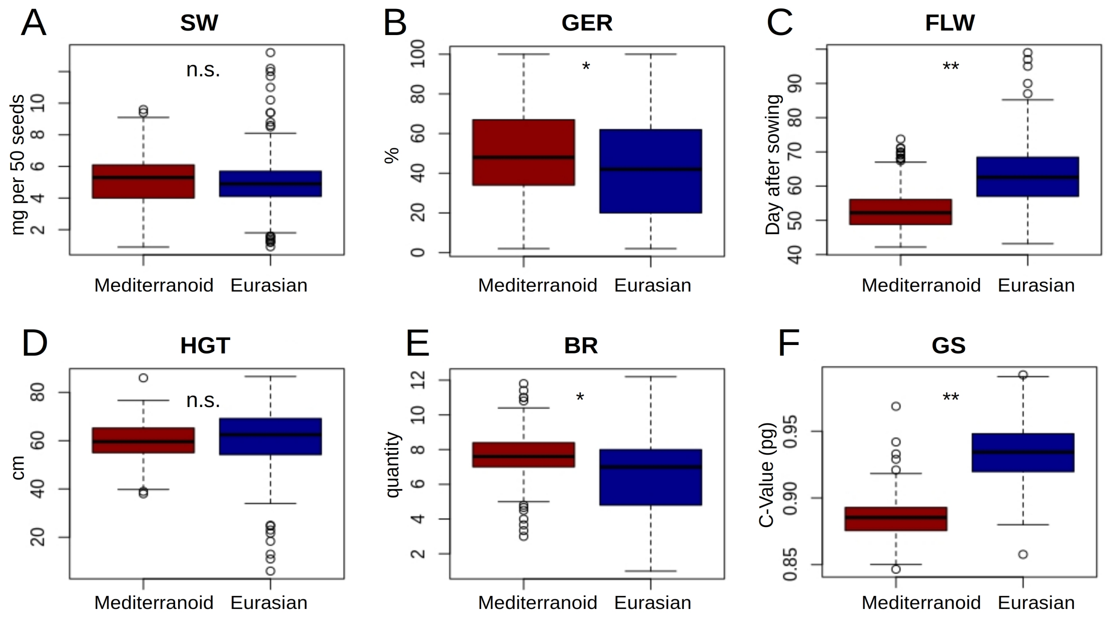

Figure III.3: Comparison of means of numerical phenotypic parameters. Red: Mediterranoid cluster. Blue: Eurasian cluster. SW: seed weight. GER: germination percentage. FLW: flowering day after sowing. HGT: plant height. BR: numer of branches. GS: genome size.

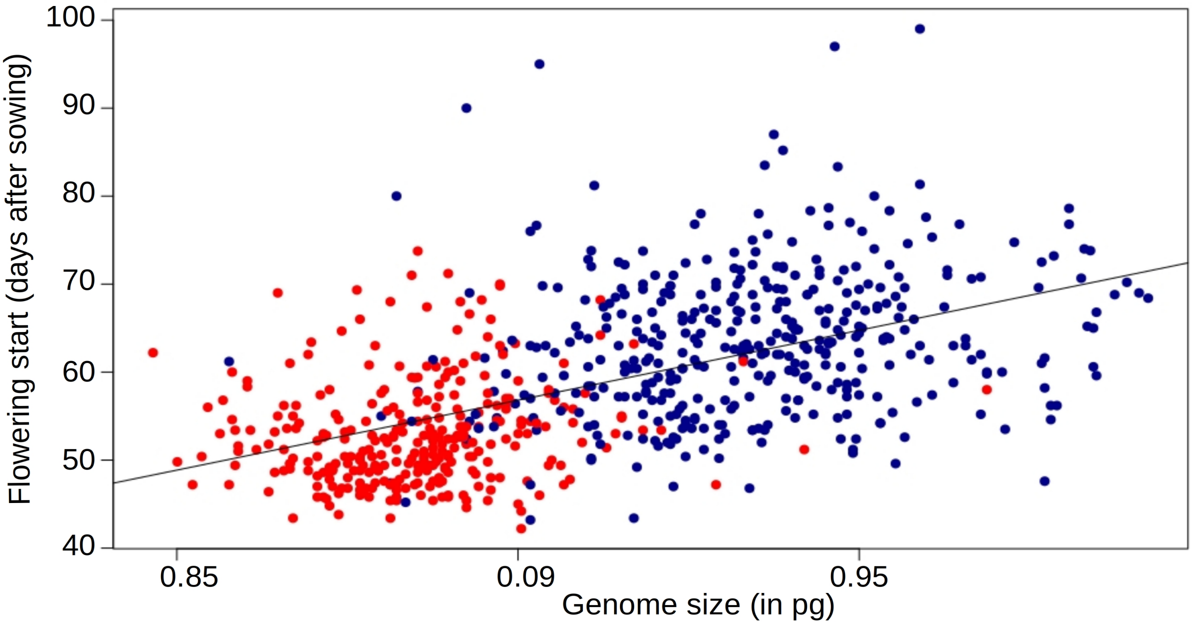

Figure III.4: Correlation plot of genome size and onset of flowering. Colors refer to cluster affiliation (red: Mediterranoid, blue: Eurasian).

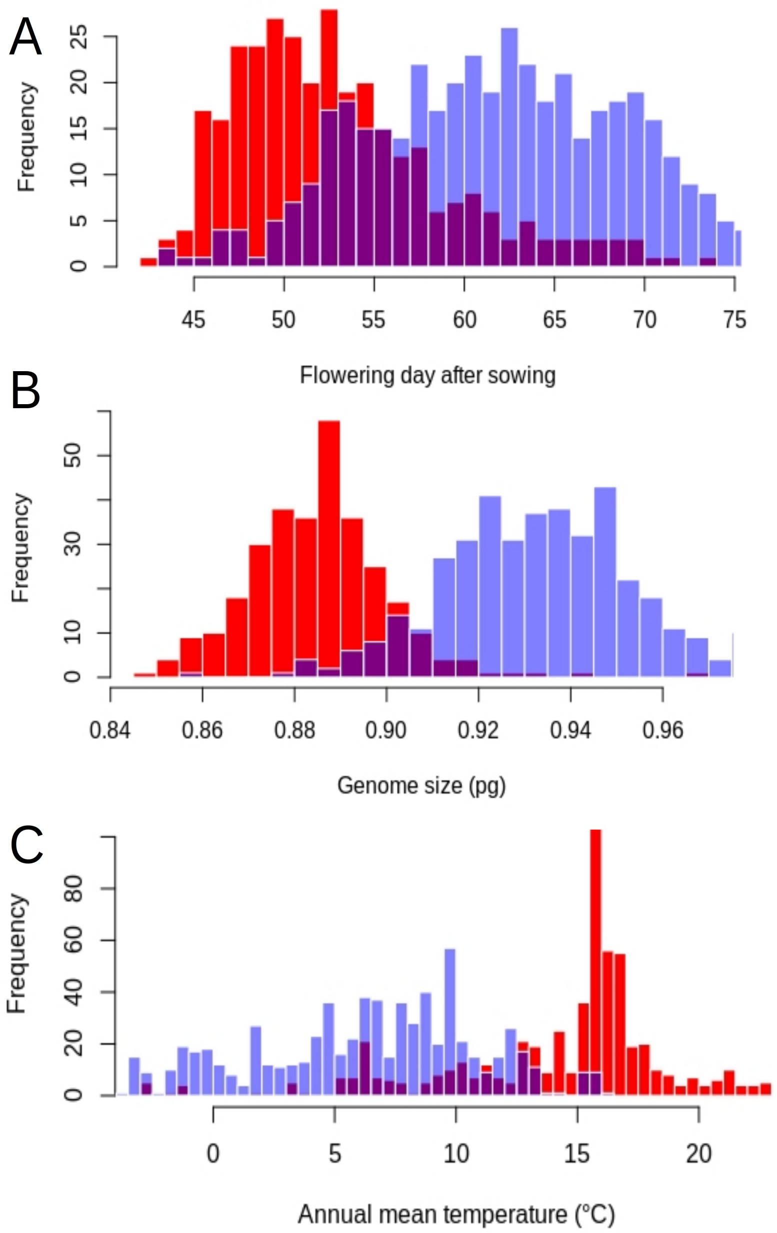

Figure III.5: Histograms of the two clusters. Blue: Eurasian cluster. Red: Mediterranoid cluster. A: Flowering day after sowing. B: Genome size. C: Annual mean temperature of source habitat.

Discussion

Genome size

The “large genome constraint hypothesis” asserts that plant species with generally small genomes are geographically more widely distributed and that invasive plant species privilege small genomes in particular (Knight et al., 2005; Rejmánek, 1996; Bennett et al., 1998; Suda et al., 2014) – a hypothesis which suits well for the great colonizing ability and the world-wide distribution of *C. bursa-pastoris*. This colonizer inhabits a variety of humid to semiarid habitats in alpine regions as well as coniferous woodland and pampas.

Within the angiosperms it is common to increase genome size due to polyploidization events, but reduce it in the course of evolution to curtail number of gene copies (Suda et al., 2014; Soltis et al., 2003; Kashkush et al., 2002; Lysak et al., 2009). Either way, genome size varies greatly between species of the Brassicaceae family at least 16.2-fold across the family between *Sphaerocardamum sp.* (1C = 0.15 pg; Bailey, 2001) and *Bunias orientalis* (1C = 2.43 pg; Lysak et al., 2009). Previous studies indicated not only great variability between but also within species (Long et al., 2013; Ŝmarda & Bureš, 2010). Variability in the amount of DNA is thought to play an important role in plant phenotypic evolution of species (Knight et al., 2005; Meagher and Vassiliadis, 2005) and has been shown to correlate with environmental parameters (e.g. Long et al., 2013; Díez et al., 2013; Albach & Greilhuber, 2004; Kang et al., 2014) as well as phenotypes like flowering time and seed weight (Meagher & Vassiliadis, 2005; Lavergne et al., 2010). Our data shows great intraspecific variability in genome size in *C. bursa-pastoris*. Samples varied from minimum to maximum 1.16-fold (≙ 16.5 %). FCM is a well-established and solid method for estimation of nuclear DNA content (e.g. Doležel et al., 2007; Suda et al., 2014), so we assume that our data is valid. The observed variation is unlikely to be explained by differences in chromosome numbers, because *C. bursa-pastoris* is described as a cytologically uniform species with *n* = 16 chromosomes (Raj, 1965). Chromosome countings of individuals from different populations did not deliver deviating results (Neuffer, personal comment). Genome size variations are generally small within plant species and on the contrary enormous between different species (Gregory, 2011; Greilhuber & Leitch, 2013). Nevertheless, intraspecific genome size variation in plants with the same ploidy is known and is primarily explained by variation in amounts of transposonal elements (TE) and repetitive sequences (Šmarda & Bureš, 2010; Muñoz-Diez et al., 2012), for instance the genome size of *Arabidopsis thaliana* showed more than 10 % variation (Long et al., 2013) or at least 30 % in *Zea mays sensu lato* (Muňoz Díez et al., 2012). It is assumed that intraspecific genome size variation is primarily found in young radiating species (Šmarda & Bureš, 2010), however, the significance for the high variability in genome size for the invasiveness of *C. bursa-pastoris* has yet to be investigated further.

Population structure

According to the native range of the species set in Europe and Asia, we hereby refer to the two *C. bursa-pastoris* clusters obtained from ADMIXTURE as "Mediterranoid" (warm climate, red *n* = 578 individuals) and "Eurasian" (cold climate, blue, *n* = 695 individuals) groups. Little admixture was detected between these groups, reflecting the selfing mating system of this species. However, some hybrids indicating gene flow between populations do occur in North America, South America, Europe and Asia (Fig. III.2a). Although predominantly selfing, crossing rates within *C. bursa-pastoris* around 12 % are reported (Hurka & Neuffer, 1997). Previous studies indicated that individuals with typical mediterranoid and temperate genotypes have lower crossing success if the Eurasian type is the mother plant (Linde, 1999, unpublished), which could explain the little admixture observed.

We found a strong correlation between cluster affiliation and climate, so we assume that these clusters are highly adapted to their environment. An important trait for annual plants is onset of flowering. Too early or too late flowering can cause elimination or reduced seed production and therefore drastic cut in population size. In this study, we also found a strong correlation between cluster affiliation and flowering time, meaning that individuals from warm climates showed early flowering and vice versa, and we found associated certain SNPs with measured flowering time, which confirms genetic ecotypic adaptation of the clusters (see *2. Genome-wide association mapping*). Our data shows also high correlation with flowering time and estimated genome size, showing that individuals from the warm cluster not only flower early but also have smaller genomes on average. A positive correlation between flowering time and genome size has also been shown before in maize (Rayburn et al., 1994; Bilinksi et al., 2018). This leads to the assumption that genome size might be somehow linked to flowering. However, the correlation between genome size and flowering start might be completely random. Genome size evolution is dynamic with both increases and decreases being reported within Brassicaceae (Johnston et al., 2005; Lysak et al., 2009), and statistical analysis suggested that genome size is not strongly influenced by selection and evolves most likely passively (Lysak et al., 2009). It is likely that the genome sizes of the two clusters developed independently because the clusters diverged very early. As *C. bursa-pastoris* is an allotetraploid (Douglas et al., 2015), one of the clusters may have inherited the early flowering trait from one of the two parent species *a priori*, but it is likely that the parent species also had an adaption of flowering time. However, it is also possible that the flowering time adaption happened after the origination of *C. bursa-pastoris*.

Colonization origin and migration patterns

For approximately 500 years, vascular plants species have been migrating between and within continents, mainly anthropogenically. Selfing, relatively small genome size and polyploidy of *C. bursa-pastoris* might be reasons colonization success, but rapid expansion is also heavily influenced by human activities. Cornille et al. (2016) hypothesized a colonization origin of *C. bursa-pastoris* in the Middle East, whereas previous studies put a focus on the Eurasian steppe belt (Hurka et al., 2012). Long distance dispersal of the Shepherd‘s Purse has been enabled by human migration. First records of *Capsella* in South Australasia are from 1847 (Kloot, 1847) and in South America from 1877 in Patagonia (see Neuffer et al., 1999). The first narrations of occurence of this species alongside other European weeds in the New World date back to the first half ot the 17th century (e.g. reviewed in Crosby, 1986). Previous studies explained introduction of European genotypes into North American sites (i.e. California Central valley) by early Spaniards (Neuffer & Hurka, 1999). The data presented here reaffirms genetic similarity of individuals from the Iberian peninsula and lowland California (Fig. III.2b), where pre-adapted individuals found their preferred niche. However, we also found evidence for new adaptations of a Mediterranoid subcluster occuring in the introduced but not the native range (Fig. III.3e).

The origin of *C. bursa-pastoris* is still quite controversially disputed: First, it was assumed that *C. bursa-pastoris* is an allotetraploid due to hybridization between *C. rubella* and *C. grandiflora* (Hurka et al., 1989). Later, this species was hypothesised as an ancient autopolyploid from a *C. grandiflora* ancestor (Hurka & Neuffer, 1997). Another supposition is that *C. bursa-pastoris* originated due to autopolyploidization of a diploid selfincompatible ancestor of a *Capsella* lineage by resulting in the tetraploid selfcompatible *C. bursa-pastoris* (Hurka et al., 2012). One newer and more tightened hypothesis assumes that *C. bursa-pastoris* is an allotetrapolyploid through hybridization of ancestral lineages of diploid *C. grandiflora* and *C. orientalis* 100 – 300 kya (Douglas et al., 2015). Origination of *C. bursa-pastoris* is suggested in Eurasia according to occurrence of these parental species (Hurka et al., 2012), or in the Middle East (Cornille et al., 2016). However, there is discussion if this species has single (e.g. Guo et al., 2009) or multiple origins (e.g. Hurka et al., 2012).

According to our findings presented here, we hypothesize multiple origination of *C. bursa-pastoris* at least twice, giving rise to the temperate cluster in the Eurasian Steppe belt and the warm cluster in the European Mediterranean areas. Both clusters show great differences in climate adaption, onset of flowering time and measured genome size. Another hypothesis for the observed pattern could also be due to single origination and very early divergence afterwards with adaptation of each group to either warm or temperate climate. Apparently, our analyses show that the warm cluster is genetically more diverse than the temperate cluster (Fig. III.2c). One possible explanation for this observation is unequal introgressions into both clusters. Unidirectional gene flow from *C.*

rubella to *C. bursa-pastoris* as well as different proportions of introgression into different subpopulations of *C. bursa-pastoris* has been demonstrated in the past (Han et al., 2015; Slotte et al., 2006, 2008; Kryvokhyzha et al., 2019). However, if one assumes different originations of the clusters, the warm cluster could be the older cluster and therefore had more time to accumulate mutations and diversify. If bottleneck effects alternate with periods of immense population growth, this has an effect on evolutionary processes, therefore it is also possible that the temperate cluster underwent a bottleneck in the past, which led to reduction of its gene pool. A single origin seems unlikely (e.g. Douglas et al., 2015; Wesse et al., 2019; Kryvokhyzha et al., 2019), but it is difficult to determine the exact number of founding lineages. However, maybe more than one hypothesis might be true: Since we found two distinctive groups via ADMIXTURE within our data and the differences in matters of genome size as well as flowering start and climate adaptation lead to the assumption that we have indeed two clusters within the studied *C. bursa-pastoris* populations and therefore maybe two groups with different spatial and temporal origins, but introgressions also play a role.

Acknowledgement

We thank our many colleagues for collecting seeds over the past decades. We also thank the Botanical Garden in Osnabrück for growing the plants and Annica Härig, Madita Knieper, Denise Pietzka, Martin Rebilas and Johanne Tietmeyer for help during the common garden experiments. Thanks to Danelle Seymour for help with RADseq. Financial support by the German Research Foundation DFG (NE 314/11-2) is greatly acknowledged.

Literature Cited

Ågren, J. A., Huang, H. R., & Wright, S. I. (2016). Transposable element evolution in the allo-tetraploid *Capsella bursa-pastoris*. *American Journal of Botany*, *103*(7), 1197– 1202.

Albach, D. C., & Greilhuber, J. (2004). Genome size variation and evolution in *Veronica*. *Annals of Botany*, *94*(6), 897–911.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655-1664.

Baker, H. G., & Stebbins, G. L. (Eds.), (1965): *The Genetics of Colonizing Species*. New York: Academic Press.

Barrett, S. C. H. (2015). Foundation of invasive genetics: The Baker and Stebbins legacy. M*olecular Ecology*, *24*(9), 1927-1941.

Bennett, M. D., Leitch, I. J., & Hanson, L. (1998). DNA Amounts in Two Samples of Angiosperm Weeds. *Annals of Botany*, *82*(1), 121–134.

Bennett, M. D., Leitch, I. J., Price, H. J., & Johnston, J. S. (2003). Comparisons with *Caenorhabditis* (∼100 Mb) and *Drosophila* (∼175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ∼157 Mb and thus ∼25% larger than the *Arabidopsis* genome initiative estimate of ∼125 Mb. *Annals of Botany*, *91*(5), 547–557.

Bilinski, P., Albert, P. S., Berg, J.J., Birchler, J. A., Grote, M. N., Lorant, A., … & Ross-Ibarra, J. (2018). Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays. *PLoS Genetics*, *14*(5), e1007162.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

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

Molecular Ecology, *14*(14), 4221-4233.

Coquillat, M. (1951): Sur les plantes les plus communes de la surface du globe. *Bulletin Mensual Societé Linnéenne*, Lyon, *20*(20), 165–170.

Cornille, A., Salcedo, A., Kryvokhyzha, D., Glémin, S., Holm, K., Wright, S. I., & Lascoux, M. (2016). Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (*Capsella bursa-pastoris*). M*olecular Ecology, 25*(2), 616-629.

Cronon, W. (1983): *Changes in the Land* (p. 132). New York: Hill and Wang.

Crosby, A. W. (1986). *Ecological Imperialism. The Biological Expension of Europe, 900–1900*. Cambridge University Press, Cambridge.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. B*riefings in functional genomics*, *9*(5-6), 416-423.

Díez, C. M., Gaut, B. S., Meca, E., Scheinvar, E., Montes-Hernandez, S., Eguiarte, L. E., & Tenaillon, M. I. (2013). Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytologist*, *199*(1), 264–276.

Doležel, J., Greilhuber, J., & Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, *2*(9), 2233–2244.

Douglas, G. M., Gos, G., Steige, K. A., Salcedo, A., Holm, K., Josephs, E. B., ... & Platts, A. E. (2015). Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris. Proceedings of the National Academy of Sciences*, *112*(9),2806-2811.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint* arXiv:1207.3907.

Gregory, T. R. (Ed.). (2011). T*he evolution of the genome*. Elsevier.

Greilhuber, J., & Leitch, I. J. (2013). Genome size and the phenotype. In *Plant Genome Diversity Volume 2* (pp. 323-344). Springer, Vienna.

Guo, Y.-L., Bechsgaard, J. S., Slotte, T., Neuffer, B., Lascoux, M., Weigel, D., & Schierup, M. H. (2009). Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. *Proceedings of the National Academy of Sciences*, *106*(13), 5246–5251.

Han, T.-S., Wu, Q., Hou, X.-H., Li, Z.-W., Zou, Y.-P., Ge, S., & Guo, Y.-L. (2015). Frequent Introgressions from Diploid Species Contribute to the Adaptation of the Tetraploid Shepherd's Purse (*Capsella bursa-pastoris*). *Molecular Plant*, *8*(3), 427–438.

Hornbeck, D. (1983). *California patterns: a geographical and historical atlas*. Mayfield Pub Co, 1983.

Hurka, H., Freundner, S., Brown, a. H. D., & Plantholt, U. (1989). Aspartate aminotransferase isozymes in the genus *Capsella* (Brassicaceae): Subcellular location, gene duplication, and polymorphism. *Biochemical Genetics*, 27(1–2), 77-90.

Hurka, H., & Haase, R. (1982). Seed Ecology of *Capsella bursa-pastoris* (Cruciferae): Dispersal Mechanism and the Soil Seed Bank1. *Flora*, *172*(1), 35-46.

Hurka, H., & Neuffer, B. (1991): Colonizing success in plants: genetic variation and phenotypic plasticity in life history traits in *Capsella bursa-pastoris*. In: Esser, G., Overdieck, D. (eds.) *Modern Ecology - Basic and Applied Aspects*, Amsterdam, London, New York, Tokyo, Elsevier-Vlg., 77-96.

Hurka, H., & Neuffer, B. (1997). Evolutionary processes in the genus *Capsella* (Brassicaceae)*. *Plant Systematics and Evolution, 206(1-4), 295–316.*

Johnston, J. S., Pepper, A. E., Hall, A. E., Chen, Z. J., Hodnett, G., Drabek, J., … & Price, H. J. (2005). Evolution of genome size in Brassicaceae. *Annals of Botany*, *95*(1), 229–235.

Kang, M., Tao, J., Wang, J., Ren, C., Qi, Q., Xiang, Q. Y., & Huang, H. (2014). Adaptive and non-adaptive genome size evolution in Karst endemic flora of China. *New Phytologist*, *202*(4) 1371–1381.

Kashkush, K., Feldman, M., & Levy, A. A. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics*, *160*(4), 1651–1659.

Kloot, P. M. (1983). Early records of alien plants naturalized in South Australia. *Journal of the Adelaide Botanic Gardens*, 6, 93-131.

Knight, C. A., Molinari, N. A., & Petrov, D. A. (2005). The large genome constraint hypothesis: Evolution, ecology and phenotype. A*nnals of Botany*, *95*(1), 177–190.

Kryvokhyzha, D., Holm, K., Chen, J., Cornille, A., Glémin, S., Wright, S. I., ... & Lascoux, M. (2016): The influence of population structure on gene expression and flowering time variation in the ubiquitous weed C*apsella bursa-pastoris* (Brassicaceae). *Molecular Ecology*, *25*(5), 1106-1121.

Kryvokhyzha, D., Salcedo, A., Eriksson, M. C., Duan, T., Tawari, N., Chen, J., ... & Stinchcombe, J. R. (2019). Parental legacy, demography, and admixture influenced the evolution of the two subgenomes of the tetraploid *Capsella bursa-pastoris* (Brassicaceae). *PLoS genetics*, *15*(2), e1007949.

Lavergne, S., Muenke, N. J., & Molofsky, J. (2010). Genome size reduction can trigger rapid phenotypic evolution in invasive plants. *Annals of Botany*, *105*(1), 109–116.

Lawson, D. J., Van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature communications*, *9*(1), 3258.

Levin, D. (2002). *The role of chromosomal change in plant evolution*. Oxford University Press.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., … & Nordborg, M. (2013). Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics*, *45*(8), 884–890.

Lysak, M. A., Koch, M. a., Beaulieu, J. M., Meister, A., & Leitch, I. J. (2009). The dynamic ups and downs of genome size evolution in Brassicaceae. *Molecular Biology and Evolution, 26(1), 85–98.*

Mack, R. N. & Lonsdale, W. M. (2001). Humans as Global Plant Dispersers: Getting More Than We Bargained For: Current introductions of species for aesthetic purposes present the largest single challenge for predicting which plant immigrants will become future pests. *AIBS Bulletin*, *51*(2), 95-102.

Meagher, T. R., & Vassiliadis, C. (2005). Phenotypic impacts of repetitive DNA in flowering plants. *New Phytologist*, *168*(1), 71–80.

Mooney, H. A., Mack, R. N., McNeely, J. A., Neville, L. E., Schei, P. J., Waage, J. K. (2005). I*nvasive alien species*. Washington, D.C.: Island Press.

Muñoz-Diez, C., Vitte, C., Ross-Ibarra, J., Gaut, B. S., & Tenaillon, M. I. (2012). Using nextgen sequencing to investigate genome size variation and transposable element content. In *Plant transposable elements* (pp. 41-58). Springer, Berlin, Heidelberg.

Neuffer, B. (1996). RAPD analyses in colonial and ancestral populations of *Capsella bursa-pastoris* (L.) Med. (Brassicaceae). *Biochemical Systematics and Ecology*, *24*(5), 393-403.

Neuffer B, Bartelheim S (1989): Gen-ecology of *Capsella bursa-pastoris* from an altitudinal transsect in the Alps. *Oecologia*, *81*(4), 521-527.

Neuffer, B., Bernhardt, K. G., Hurka, H., & Kropf, M. (2011). Monitoring population and gene pool dynamics of the annual species *Capsella bursa- pastoris* (Brassicaceae): a review of relevant species traits and the initiation of a long- term genetic monitoring programme. B*iodiversity and Conservation*, *20*(2), 309-323.

Neuffer, B., Hirschle, S., Jäger, S. (1999). The colonizing history of Capsella in Patagonia (South America) – Molecular and adaptive significance. *Folia Geobotanica*, *34*(4), 435-450.

Neuffer, B., Linde, M. (1999). Capsella bursa-pastoris - Colonization and adaptation; a globetrotter conquers the world. In: van Raamsdonk, L. W. D., den Nijs, J.C.M. (eds.) Plant Evolution in Man-Made Habitats. *Proceedings of the VIIth International IOPB Symposium 1998*, 49-72.

Neuffer, B., & Hurka, H. (1999). Colonization history and introduction dynamics of *Capsella bursa-*pastoris (Brassicaceae) in North America: Isozymes and quantitative traits. *Molecular Ecology, 8*(10), 1667–1681.

Neuffer, B. (2002). Der weltweite Erfolg des Hirtentäschelkrautes *Capsella bursa-pastoris* (L.) Medik. *Biologische Invasionen. Herausforderung Zum Handeln?,* 235–255.

Neuffer, B., & Hurka, H. (1988). Germination behaviour in populations of *Capsella bursa-pastoris* (Cruciferae). *Plant Systematics and Evolution*, *161*(1–2), 35–47.

Raj, B. (1965). Chromosome numbers in some Indian Angiosperms - II. *Proceedings of the Indian Academy of Sciences - Section B*, *61*(5), 253-261.

Randall, R. P. (2012). *A global compendium of weeds.* Department of Agriculture and Food Western Australia, (Ed.2).

Rayburn, A. L., Dudley, J. W., & Biradar, D. P. (1994). Selection for early flowering results in simultaneous selection for reduced nuclear-DNA content in maize. *Plant Breeding*, *112*(4), 318–322.

Sakai, A. K., Allendorf, F. W., Holt, J. S., Lodge, D. M., Molofsky, J., With, K. A., ... & McCauley, D. E. (2001). The population biology of invasive species. *Annual review of ecology and systematics*, *32*(1), 305-332.

Shull, G. H. (1909). *Bursa bursa-pastoris* and B*ursa heegeri* biotypes and hybrids. Carnegie institution of Washington.

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. *American Journal of Botany*, *93*(11), 1714–1724.

Slotte, T., Huang, H., Lascoux, M., & Ceplitis, A. (2008). Polyploid speciation did not confer instant reproductive isolation in *Capsella* (Brassicaceae). *Molecular Biology and Evolution*, *25*(7), 1472-1481.

Šmarda, P., & Bureš, P. (2010). Understanding intraspecific variation in genome size in plants. *Preslia*, *82*(1), 41-61.

Soltis, D. E., Soltis, P. S., Bennett, M. D., & Leitch, I. J. (2003). Evolution of genome size in the angiosperms. *American Journal of Botany*, *90*(11), 1596–1603.

Suda, J., Meyerson, L. A., Leitch, I. J., & Py, P. (2014). The hidden side of plant invasions: the role of genome size. *New Phytologist*, *205*(3), 994– 1007.

Vitousek, P. M., D’Antonio, C. M., Loope, L. L., Westbrooks, R. (1996). Biological invasions as global environmental change. *American Scientist, 84*, 218–28.

Yokoya, K., Roberts, A. V., Mottley, J., Lewis, R., & Brandham, P. E. (2000). Nuclear DNA amounts in roses. *Annals of Botany*, *85*(4), 557-561.

Zhou, T. Y., Lu, L. L., Yang, G., Al-Shehbaz, I. A. (2001). Brassicaceae (Cruciferae). In: Wu, Z. Y., Raven, P. H., eds. *Flora of China 8*. Beijing, StLouis: Science Press/Missouri Botanical Garden Press, 1–193.

2. Genome-wide association mapping

In order to find out whether and which SNPs are subject to natural selection, a GWAS was applied. The program *BEAGLE* uses an imputation algorithm to replenish missing data and calculates a statistically supposable genotype for the relevant loci (Browning & Browning, 2007; 2016). Data imputation is a crucial step before GWAS. *BEAGLE* version *4.1* was executed on the filtered vcf-file before filtering for MAF = 0.05 (see *5.2.12 Quality calculations and filtering*) with the following parameters: impute = true, nthreads = 20, window = 200, overlap = 5. This resulted in a vcf-file containing filtered and imputed 82,370 SNPs. After imputation, the file was filtered with *PLINK* version *1.90b3.38* (Purcell et al., 2007) for MAF = 0.01, retaining 25,851 SNPs.

GWAS itself was performed using *EMMAX* version *07032010* (Kang et al., 2010). EMMAX uses a statistical process (EMMA algorithm) for large scale association mapping accounting for the sample structure. The general principle of GWAS is to find associations between the SNP genotypes and the recorded phenotypes provided in seperate text files (each phenotype has to be tested separatedly). If genotypes coincide with a certain phenotype because of population structure in the data (because of pedigree and not “true“ locus association), false-positives might occur. Therefore, EMMAX uses an algorithm to mathematically surpress kinship effects. Using the emma-kin kinship matrix algorithm, high correlations coinciding with high degree of pedigree will be mathimatically surpressed.

GWAS was performed with the measured phenotypic parameters from the common garden experiments (see *5.1 Common garden experiment*). Additionally, the bioclimatic variables from Worldclim were used as variables to find SNPs associated with the respective parameter (e.g. Annual mean temperature (BIO1) or isothermality (BIO3); for explaination of all variables see *5.2.11 R-Bioclim.R*). SNPs outside the 16 chromosomes of *C. bursa-pastoris* were excluded from the analysis, retaining 25,833 SNPs. The SNPs were rather evenly distributed across the genome, varying from min = 1,233 SNPs on chromosome 12 and max = 1,985 on chromosome 2 (Fig. 3).

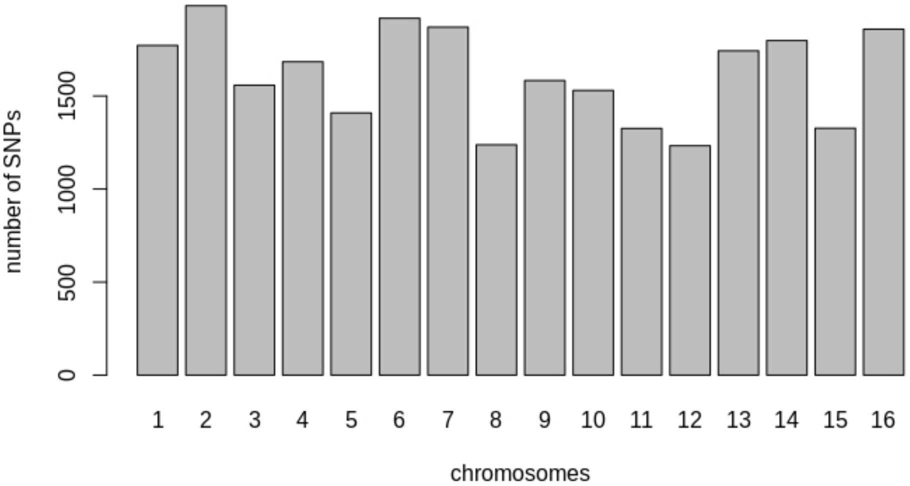

Figure 3: Number of SNPs on each chromosome of *C. bursa-pastoris*.

The flowering time, genome size, germination percentage and number of basal branches showed significant differences between the two obtained clusters from the ADMIXTURE analysis (see Tab. III.1), and some SNPs associated with these parameters could be detected with GWAS, indicating that these loci are under natural selection:

2.1 Flowering time

2 highly significantly associated SNPs (Fig. 4):

- SNP 14:5913350, -log$_{10}$(p) = 8.607
- SNP 15:13308975, -log$_{10}$(p) = 7.623

Several SNPs are associated with the flowering time (Fig. 4). This seems reasonable, since some vernalization sensitive flowering time genes like *FLC*, *FCA* or *FRI* have been described in the literature so far (Lempe et al., 2005; Linde et al., 2001; Shindo et al., 2006; Neuffer et al., 2011). Genes involved in regulation of the circadian clock seem also to play a role in ecotypic differentiation of flowering time in *C. bursa-pastoris* (eg. *CCA1* and *TOC1*; Slotte et al., 2007). Aside from the molecular genetic background, the flowering of the Shepherd’s Purse seems to be basically influenced by day length (Hurka et al., 1976) and temperature conditions, and epigenetic effects may also be involved (Shindo et al., 2006). Punctual flowering is crucial for an annual plant. If a plant starts flowering during the wrong time of the year, seed production is radically reduced, resulting in lower fitness of this particular flowering ecotype. In the worst scenario of the individual, it will be eliminated completely, leading to an allelic shift to ecotypes that flower earlier or later within the population. Therefore, it was much expected to find certain SNPs highly accociated with flowering.

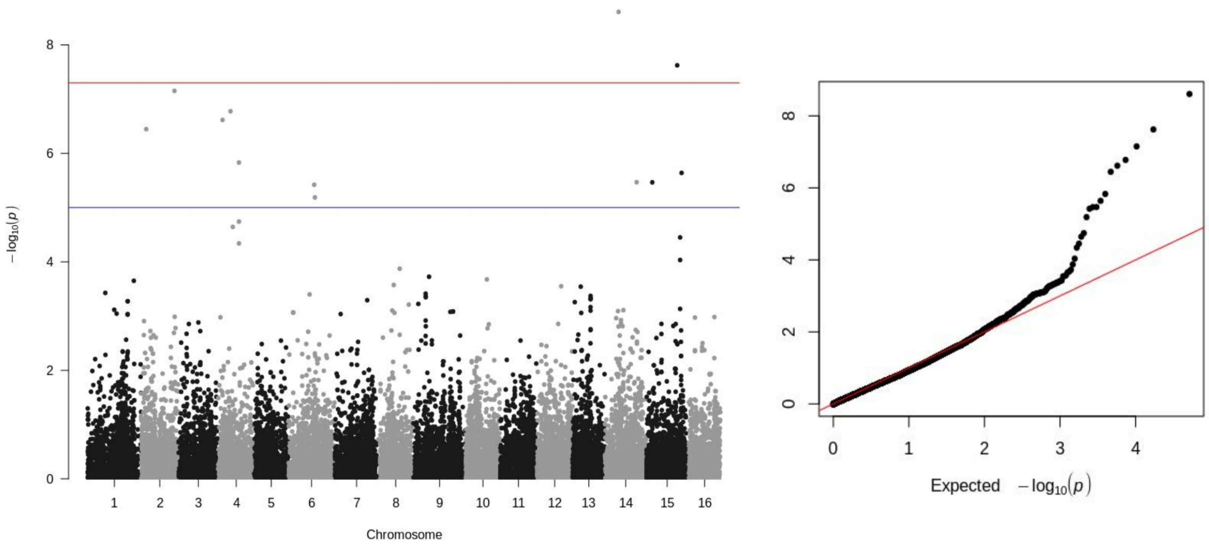

Figure 4: Manhattan and qq-plot of flowering time of *C. bursa-pastoris*.

2.2 Genome size

4 associated SNPs (Fig. 5):

- SNP 9:14673775, -$\log_{10}(p)$ = 5.389
- SNP 7:13244283, -$\log_{10}(p)$ = 5.262
- SNP 9:14622798, -$\log_{10}(p)$ = 5.255
- SNP 9:14622831, -$\log_{10}(p)$ = 5.255
- SNP 9:15192568, -$\log_{10}(p)$ = 5.185

The genome size estimated via FCM varied tremendously within *C. bursa-pastoris* (see *5.1.2.7 Genome size*). According to the results shown here, there are some SNPs associated with the genome size (Fig. 5). However, no SNPs could be found that were above the highly significant threshold, and the qq-plot shows the majority of SNPs deviating from the expected line, indicating either a high number of false positives or cryptic population structure. Since high intraspecific variation of genome size is described as the result of high TE abundance in the literature this result should not be overinterpreted (Šmarda & Bureš, 2010; Muñoz-Diez et al., 2012; see *III. Taking the long way around – Worldwide geographical structure of the cosmopolitan weed Capsella bursa-pastoris (Brassicaceae)*, discussion). A direct influence of SNPs on the genome size seems unlikely. Nevertheless, further studies on putative genes responsible for genome size, might be promising after all.

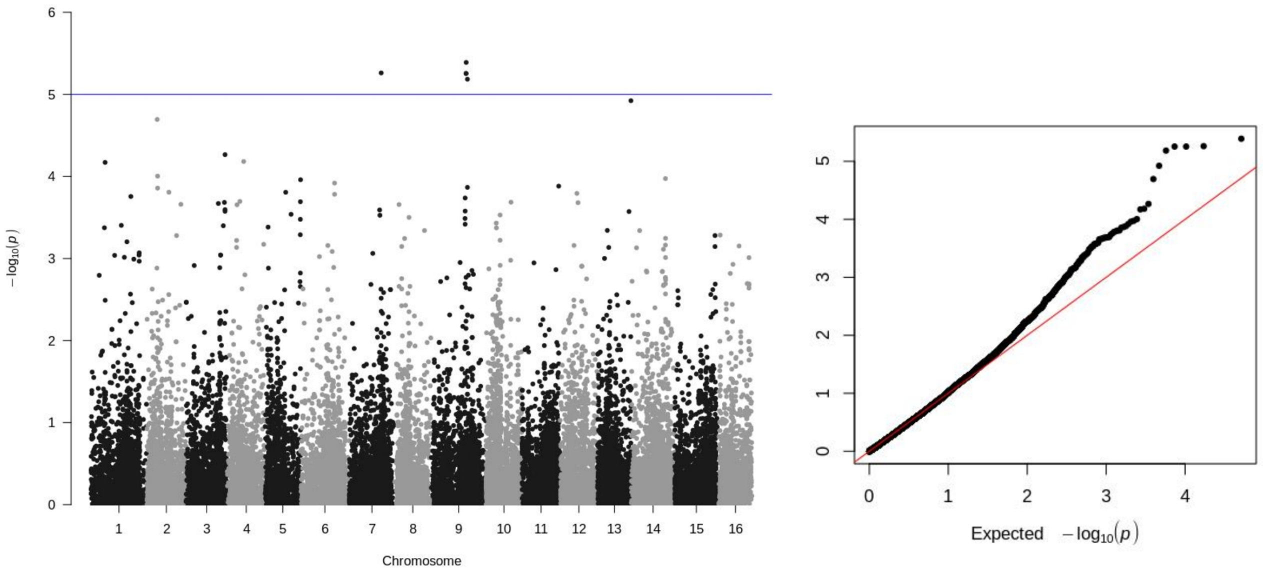

Figure 5: Manhattan and qq-plot of genome size of *C. bursa-pastoris*.

2.3 Germination

1 highly significantly associated SNP (Fig. 6):

- SNP 15:15199637, -$\log_{10}(p)$ = 7.877

One SNP was found to be highly associated with the germination potential (Fig. 6). The *p*-value is convincing and the qq-plot looks quite promising. Although this SNP was far above the significance level, it is rather doubtful that the germination rate is subject to a genetic component, and association of SNPs with the germination percentage has to be interpreted carefully. The seeds were directly taken from the natural collection sites, and the ripening degree at the moment of collection is likely to play a greater role in the germination behaviour than the molecular genetic background of the individuals. It is possible that an optimal harvest time could not always be maintained and that the plants had different substrate conditions. All this influences the germination rate. Therefore, the correlation between the germination behaviour and the cluster affiliation as obtained from the ADMIXTURE analysis (see tab. III.1) might be completely random.

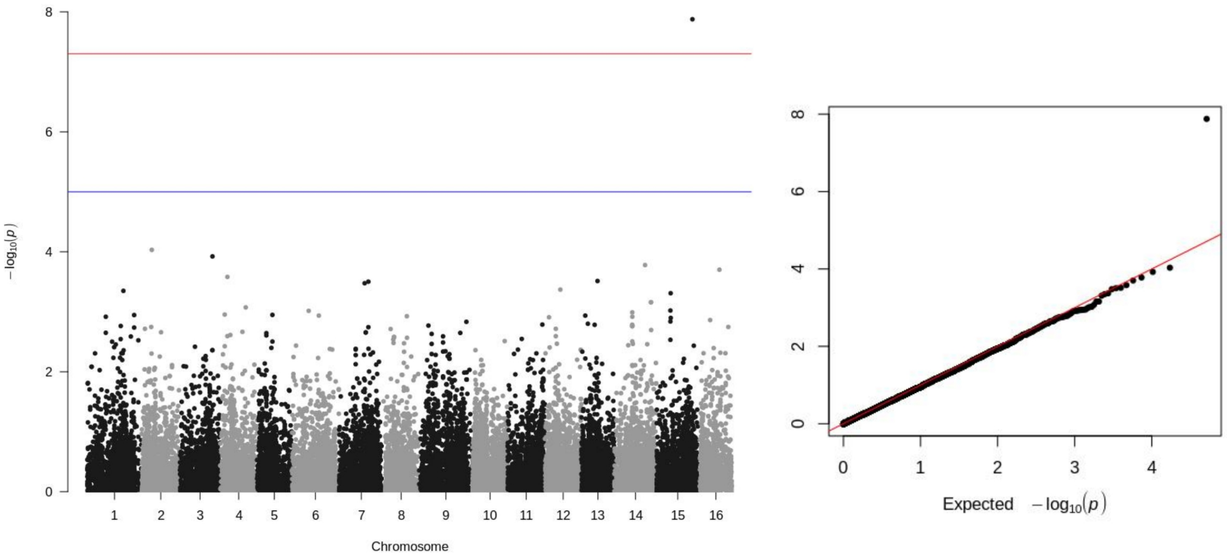

Figure 6: Manhattan and qq-plot of germination percentage of *C. bursa-pastoris*.

2.4 Number of basal branches

5 associated SNPs (Fig. 7):

- SNP 6:11529601, -log$_{10}$(p) = 6.957
- SNP 9:20454277, -log$_{10}$(p) = 6.170
- SNP 2:14467981, -log$_{10}$(p) = 5.471
- SNP 15:13309003, -log$_{10}$(p) = 5.261
- SNP 15:13308975, -log$_{10}$(p) = 5.242

According to the results shown here, 5 SNPs show association with the number of branches of the individuals, but no SNP was above the strict significance level (Fig. 7). The number of branches (or inflorescences) can be used as a parameter of fitness (Cornille et al., 2018), so an association with certain SNPs would not be fallacious, since this parameter showed a significant differences between the two obtained clusters (see tab. III.1). However, the qq-plot shows the majority of SNPs deviating from the expected line, indicating either a high number of false positives or cryptic population structure.

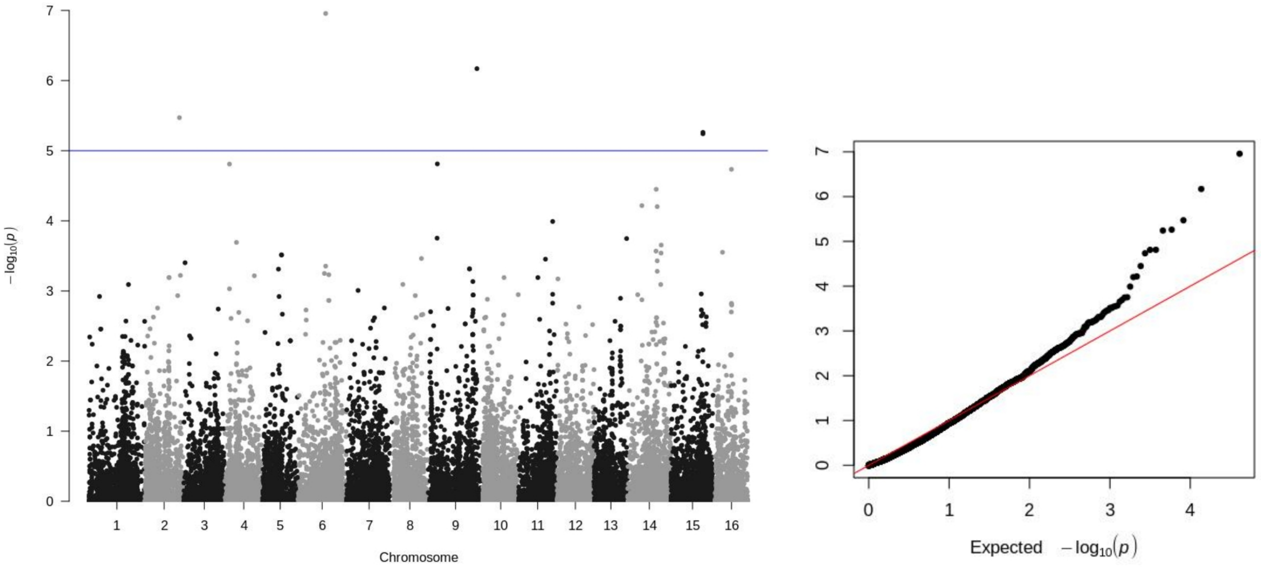

Figure 7: Manhattan and qq-plot of number of branches of *C. bursa-pastoris*.

2.5 Plant height and seed weight

The planth height and the seed weight did not show significant differences between the Eurasian and the Mediterranoid cluster (see Tab. III.1). However, the GWAS revealed the following associated hyplotypes (Fig 8):

Plant height (5 associated SNPs; Fig. 8a):

- SNP 3:13900310, -$\log_{10}(p)$ = 6.062
- SNP 5:2530599, -$\log_{10}(p)$ = 6.062
- SNP 14:13734322, -$\log_{10}(p)$ = 5.184
- SNP 8:1507773, -$\log_{10}(p)$ = 5.057
- SNP 11:14177173, -$\log_{10}(p)$ = 5.043

Seed weight (3 highly significantly associated SNPs; Fig. 8b):

- SNP 11:13245744, -$\log_{10}(p)$ = 11.013
- SNP 4:1929647, -$\log_{10}(p)$ = 8.156
- SNP 12:6626936, -$\log_{10}(p)$ = 7.658

It is not certain to what extent the seed weight in particular can be used as an unbiased characteristic. It is possible that the time of harvest varies too much. To rule this out, the seed harvested after the common garden experiment would have to be weighed. Here at least the substrate and weather conditions were the same for all plants. There might or might not be a genetic component in seed weight, however, the qq-plot is not very convincing (Fig. 8a, right).

Since no SNP was above the highly significant level when considering plant height, it can rather be assumed that the growth is not necessarily genetically determined; at least this data basis does not permit this conclusion here (Fig 8b).

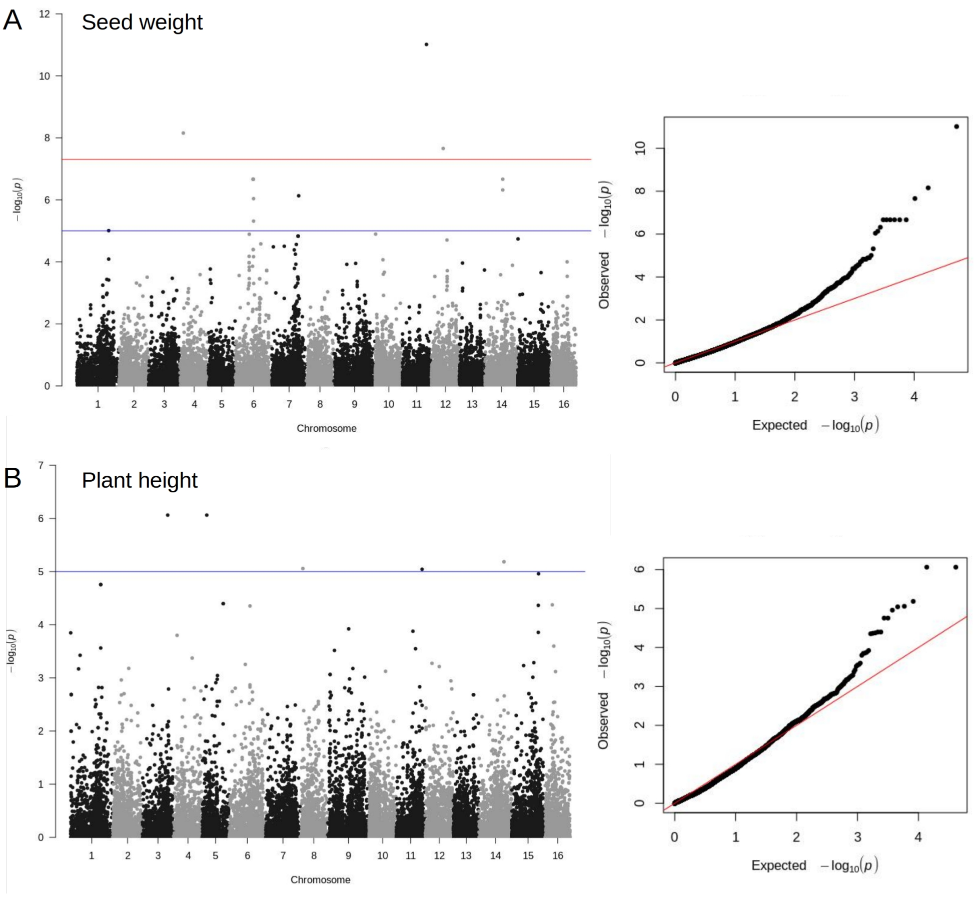

Figure 8: Manhattan and qq-plot of A: seed weight and B: plant height of *C. bursa-pastoris*.

2.6 Leaf shape

Since GWAS was designed for case-control studies, associations can only be performed on presence/absence data. Therefore, each of the four leaf types had to be examined seperatedly (Fig. 9):

Heteris (4 highly significantly associated SNPs; Fig. 9a):

- SNP 2:8227155, -$\log_{10}(p)$ = 16.207
- SNP 9:3236889, -$\log_{10}(p)$ = 11.757
- SNP 5:8079932, -$\log_{10}(p)$ = 11.002
- SNP 15:2823626, -$\log_{10}(p)$ = 7.968

Simplex (4 associated SNPs; Fig. 9b):

- SNP 3:9452889, -$\log_{10}(p)$ = 7.095
- SNP 11:11994513, -$\log_{10}(p)$ = 6.374
- SNP 2:8227155, -$\log_{10}(p)$ = 6.347
- SNP 5:8079932, -$\log_{10}(p)$ = 5.206

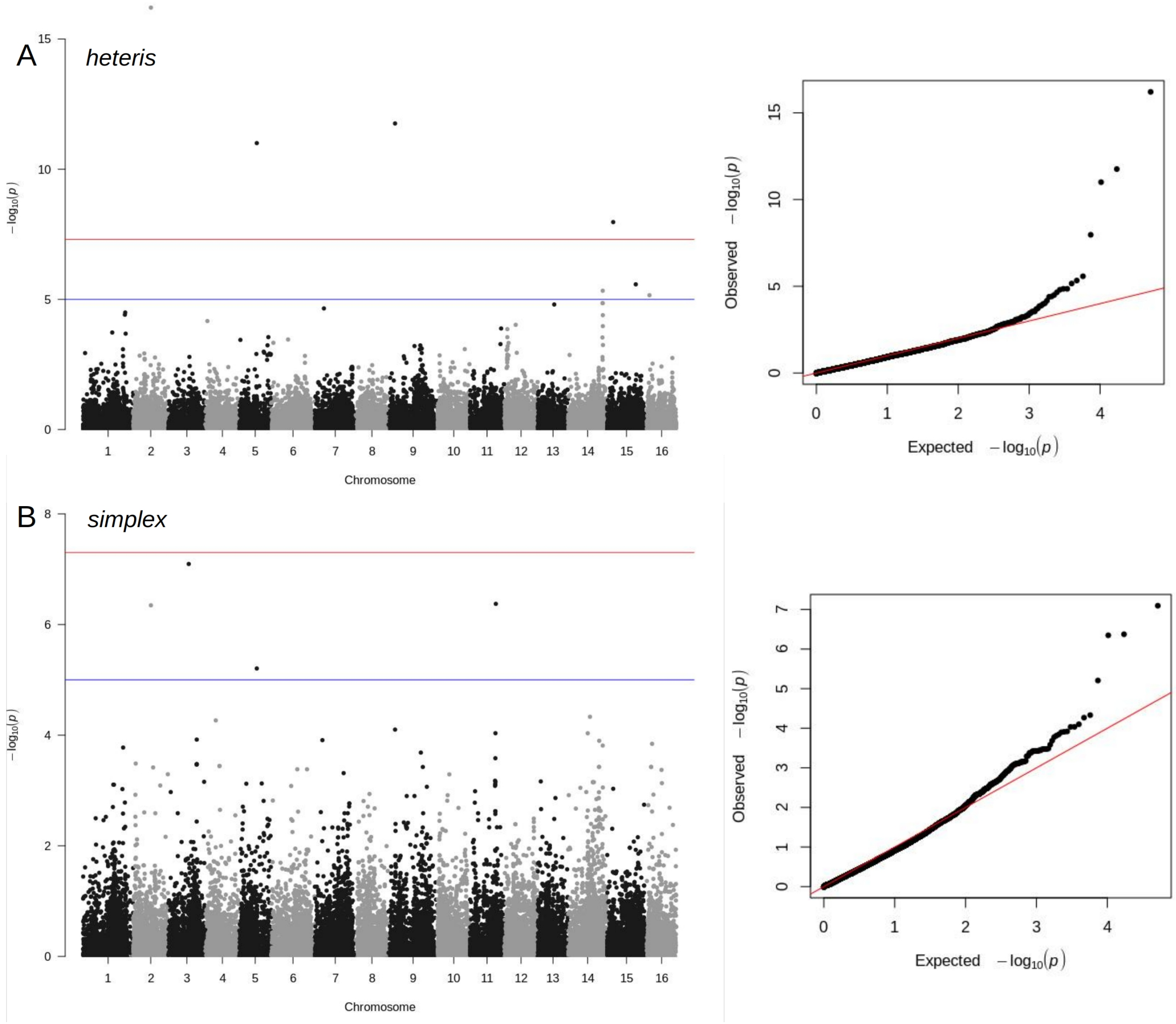

Figure 9: Manhattan and qq-plots of the different leaf types of *C. bursa-pastoris*. A: *heteris*. B: *simplex*.

Tenuis (3 associated SNPs; Fig. 9c):

- SNP 14:2135722, $-\log_{10}(p)$ = 5.997
- SNP 14:11331800, $-\log_{10}(p)$ = 5.879
- SNP 13:5185253, $-\log_{10}(p)$ = 5.316

Rhomboidea (1 associated SNP; Fig. 9d):

- SNP 12:6626936, $-\log_{10}(p)$ = 5.758

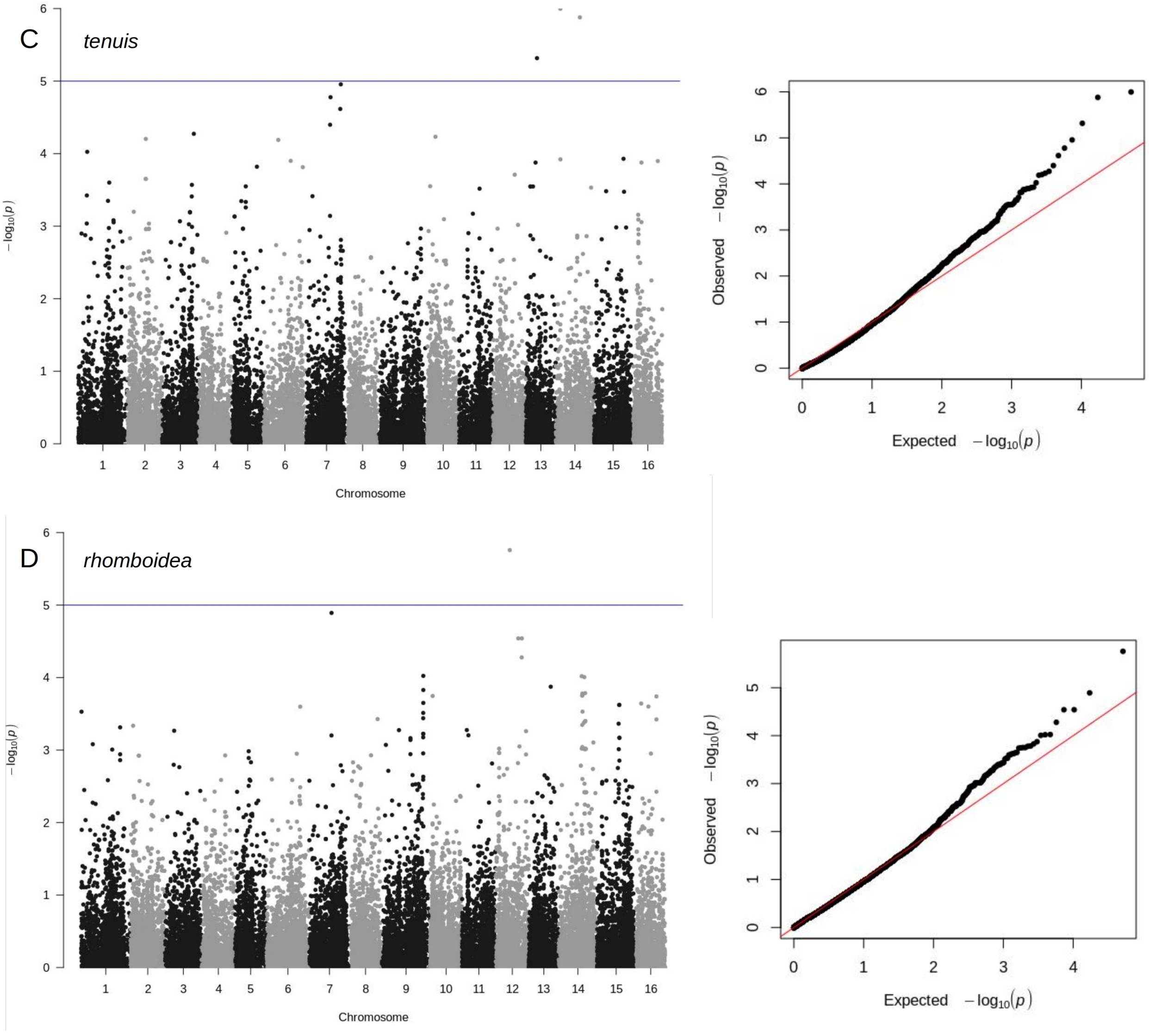

Figure 9: Manhattan and qq-plots of the different leaf types of *C. bursa-pastoris*. C: *tenuis*. D: *rhomboidea*.

If the leaf shape of *C. bursa-pastoris* would not be an adaptive trait, one would expect an even distribution of the leaf types. However, this study showed that this is not the case (see *5.1.2.4 Leaf shapes*). The leaf shape of the Shepherd‘s Purse plays a role, and several genes are hypothesized to be involved in the leaf development: The mendelian inheritance, in which two loci with two alleles are mainly involved in the leaf shapes, has been known for 100 years (Shull, 1909; 1911; see *I. The role of ecotypic variation in driving worldwide colonization by a cosmopolitan plant*, Fig. I.1).

In addition, other factors have been described in the past which have an influence on the leaf margin of *Capsella*: Among others, *REDUCED COMPLEXITY* (*RCO-A* and *RCO-B*) seems to be involved in the dissection of the leaves (Sicard et al., 2014). Surprisingly, only two SNPs were found in this GWAS that were highly associated with more than one of the leaf types tested (2:8227155 and 5:8079932 for *heteris* and *simplex* each). If a limited number of candidate genes were involved, it would have been expected that the same coordinates would occur with high association values for all leaf forms. However, the qq-plots of this GWAS indicate that these results should be treated with caution. Nontheless, the complete molecular basis of the leaf is yet to be discovered, and it is an interesting question which of the sequenced SNPs are associated with the leaf types. To identify the exact genes affected by these highly associated SNPs, further studies have to be done.

2.7 Climate variables and coordinates

GWAS was also performed with the geographical coordinates and altitude of the source area of the plants as well as the bioclimatic variables as “phenotypes”. The Manhattan and qq-plots of these GWAS results can be found in the sigital supplement of this thesis.

Altitude (5 highly significantly associated SNPs):

- SNP 12:6626936, $-\log_{10}(p) = 73.615$
- SNP 1:485815, $-\log_{10}(p) = 18.665$
- SNP 1:18109736, $-\log_{10}(p) = 9.561$
- SNP 1:18535379, $-\log_{10}(p) = 8.406$
- SNP 9:20672698, $-\log_{10}(p) = 7.871$

Longitutde (>15 highly significantly associated SNPs)

Old World vs. New World (4 highly significantly associated SNPs):

- SNP 15:15199637, $-\log_{10}(p) = 46.384$
- SNP 15:12445415, $-\log_{10}(p) = 7.840$
- SNP 15:12445373, $-\log_{10}(p) = 7.776$
- SNP 15:12476704, $-\log_{10}(p) = 7.584$

Latitude (15 highly significantly associated SNPs):

- SNP 15:15199637, $-\log_{10}(p) = 23.190$
- SNP 7:14404861, $-\log_{10}(p) = 11.824$
- SNP 1:11543999, $-\log_{10}(p) = 9.924$
- SNP 8:2264297, $-\log_{10}(p) = 9.485$
- SNP 9:7256770, $-\log_{10}(p) = 9.385$
- SNP 6:6967676, $-\log_{10}(p) = 8.775$
- SNP 5:11502211, $-\log_{10}(p) = 8.196$
- SNP 5:13569473, $-\log_{10}(p) = 8.114$
- SNP 4:11642182, $-\log_{10}(p) = 8.043$
- SNP 5:11271627, $-\log_{10}(p) = 7.915$
- SNP 4:5388860, $-\log_{10}(p) = 7.898$
- SNP 7:10828335, $-\log_{10}(p) = 7.881$
- SNP 15:13246459, $-\log_{10}(p) = 7.861$
- SNP 7:11910153, $-\log_{10}(p) = 7.478$
- SNP 9:8456679, $-\log_{10}(p) = 7.302$

BIO1 (Annual Mean Temperature;
9 highly significantly associated SNPs):

- SNP 6:9240484, -log$_{10}$(p) = 11.798
- SNP 7:11910153, -log$_{10}$(p) = 8.566
- SNP 12:6626936, -log$_{10}$(p) = 8.247
- SNP 3:9448644, -log$_{10}$(p) = 8.022
- SNP 5:10315237, -log$_{10}$(p) = 7.948
- SNP 6:16197632, -log$_{10}$(p) = 7.849
- SNP 4:7440628, -log$_{10}$(p) = 7.845
- SNP 8:13172888, -log$_{10}$(p) = 7.538
- SNP 16:1158642, -log$_{10}$(p) = 7.104

BIO2 (Mean Diurnal Range;
3 highly significantly associated SNPs):

- SNP 6:9240484, -log$_{10}$(p) = 9.033
- SNP 16:1158642, -log$_{10}$(p) = 8.585
- SNP 12:6626936, -log$_{10}$(p) = 7.973

BIO3 (Isothermality; 8 highly
significantly associated SNPs):

- SNP 12:6626936, -log$_{10}$(p) = 82.310
- SNP 1:485815, -log$_{10}$(p) = 24.104
- SNP 8:9819672, -log$_{10}$(p) = 10.892
- SNP 1:18109736, -log$_{10}$(p) = 10.855
- SNP 1:18535379, -log$_{10}$(p) = 9.487
- SNP 12:7415395, -log$_{10}$(p) = 8.278
- SNP 1:2487645, -log$_{10}$(p) = 7.454
- SNP 9:20672698, -log$_{10}$(p) = 7.418

BIO4 (Temperature Seasonality; 8 highly
significantly associated SNPs):

- SNP 12:6626936, -log$_{10}$(p) = 42.614
- SNP 1:485815, -log$_{10}$(p) = 13.192
- SNP 7:8624829, -log$_{10}$(p) = 9.206
- SNP 15:9812857, -log$_{10}$(p) = 9.103
- SNP 15:2756773, -log$_{10}$(p) = 8.622
- SNP 4:7440628, -log$_{10}$(p) = 8.471
- SNP 6:9240484, -log$_{10}$(p) = 7.970
- SNP 16:1158642, -log$_{10}$(p) = 7.661

BIO5 (Max Temperature of Warmest Month;
12 highly significantly associated SNPs):

- SNP 12:6626936, -log$_{10}$(p) = 30.048
- SNP 6:9240484, -log$_{10}$(p) = 10.916
- SNP 1:485815, -log$_{10}$(p) = 10.070
- SNP 5:13569473, -log$_{10}$(p) = 9.837
- SNP 16:1158642, -log$_{10}$(p) = 9.461
- SNP 7:11910153, -log$_{10}$(p) = 9.342
- SNP 8:1072714, -log$_{10}$(p) = 9.182
- SNP 4:6106641, -log$_{10}$(p) = 9.170
- SNP 9:7256770, -log$_{10}$(p) = 8.943
- SNP 5:10315237, -log$_{10}$(p) = 8.925
- SNP 6:12429788, -log$_{10}$(p) = 8.011
- SNP 16:7771671, -log$_{10}$(p) = 7.526

BIO6 (Min Temperature of Coldest Month;
10 highly significantly associated SNPs):

- SNP 4:7440628, -log$_{10}$(p) = 14.919
- SNP 15:9812857, -log$_{10}$(p) = 12.897
- SNP 6:16197632, -log$_{10}$(p) = 12.240
- SNP 7:8624829, -log$_{10}$(p) = 11.982
- SNP 6:6589152, -log$_{10}$(p) = 11.464
- SNP 15:2756773, -log$_{10}$(p) = 10.467
- SNP 1:17865163, -log$_{10}$(p) = 10.093
- SNP 14:5734937, -log$_{10}$(p) = 9.487
- SNP 8:12345623, -log$_{10}$(p) = 9.222
- SNP 5:4416626, -log$_{10}$(p) = 7.768

BIO7 (Temperature Annual Range;
5 highly significantly associated SNPs):

- SNP 12:6626936, $-\log_{10}(p) = 32.451$
- SNP 1:485815, $-\log_{10}(p) = 10.908$
- SNP 16:1158642, $-\log_{10}(p) = 8.808$
- SNP 6:9240484, $-\log_{10}(p) = 8.686$
- SNP 7:8624829, $-\log_{10}(p) = 7.416$

BIO8 (Mean Temperature of Wettest Quarter;
4 highly significantly associated SNPs):

- SNP 6:9240484, $-\log_{10}(p) = 33.663$
- SNP 16:1158642, $-\log_{10}(p) = 19.252$
- SNP 6:12429788, $-\log_{10}(p) = 9.354$
- SNP 8:5642924, $-\log_{10}(p) = 8.939$

BIO9 (Mean Temperature of Driest Quarter;
1 highly significantly associated SNPs):

- SNP 12:6626936, $-\log_{10}(p) = 16.756$

BIO10 (Mean Temperature of Warmest Quarter;
13 highly significantly associated SNPs):

- SNP 12:6626936, $-\log_{10}(p) = 29.015$
- SNP 6:9240484, $-\log_{10}(p) = 15.889$
- SNP 16:1158642, $-\log_{10}(p) = 11.651$
- SNP 1:485815, $-\log_{10}(p) = 10.411$
- SNP 6:12429788, $-\log_{10}(p) = 9.535$
- SNP 7:11910153, $-\log_{10}(p) = 9.112$
- SNP 5:10315237, $-\log_{10}(p) = 8.479$
- SNP 5:13569473, $-\log_{10}(p) = 8.065$
- SNP 16:2893685, $-\log_{10}(p) = 7.858$
- SNP 9:7256770, $-\log_{10}(p) = 7.700$
- SNP 8:5642924, $-\log_{10}(p) = 7.557$
- SNP 8:1072714, $-\log_{10}(p) = 7.415$
- SNP 1:18535379, $-\log_{10}(p) = 7.351$

BIO11 (Mean Temperature of Coldest Quarter;
12 highly significantly associated SNPs):

- SNP 4:7440628, $-\log_{10}(p) = 14.713$
- SNP 15:9812857, $-\log_{10}(p) = 12.982$
- SNP 6:16197632, $-\log_{10}(p) = 12.262$
- SNP 7:8624829, $-\log_{10}(p) = 12.026$
- SNP 6:6589152, $-\log_{10}(p) = 10.952$
- SNP 15:2756773, $-\log_{10}(p) = 10.933$
- SNP 1:17865163, $-\log_{10}(p) = 9.835$
- SNP 14:5734937, $-\log_{10}(p) = 9.678$
- SNP 8:12345623, $-\log_{10}(p) = 9.536$
- SNP 4:4279152, $-\log_{10}(p) = 7.540$
- SNP 5:4416626, $-\log_{10}(p) = 7.521$
- SNP 8:12345859, $-\log_{10}(p) = 7.325$

BIO12 (Annual Precipitation;
>15 highly significantly associated SNPs)

BIO13 (Precipitation of Wettest Month;
>15 highly significantly associated SNPs)

BIO14 (Precipitation of Driest Month;
12 significantly associated SNPs):

- SNP 4:5388860, $-\log_{10}(p) = 6.846$
- SNP 8:2264297, $-\log_{10}(p) = 6.207$
- SNP 8:2336264, $-\log_{10}(p) = 6.108$
- SNP 4:6497016, $-\log_{10}(p) = 6.020$
- SNP 4:6106605, $-\log_{10}(p) = 5.946$
- SNP 11:11352273, $-\log_{10}(p) = 5.848$
- SNP 12:11744648, $-\log_{10}(p) = 5.831$
- SNP 8:7130753, $-\log_{10}(p) = 5.653$
- SNP 4:6106641, $-\log_{10}(p) = 5.343$
- SNP 4:8893622, $-\log_{10}(p) = 5.162$
- SNP 2:5871116, $-\log_{10}(p) = 5.138$
- SNP 8:2336253, $-\log_{10}(p) = 5.119$

BIO15 (Precipitation Seasonality;
8 highly significantly associated SNPs):

- SNP 12:6626936, $-\log_{10}(p)$ = 11.724
- SNP 9:8456679, $-\log_{10}(p)$ = 9.285
- SNP 5:7822853, $-\log_{10}(p)$ = 8.945
- SNP 15:13246459, $-\log_{10}(p)$ = 8.702
- SNP 7:14404861, $-\log_{10}(p)$ = 8.421
- SNP 5:11502211, $-\log_{10}(p)$ = 7.934
- SNP 14:15975442, $-\log_{10}(p)$ = 7.656
- SNP 1:485815, $-\log_{10}(p)$ = 7.414

BIO16 (Precipitation of Wettest Quarter;
>15 highly significantly associated SNPs)

BIO17 (Precipitation of Driest Quarter;
10 significantly associated SNPs):

- SNP 4:5388860, $-\log_{10}(p)$ = 6.454
- SNP 8:2264297, $-\log_{10}(p)$ = 6.237
- SNP 12:11744648, $-\log_{10}(p)$ = 6.080
- SNP 12:6626936, $-\log_{10}(p)$ = 6.055
- SNP 8:7130753, $-\log_{10}(p)$ = 5.947
- SNP 4:6497016, $-\log_{10}(p)$ = 5.841
- SNP 4:6106605, $-\log_{10}(p)$ = 5.695
- SNP 11:11352273, $-\log_{10}(p)$ = 5.660
- SNP 8:2336264, $-\log_{10}(p)$ = 5.553
- SNP 4:6106641, $-\log_{10}(p)$ = 5.053

BIO18 (Precipitation of Warmest Quarter;
>15 highly significantly associated SNPs)

BIO19 (Precipitation of Coldest Quarter;
11 highly significantly associated SNPs):

- SNP 8:7772138, $-\log_{10}(p)$ = 14.933
- SNP 8:7772192, $-\log_{10}(p)$ = 13.093
- SNP 5:3957261, $-\log_{10}(p)$ = 12.002
- SNP 11:11994513, $-\log_{10}(p)$ = 10.351
- SNP 4:4815357, $-\log_{10}(p)$ = 9.630
- SNP 3:5651295, $-\log_{10}(p)$ = 9.371
- SNP 3:9452889, $-\log_{10}(p)$ = 8.715
- SNP 7:15272834, $-\log_{10}(p)$ = 7.859
- SNP 7:16925691, $-\log_{10}(p)$ = 7.756
- SNP 1:12686276, $-\log_{10}(p)$ = 7.507
- SNP 6:9240484, $-\log_{10}(p)$ = 7.399

In my opinion, the most interesting results were provided by the GWAS using different climate parameters and altitudinal ranges. Climate probably has the greatest influence on natural selection, as the survival and reproduction probability of plants depends most on temperature and humidity. Highly associated SNPs could be identified for almost all climate variables here. The fact that altitude in particular is also decisive was shown, for example, in a study in which the genetic adaptation of humans to altitude could be demonstrated, which in this case was due to the different oxygen content at different altitudes (Beall et al., 2010).

Finding out which gene is influenced by SNPs is not entirely trivial, since the SNPs do not necessarily have to be located in the corresponding gene. Therefore, it is not enough to know the SNP's position in the genome. If one is looking for certain genes that are targeted by this SNP, one has to consider loci that are in linkage disequilibrium (LD) with it (LD approaches 1: complete link, LD approaches 0: no link). Interesting SNPs in the proximity (50kb window up- and downstream) should have a LD of > 0.7 and can be examined in an annotated reference genome. The window of interest is dependant on the expected "LD decay”, since SNPs that are somehow linked to each other lose this linkage as the spatial distance within the genome increases. In outcrossing species, one would expect a faster LD decay due to recombination.

One single SNP can affect several phenotypes (pleiotropy). Some SNPs had high associations with more than one tested parameter (Tab. 2):

Table 2: Phenotypes with highly associated SNPs.

SNP	Parameters
12:6626936	Altitude, BIO10, 15, 17, 1, 2, 3, 4, 5, 7, 9, *rhomboidea*, seedweight
1:485815	BIO10, 3, 4, 5, 7, 9, altitude
6:16197632	BIO11, 12, 13, 16, 18, 1, 6
6:9240484	BIO7, 8, 10, 1, 2, 5
16:1158642	BIO5, 7, 8, 10, 2
4:7440628	BIO4, 6, 7, 9, 11
15:15199637	germination, latitude, longitude
7:11910153	BIO10, 1, 5, 9
7:8624829	BIO11, 4, 6, 7
11:7800670	BIO13, 16, 18
1:18575244	BIO13, 16, 18
15:13246459	BIO15, *heteris*, longitude
15:2756773	BIO11, 4, 6
15:9812857	BIO11, 4, 6
4:15137325	BIO12, 13, 16
6:12062108	BIO16, 18, 12
8:2264297	BIO17, latitude, BIO14
11:10890643	BIO18, 13
11:11994513	BIO19, *simplex*
1:18109736	Altitude, BIO3
1:18535379	Altitude, BIO3
14:13958632	BIO12, 16
14:15696417	BIO12, 13
14:15975442	*heteris*, longitude
15:13308975	branches, flowering
2:100740	BIO2, 8
2:14467981	branches, flowering
2:8227155	*heteris*, *simplex*
4:4815357	*simplex*, BIO19
4:5388860	BIO14, 17
4:6106641	BIO2, 9
4:6497016	BIO14, 17
5:11502211	longitude, BIO15
5:8079932	*simplex*, *heteris*
6:12429788	BIO10, 8
6:6589152	BIO6, 11
7:14404861	BIO15, latitude

In order to find the exact genes influenced by these SNPs, further investigations have to be carried out. However, it is not always possible to find coherent SNPs using RADseq, as this sequencing technique is a RRLS technique.

3. Conclusions

The work proposed here generates a comprehensive picture of phenotypic diversity in relationship to genetic variation within *Capsella bursa-pastoris*. With the novel RADseq method it was possible to perform population genetic studies of unprecedented depth and complexity and allowed the exploration of evolutionary history, range expansion and invasion patterns of this plant species.

3.1 Evolutionary history of the Shepherd's Purse

C. bursa-pastoris originated 100 – 300 kya from the hybridization between an ancestral *C. orientalis* and an ancestor from the *C. grandiflora/rubella* lineage according to the current literature (Douglas et al., 2015). However, the exact origination area is still quite controversially disputed. It has been argued that this species originated in the Eurasian steppe belt (Hurka et al., 2012) whereas Slotte et al. (2006) and Douglas et al. (2015) discuss an east Mediterranean origin. Due to the very clear adaptation of the here found two subpopulations to two very different climatic locations, this thesis suggests that the allopolyploid *C. bursa-pastoris* may have emerged as a species several times, at least once in the temperate climate of the Central Asian steppe and at least once in the warmer climates of the European Mediterranean.

3.2 Worldwide population structure of *Capsella bursa-pastoris*

The present-day variation enables the investigation of population structure and demographic history of *C. bursa-pastoris*. The colonization success of this species is achieved by remarkable adaptability and ecotypic variation, and the colonization process was supported anthropogenically, when the Europeans settled the New World. This is clearly evident from herbarium records. Earlier studies already postulated the dispersal of certain genotypes by certain groups of people who introduced their native plants into new settlement areas, in particular the Spaniards in certain parts of the New World (e.g. Hurka & Neuffer, 1999; Neuffer & Linde, 1999). Nevertheless, this thesis shows that an environmental filter must have had a strong impact on the global population structure. The clearly visible adaptation to cold and warm climates of the two subpopulations of the Shepherd's Purse indicates that certain pre-adapted ecotypes have been able to establish themselves at the new locations where they found similar climate conditions. However, this work also shows that there have been new adaptations in the New World (Founder Effect).

Population structure analyses of *C. bursa-pastoris* have been done before in the literature. One recent study found three prevalent populations within this species: “European“ (EUR, *n* = 76 individuals), “Middle Eastern“ (ME, *n* = 42) and “Asian“ (ASI, *n* = 143 individuals) (Cornille et al., 2016). However, sampling numbers are comparably low, and only a few localities have been sampled by this working group (Cornille et al., 2016). We present here a more extensive sampling from every continent except Antarctica. The fact that we received $K = 2$ with two different marker systems (isozymes in *II. Geographical structure of genetic diversity in Shepherd‘s Purse, Capsella bursa-pastoris – a global perspective* and SNPs in *III. Taking the long way around – Worldwide geographical structure of the cosmopolitan weed Capsella bursa-pastoris (Brassicaceae)*) shows that the Shepherd‘s Purse probably consists of two main populations instead of three. Admittedly, when both results of both working groups are compared, it is apparent that the other group probably found similar cluster affiliations than we did with $K = 4$ (see Fig. III.2h), because their dataset comprises mainly of individuals from middle and western Europe, southeastern China and eastern Russia, which is mostly a subset of our dataset. The fact that the other study lacks samples from Central Asia, South America and Australia shows an obvious sampling gap and points to a sampling bias.

3.3 Phenotypes and GWAS

The Shepherd’s Purse is a successful colonizer and characterized by great phenotypic plasticity. The results of this work indicate that there are SNPs associated with a variety of phenotypes and therefore likely under the influence of natural selection (see *2. Genome-wide association mapping*). Some SNPs were associated with the onset of flowering, and some genes are already known for this phenotype (see *2.1 Flowering time*). Although several highly associated SNPs were found for the shape of the leaves, they did not always coincide for every leaf type. Further investigations would have to be done in the future. It is also suggested that climate has a major influence on natural selection (see *2.7 Climate variables and coordinates*). In order to identify the genes, however, further investigations have to be carried out.

Although GWAS is a promising new method to find genes responsible for a certain trait expression, there are potential pitfalls in common garden experiments *per se*. The observations from only one garden might be problematic in interpretation if the phenotypic plasticity is quite high between populations from different environments (Williams et al., 2008): If individuals from introduced populations exceed those from native populations, the reverse might be true under different garden conditions. For example, introduced and native populations of *Cynoglossum officinale* differed substantial in size and fecundity between gardens, because they reacted differently to the particular growing conditions (Williams et al., 2008). In this thesis, experiments have been conducted only in Germany, so the explanation for the success of one particular population might not necessarily be true. It would be advisable to carry out comparative common garden experiments at completely different locations in parallel and compare the recorded phenotypes.

4. Literature Cited

Alexander, D. H., Novembre, J., & Lange, K. (2009): Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655-1664.

Almquist, E. B. (1926): Zur Artbildung in der freien Natur. *Acta Horti Bergiani*, *9* (1929), 37-76.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016): Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature ReviewsGenetics*, *17*(2), 81.

Baird, N. A., Etter P. D., Atwood, T. S., Currey, M. C., Shiver A. L., Lewis Z. A., ... &Johnson E. A. (2008): Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one, *3*(10), e3376.

Baker, H. G., & Stebbins, G. L. (Eds.), (1965): *The Genetics of Colonizing Species*. New York: Academic Press.

Barrett, S. C. (2015): Foundations of invasion genetics: the Baker and Stebbins legacy. Molecular Ecology, *24*(9), 1927-1941.

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., ... & Montgomery, H. E. (2010). Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*, *107*(25), 11459-11464.

Berger, J., Suzuki, T., Senti, K. A., Stubbs, J., Schaffner, G., & Dickson, B. J. (2001): Genetic mapping with SNP markers in Drosophila. *Nature Genetics*, *29*(4), 475-481.

Bolger, A. M., Lohse, M., & Usadel, B. (2014): Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bradshaw, A. D. (1965): Evolutionary significance of phenotypic plasticity in plants. In *Advancesin Genetics*, *13*, 115–155.

Browning, S. R., & Browning, B. L. (2016): Genotype imputation with millions of reference samples. *American Journal of Human Genetics 98*(1), 116-126.

Browning, S. R., & Browning, B. L. (2007): Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084-97.

Buffalo, V. (2015): *Bioinformatics Data Skills*. Sebastopol, California: O‘Reilly Media Inc..

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

Coquillat, M. (1951): Sur les plantes les plus communes de la surface du globe. *Bulletin Mensual Societé Linnéenne, Lyon*, *20*(20), 165–170.

Cornille, A., Salcedo, A., Kryvokhyzha, D., Glémin, S., Holm, K., Wright, S. I., & Lascoux, M. (2016): Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (*Capsella bursa-pastoris*). *Molecular Ecology*, *25*(2), 616-629.

Cornille, A., Salcedo, A., Huang, H., Kryvokhyzha, D., Holm, K., Ge, X. J., ... & Lascoux, M. (2018): Local adaptation and maladaptation during the worldwide range expansion of a self-fertilizing plant. *bioRxiv*, 308619.

Cronon, W. (1983): *Changes in the Land* (p. 132). New York: Hill and Wang.

Crosby, A. W. (1986): *Ecological Imperialism. The biological expansion of Europe, 900-1900*. Cambridge University Press, Cambridge.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011): The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

Davey, J.W., & Blaxter, M. L. (2010): RADSeq: next-generation population genetics. *Briefings in functional genomics*, *9*(5-6), 416-423.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011): Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499.

de Villemereuil, P., Gaggiotti, O. E., Mouterde, M., & Till-Bottraud, I. (2016): Common garden experiments in the genomic era: new perspectives and opportunities. *Heredity*, *116*(3), 249.

Doležel, J., Greilhuber, J., & Suda, J. (2007): Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, *2*(9), 2233-2244.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011): A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, *6*(5), e19379.

Garrison, E., & Marth, G. (2012): "Haplotype-based variant detection from short-read sequencing." *arXiv preprint arXiv*:1207.3907.

Harvey, W. H. & Sonder, O. W. (1860): *Flora Capensis*. Vol. 1. Cambridge University Press, Cambridge.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010): Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS genetics*, *6*(2), e1000862.

Holderegger, R., Herrmann, D., Poncet, B., Gugerli, F., Thuiller, W., Taberlet, P., ... & Manel, S. (2008): Land ahead: using genome scans to identify molecular markers of adaptive relevance. *Plant Ecology & Diversity*, *1*(2), 273-283.

Hornbeck, D. (1983): *California patterns: a geographical and historical atlas*. Mayfield Pub Co, 1983.

Hurka, H., & Haase, R. (1982): Seed Ecology of *Capsella bursa-pastoris* (Cruciferae): Dispersal Mechanism and the Soil Seed Bank1. *Flora*, *172*(1), 35-46.

Hurka, H., Krauss, R., Reiner, T., & Wöhrmann, K. (1976): Das Blühverhalten von *Capsella bursa-pastoris* (Brassicaceae). *Plant Systematics and Evolution*, *125*(2), 87-95.

Hurka, H., & Neuffer, B. (1991): Colonizing success in plants: genetic variation and phenotypic plasticity in life history traits in *Capsella bursa-pastoris*. In: Esser, G., Overdieck, D. (eds.) *Modern Ecology - Basic and Applied Aspects*, Amsterdam, London, New York, Tokyo, Elsevier-Vlg., 77-96.

Hurka, H., & Neuffer, B. (1997): Evolutionary processes in the genus *Capsella* (Brassicaceae)*. *Plant Systematics and Evolution*, *206*(1-4), 295–316.

Josse, J. & Husson, F. (2016): missMDA a package to handle missing values in principal component methods. *Journal of Statistical Software*, *70*(1).

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., ... & Eskin, E. (2010): Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics 42*(4), 348-54.

Kloot, P. M. (1983): Early records of alien plants naturalized in South Australia. *Journal of the Adelaide Botanic Gardens*, *6*, 93-131.

Kryvokhyzha, D., Holm, K., Chen, J., Cornille, A., Glémin, S., Wright, S. I., ... & Lascoux, M. (2016): The influence of population structure on gene expression and flowering time variation in the ubiquitous weed *Capsella bursa-pastoris* (Brassicaceae). *Molecular Ecology*, *25*(5), 1106-1121.

Kryvokhyzha, D., Salcedo, A., Eriksson, M. C., Duan,T., Tawari, N., Chen, J., ... & Stinchcombe, J. R. (2019): Parental legacy, demography, and admixture influenced the evolution of the two subgenomes of the tetraploid *Capsella bursa-pastoris* (Brassicaceae). *PLoS Genetics*, *15*(2), e1007949.

Lamping, H. (1985): *Australien*. Stuttgart: Klett.

Lawson, D. J., Van Dorp, L., & Falush, D. (2018): A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature communications*, *9*(1), 3258.

Lê, S., Josse, J. & Husson, F. (2008): FactoMineR: An *R* Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1). pp. 1-18.

Lempe, J., Balasubramanian, S., Sureshkumar, S., Singh, A., Schmid, M., & Weigel, D. (2005): Diversity of flowering responses in wild *Arabidopsis thaliana* strains. PLoS genetics, *1*(1), e6.

Lepais, O. & Weir, J. T. (2014): SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources*, *14*(6), 1314–1321.

Li, H., & Durbin, R. (2009): "Fast and accurate short read alignment with Burrows–Wheeler transform." *Bioinformatics*, *25*(14), 1754-1760.

Linde, M., Diel, S., & Neuffer, B. (2001): Flowering ecotypes of *Capsella bursa-pastoris* (L.) Medik. (Brassicaceae) analysed by a cosegregation of phenotypic characters (QTL) and molecular markers. *Annals of Botany*, *87*(1), 91-99.

Mack, R. N., & Lonsdale, W. M. (2001): Humans as Global Plant Dispersers: Getting More ThanWe Bargained For: Current introductions of species for aesthetic purposes present the largest single challenge for predicting which plant immigrants will become future pests. *AIBSBulletin*, *51*(2), 95-102.

Marais, W. (1970): Cruciferae. In Codd, L. E., De Winter, B., Killick, D. J. B., Rycroft, H. B., eds., *Flora of Southern Africa*. Vol. 13. Government Printer, Pretoria.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., Johnson, E. A. (2007): Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers". *Genome Research*, *17*(2), 240–248.

Mooney, H. A., Mack, R. N., McNeely, J. A., Neville, L. E., Schei, P. J., Waage, J. K. (2005): *Invasive alien species*. Washington, D.C.: Island Press.

Neuffer, B. (1996): RAPD analyses in colonial and ancestral populations of *Capsella bursa-pastoris* (L.) Med. (Brassicaceae). *Biochemical Systematics and Ecology*, *24*(5), 393-403.

Neuffer, B. (2011): Native range variation in *Capsella bursa-pastoris* (Brassicaceae) along a 2500 km latitudinal transect. *Flora*, *206*(2), 107-119.

Neuffer, B., Bartelheim, S. (1989): Gen-ecology of *Capsella bursa-pastoris* from an altitudinal transsect in the Alps. *Oecologia 81*(4), 521-527.

Neuffer, B., Bernhardt K.-G., Hurka, H., Kropf, M. (2011): Monitoring population and gene pool dynamics of the annual species *Capsella bursa-pastoris* (Brassicaceae) – initiation of a long-term genetic monitoring and a review of relevant species traits. *Biodiversity and Conservation*, *20*(2), 309-323.

Neuffer, B., Hirschle, S., Jäger, S. (1999): The colonizing history of *Capsella* in Patagonia (South America) – Molecular and adaptive significance. *Folia Geobotanica*, *34*(5), 435-450.

Neuffer, B., & Hurka, H. (1999): Colonization history and introduction dynamics of *Capsella bursa-pastoris* (Brassicaceae) in North America: isozymes and quantitative traits. *MolecularEcology*, *8*(10), 1667-1681.

Neuffer, B., Linde, M. (1999): *Capsella bursa-pastoris* – Colonization and adaptation; a globe-trotter conquers the world. In: van Raamsdonk, L.W.D., den Nijs, J.C.M. (eds.) *Plant Evolution in Man-Made Habitats*. Proceedings of the VIIth International IOPB Symposium 1998, 49-72.

Neuffer, B., Wesse, C., Voss, I., & Scheibe, R. (2018): The role of ecotypic variation in driving worldwide colonization by a cosmopolitan plant. *AoB Plants*, *10*(1), ply005.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000): Inference of population using multilocus genotype data. *Genetics 155*(2), 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007): PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559-575.

R Core Team (2018): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Randall, R. P. (2012): A global compendium of weeds. *Department of Agriculture and Food Western Australia* (Ed.2).

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., & Foulkes, A. S. (2015): A guide to genome wide association analysis and post analytic interrogation. Statistics in medicine, *34*(28), 3769-3792.

Scaglione, D., Acquadro, A., Portis, E., Tirone, M., Knapp, S. J., & Lanteri, S. (2012): RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *BMC Genomics*, *13*(1), 3.

Shindo, C., Lister, C., Crevillen, P., Nordborg, M., & Dean, C. (2006): Variation in the epigenetic silencing of FLC contributes to natural variation in *Arabidopsis* vernalization response. *Genes & Development*, *20*(22), 3079-3083.

Shull, H. (1909): *Bursa bursa-pastoris* and *Bursa heegeri*: biotypes and hybrids. *Carnegie Institution of Washington Publication* 112, 3-56.

Shull, H. (1911): Defective inheritance-ratios in *Bursa* hybrids. Verhandlungen des natur-historische Vereins in Brünn 49, 156–168.

Sicard, A., Thamm, A., Marona, C., Lee, Y. W., Wahl, V., Stinchcombe, J. R., ... & Lenhard, M. (2014): Repeated evolutionary changes of leaf morphology caused by mutations to a homeobox gene. *Current Biology*, *24*(16), 1880-1886.

Slotte, T. (2007): Evolution of flowering time in the tetraploid *Capsella bursa-pastoris* (Brassicaceae) (Doctoral dissertation, Acta Universitatis Upsaliensis).

Šmarda, P., & Bureš, P. (2010): Understanding intraspecific variation in genome size in plants. *Preslia*, *82*(1), 41-61.

Williams, J.L., Auge, H., & Maron, J. L. (2008): Different gardens, different results: native and introduced populations exhibit contrasting phenotypes across common gardens. *Oecologia*, *157*(2), 239-248.

Zhou, T. Y., Lu, L. L., Yang, G., Al-Shehbaz, I. A. (2001): Brassicaceae (Cruciferae). In: Wu ZY, Raven PH, eds. *Flora of China 8*. Beijing, St Louis: Science Press/Missouri Botanical Garden Press, 1–193.

List of Supplementary Contents

List of Supplementary Figures

List of Supplementary Tables

5. Supplement

5.1 Common garden experiment

5.1.1 Materials and Methods

The seeds come from parental plants from populations from a variety of locations from all over the world, but mainly from North and South America (“New World“ experiment) and Eurasia (“Old World“ experiment) respectively. Some samples were added from other regions from the world (e.g. Africa and Australia) to either one of the two common garden experiments. All seed vouchers are stored at the Botanical Garden of the Osnabrück University.

Sowing and planting as well as estimation of the genome size via flow cytometry are described in *III. Taking the long way around – Worldwide geographical structure of the cosmopolitan weed Capsella bursa-pastoris (Brassicaceae)*. A thermohygrograph recorded temperature and air moisture during both common garden experiments (Fig. S.1).

The following parameters were recorded during the common garden experiments:

- Seed weight of the collected seeds before sowing. The weight was determined on 50 seeds.
- Germination rate. Percentage of germinated seeds per 50 sown seeds per plant family.
- Onset of flowering. The exact day after sowing when the first white petals showed at the bud (buds were observed daily).
- Leaf shape. Determined on fully differentiated adult leaves according to the classification of Shull (1909), primarily discriminating the dissection of leaves, which ranges from entire leaves to very deeply dissected ones: *simplex*, *tenuis*, *rhomboidea* and *heteris*.
- Plant height. Measured in cm from the ground to the hightest tip of the plant.
- Number of basal branches. The amount of branches emerging from the rosette of the plant.

The descriptive statistics have been calculated using R (R Core Team 2018). The Shapiro-Wilk normality test to test for gaussian distribution of the data was performed using the R function shapiro.test(). To test for association between paired samples, the Pearson's product moment correlation coefficient was calculated using the R function cor.test(). To test if observations from two groups are independent of each other or equal, the non-parametric Wilcoxon rank sum test is used if the assumption of normal distribution of the data is not fulfilled. The Wilcoxon rank sum test was performed using the R function Wilcox.test().

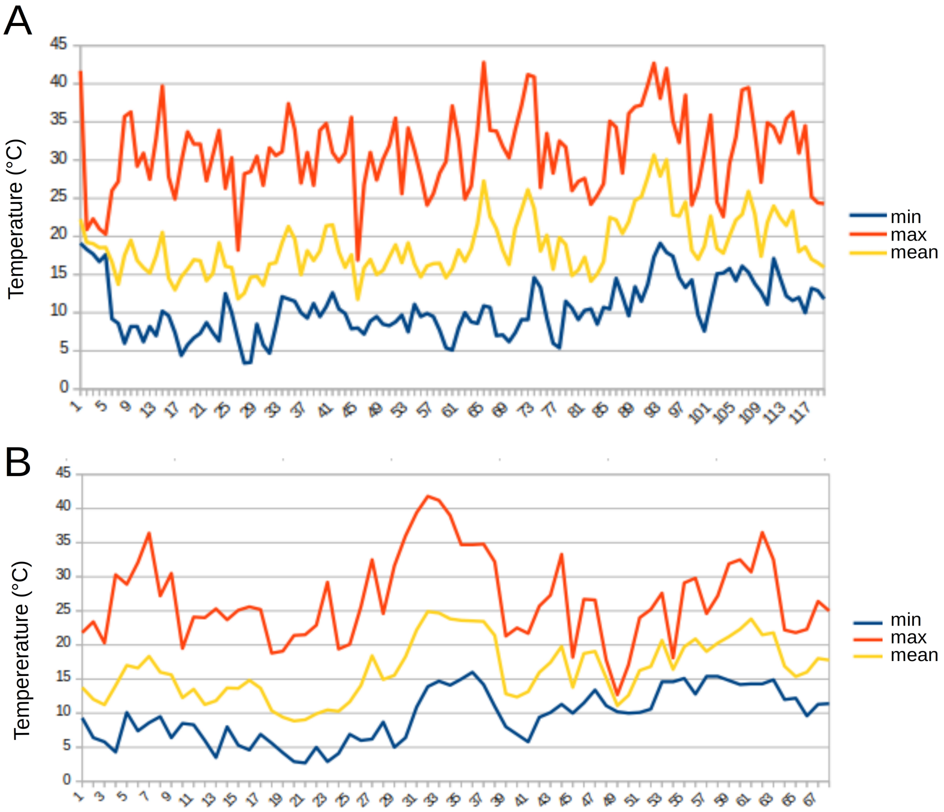

Figure S.1: Temperatures recorded during the two common garden experiments.
A: experiment in 2015 („New World“).
B: experiment in 2016 („Old World“).
Numbers on x-axis are days after sowing.

For further analysis, Multiple factor analysis (MFA) was performed with the R package FactoMineR (Lê et al., 2008). The numerical parameters (seed weight, germination percentage, onset of flowering, plant height, genome size and number of branches) were analyzed with Principal Component Analysis (PCA) using the PCA() function in FactoMineR. Since PCA is not suitable for categorical data, additionally a Factor analysis of mixed data (FAMD) was performed using the FAMD() function in FactoMineR to add information from the scored leaf types to the dataset. FAMD, like PCA, is a multiple factor analysis (MFA) to analyze a group of individuals described by variables. For both, PCA and FAMD, the dataset was imputed with the missMDA() function in FactoMineR to replace missing data with suitable values (Josse & Husson 2016).

5.1.2 Results

5.1.2.1 Seed weight

The seed weight (*n* = 4199; plant family mean) was measured in mg per 50 seeds and varied from 0.08 mg to 13.20 mg with a median of 4.90 mg, a mean of 4.907 mg and a standard deviation of 1.60 mg. A histogram is shown in Fig. S.2a. There is no normal distribution in the data according to the Shapiro-Wilk normality test (Fig. S.2b; *W* = 0.97232, *p* < 0.05).

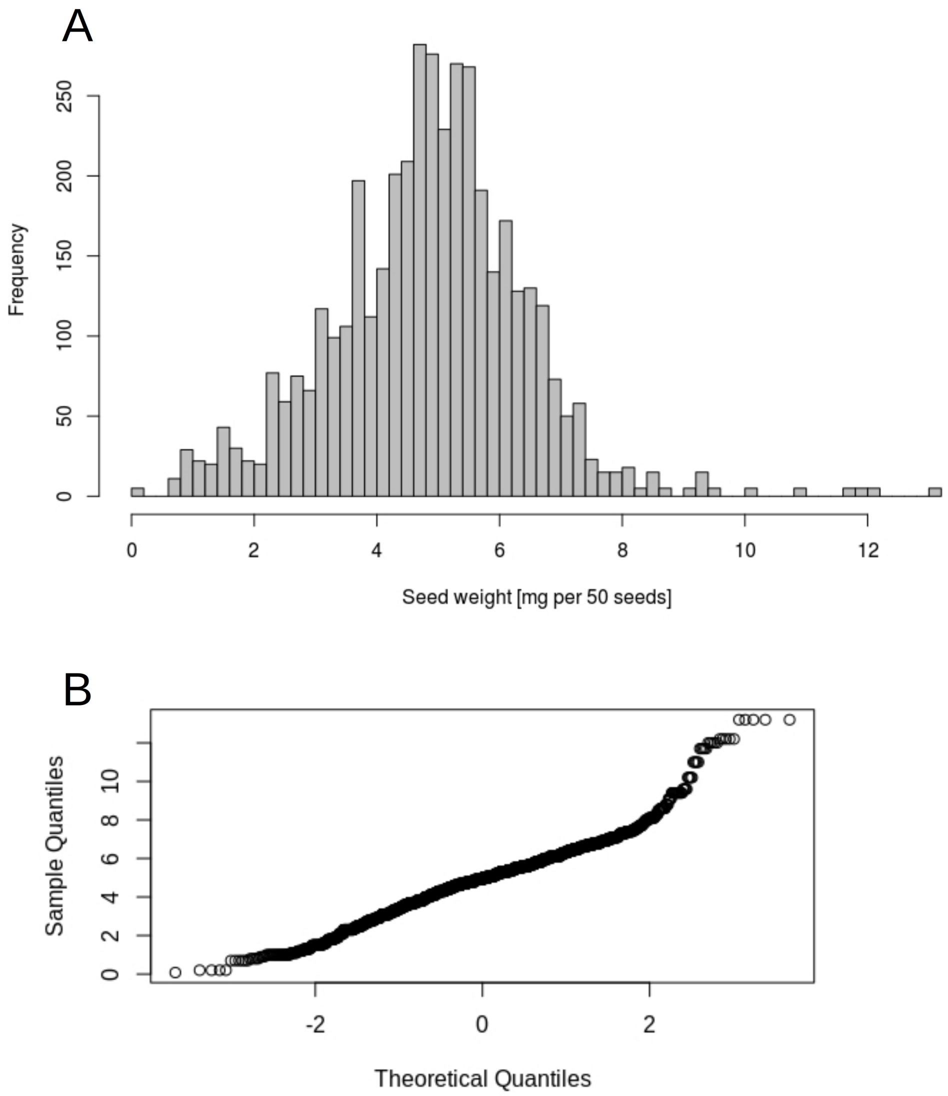

Figure S.2: Seed weight.
The seed wieght was maesured in mg per 50 seeds.
A: Histogram of the seed weight. B: Normal qq-plot.

5.1.2.2 Germination

The germination percentage (n = 4199; plant family mean) varied from 0 to 100 % with a standard deviation of 26.87 %, a median of 44.00 % and a mean of 44.99 %. A histogram is shown in Fig. S.3a. The data is not normally distributed by eye (Fig. S.3b) and by the Shapiro-Wilk normality test (W = 0.96742, $p < 0.05$).

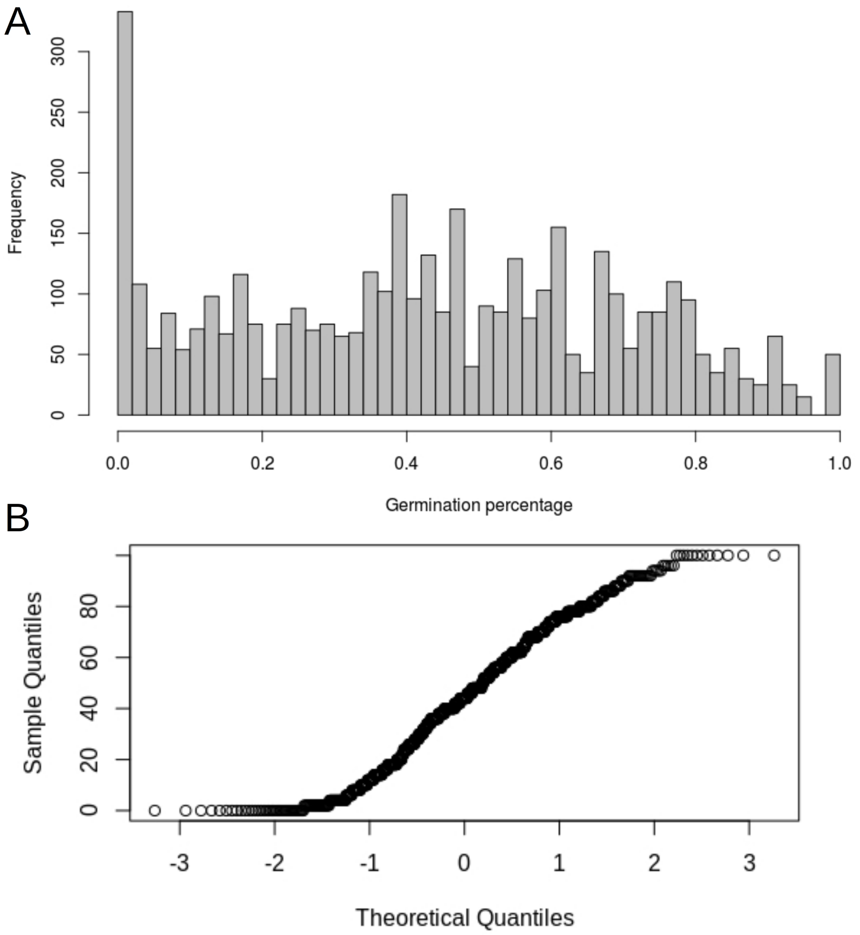

Figure S.3: Germination.
A: Histogram of the germination. B: Normal qq-plot.

5.1.2.3 Flowering

The onset of flowering (*n* = 3,888) varied from 41 days to 101 days with a standard deviation of 8.92 days, a median of 57.00 days and a mean of 58.09 days. A histogram is shown in Fig. S.4a. The flowering time is not normally distributed by eye (Fig. S.4b) and Shapiro-Wilk normality test (*W* = 0.95984, $p < 0.05$).

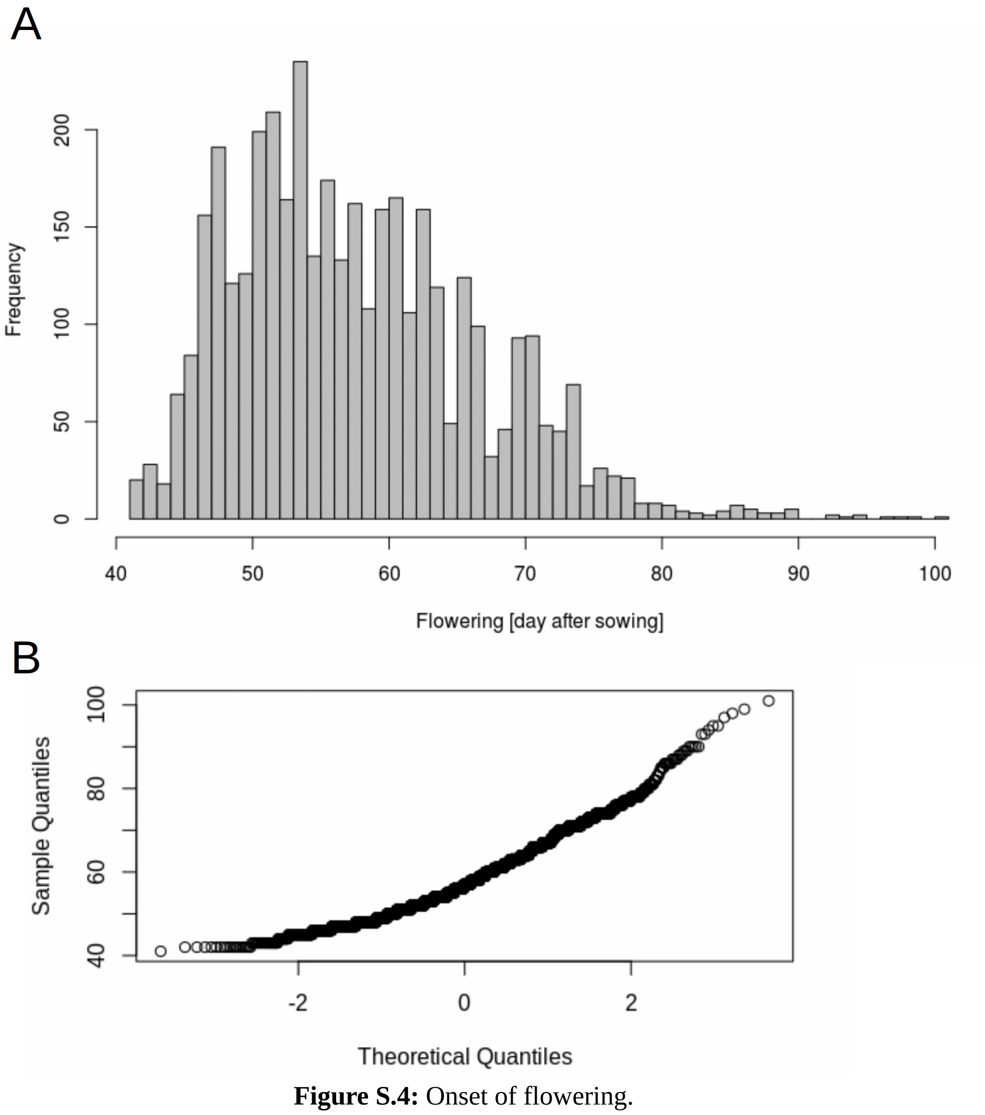

Figure S.4: Onset of flowering.
A: Histogram of the flowering. B: Normal qq-plot.

5.1.2.4 Leaf shape

Leave shapes from 2,523 individuals were recorded. 761 individuals had the *rhomboidea* leaf type (≙ 30.2 %), 509 were *heteris* (≙ 20.2 %), 317 were *tenuis* (≙ 12.6 %) and 797 were *simplex* (≙ 31.6 %). The leaf shapes of 139 individuals (≙ 5.5 %) were not detectable (Tab. S.1).

The leaf shapes were not evenly distributed over all sampled regions (Fig. S.5): In the native distribution of *C. bursa-pastoris*, Eurasia, the most common leaf type is *heteris* (≙ 36.6 %), whereas it is *simplex* in Africa (≙ 69.1 %) and North America (≙ 40.8 %) and *rhomboidea* in South America (≙ 45.9 %). In Eurasia, *rhomboidea* and *heteris* are almost evenly common (≙ 31.7 % and 36.6 % respectively). The *tenuis* type is rather rare, being the rarest in South America and the second rarest in North America and Eurasia. The *tenuis* and *heteris* type were not scored in Africa, but the sampling size is very small. The sampling size is too small in Australia to make any statements of this region.

As already stated in Neuffer et al., 2018 (*I. The role of ecotypic variation in driving worldwide colonization by a cosmopolitan plant*), the degree of leaf-margin dissection is likely to be functionally important and an adaptive trait. According to the data presented here, the leaf types seem to be adapted to the altitude of the source area of the plants (Fig. S.6; Kruskal-Wallis chi-squared = 61.623, df = 4, *p*-value = 1.322e-12).

A significant correlation between the leaf type after Shull and the leaf thickness could not be detected (Kruskal-Wallis chi-squared = 10.743, df = 4, *p*-value = 0.02961).

Table S.1: Leaf type occurences after Shull (1909).

	rhomboidea	***heteris***	***tenuis***	***simplex***	***unscorable***	***sum***
Eurasia	290	335	135	130	25	915
Africa	50	0	0	134	10	194
SouthAmerica	200	55	25	131	25	436
North America	221	119	157	397	79	973
Australia	0	0	0	5	0	5
sum	**761**	**509**	**317**	**797**	**139**	**2523**

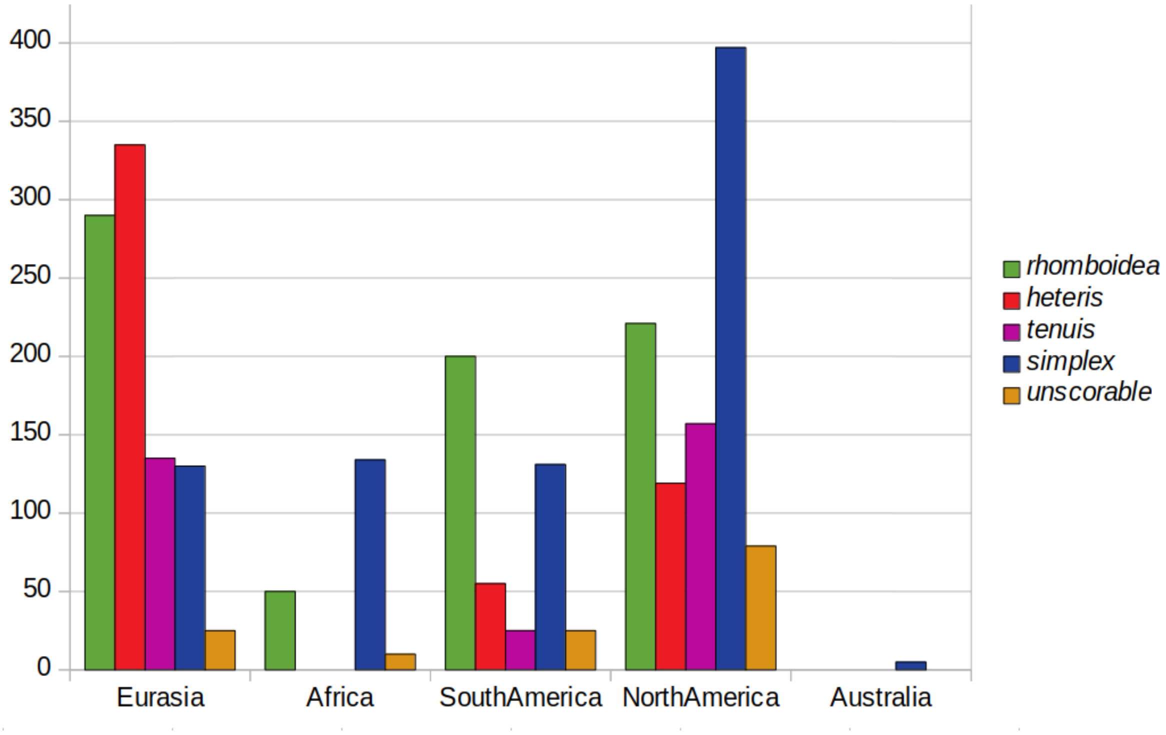

Figure S.5: Distribution of scored leaf types after Shull (1909) in various regions.

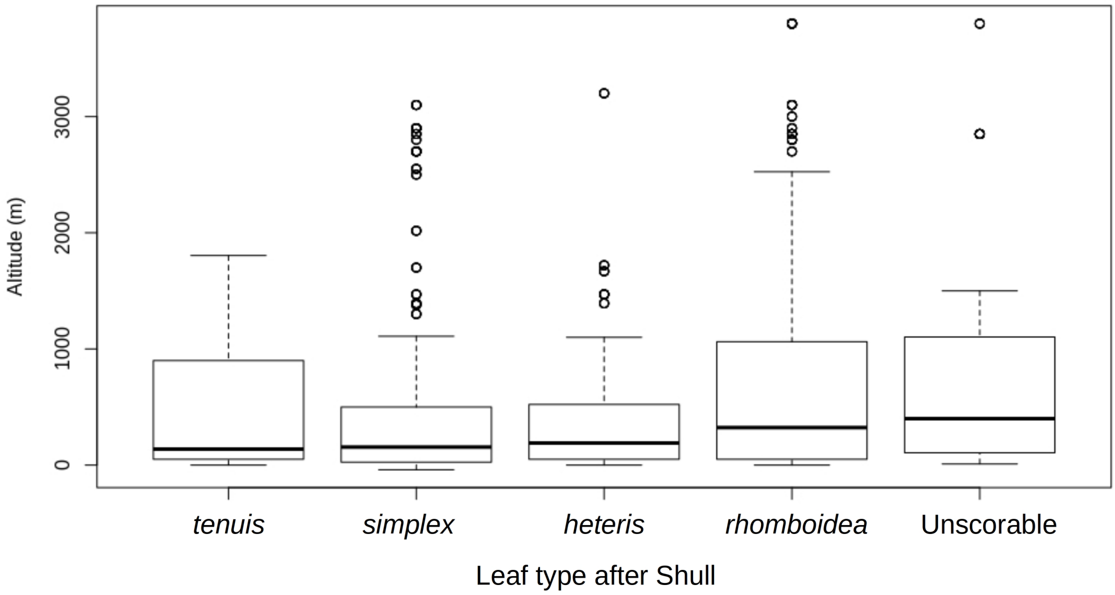

Figure S.6: Leaf types after Shull arranged according to their altitude occurences. At 0.05 significance level, the altitude of the source habitat and the leaf types after Shull are nonidentical populations.

5.1.2.5 Plant height

The measured plant height (n = 1,918) varied from 4 cm to 102 cm with a standard deviation of 13.1 cm, a median of 57 cm and a mean of 60.7 cm. A histogram is shown in Fig. 7a.

The data is not normally distributed according to the Shapiro-Wilk normality test (Fig. S.7b; W = 0.95984, p < 0.05).

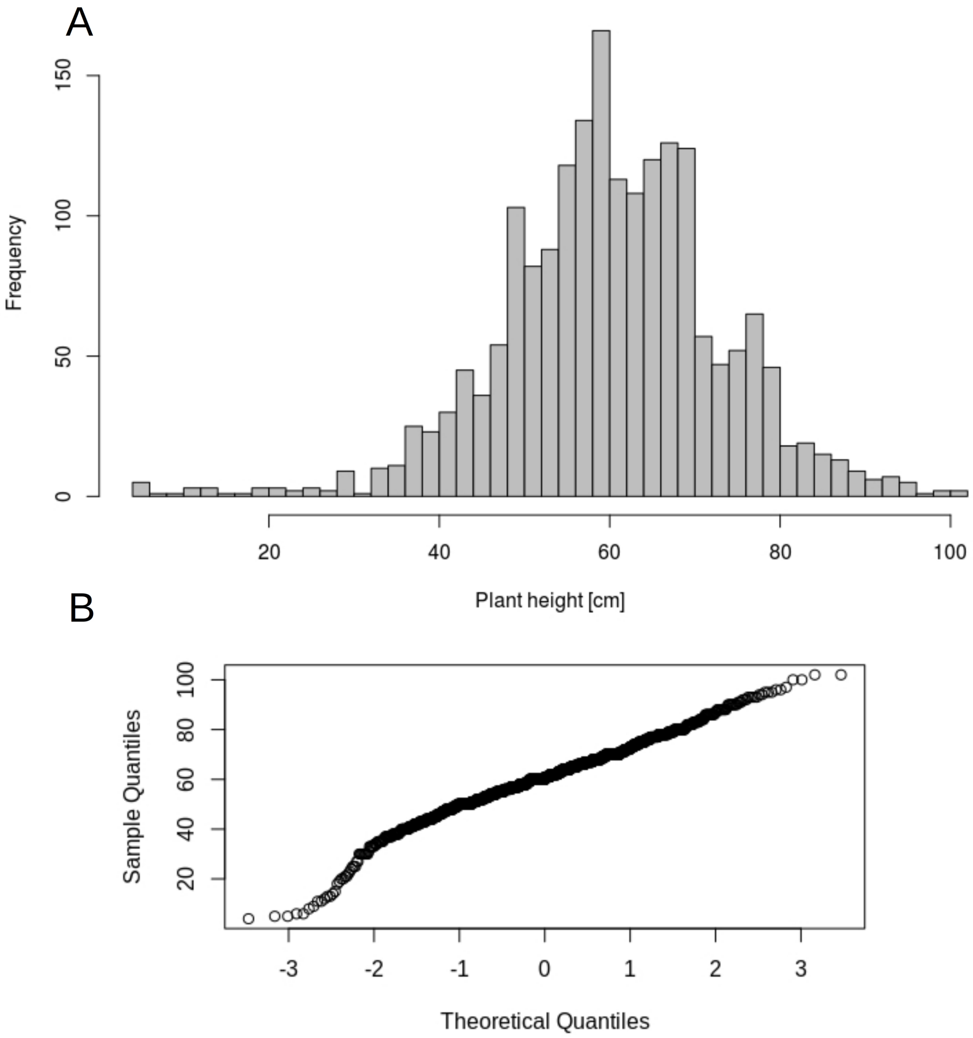

Figure S.7: Plant height.
A: Histogram of the plant height. B: Normal qq-plot.

5.1.2.6 Number of branches

The number of basal branches (n = 1,918) varied from 1 to 19 with a standard deviation of 2.35, a median of 7.00 and a mean of 7.55. A histogram is shown in Fig. S.8a. The data is not normally distributed according to the Shapiro-Wilk normality test (Fig. S.8b; W = 0.9576, $p < 0.05$).

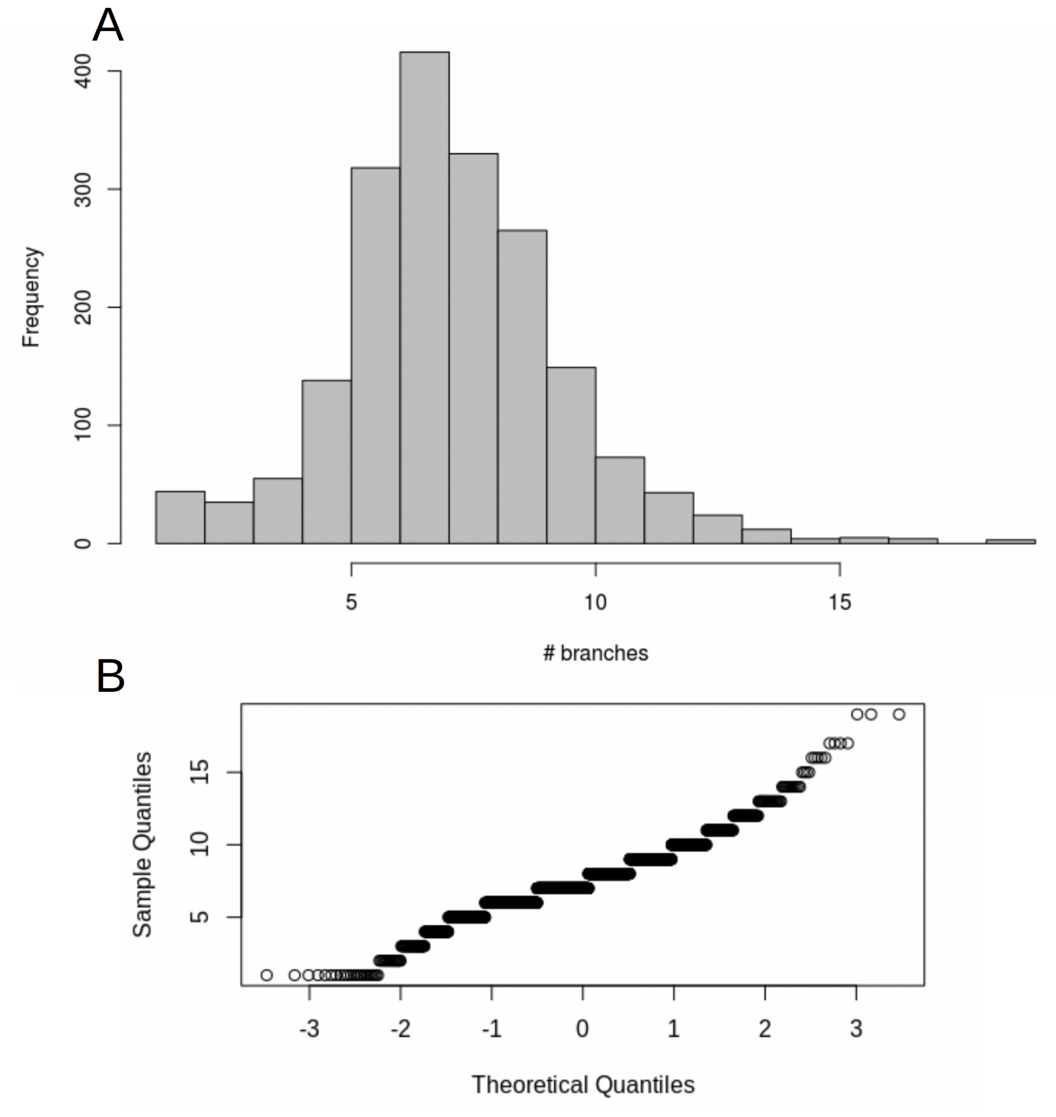

Figure S.8: Number of branches.
A: Histogram of the branch number. B: Normal qq-plot.

5.1.2.7 Genome size

Since many populations occur sympatrically with other species and flowering periods can overlap, interspecific crosses between the *Capsella* species are possible (e.g. Almquist, 1929; Hurka & Neuffer, 1997). At least two hybrid individuals could be detected during this study with the help of FCM: With a genome size of 0.73 pg and 0.65 respectively, two individuals had measurements laying in-between a typical tetraploid or diploid *Capsella* sample (e.g. Hurka et al., 2012). Other samples were identified as obviously diploid individuals according to the FCM and therefore removed from the dataset together with the triploid ones.

The genome size of *C. bursa-pastoris* estimated via flow cytometry (*n* = 3,870) varied from 0.8465 pg to 1.0008 pg with a standard deviation of 0.0315 pg, a median of 0.9134 pg and a mean of 0.9136 pg. A histogram is shown in Fig. S.9a. The data is not normally distributed by eye (Fig. S.9b) and Shapiro-Wilk test normality test (*W* = 0.97767, *p* < 0.05).

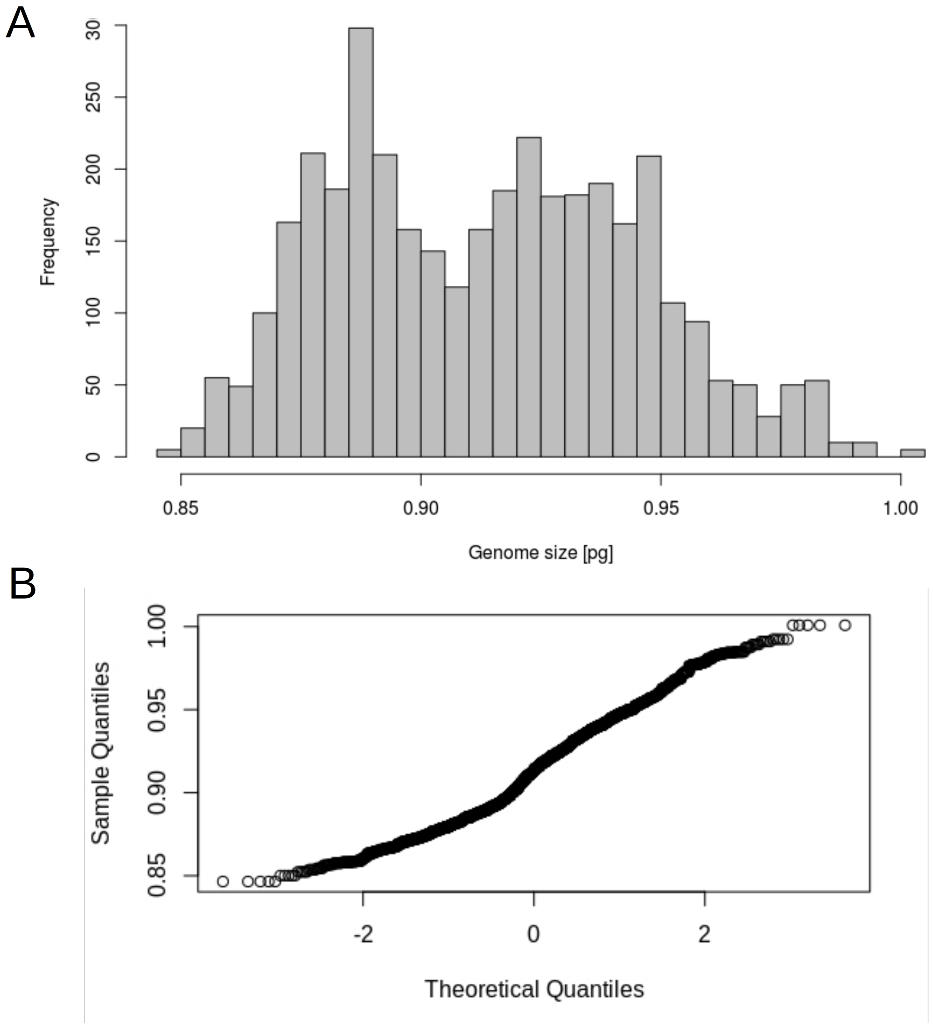

Figure S.9: Genome size.
A: Histogram of the genome size. B: Normal qq-plot.

5.1.2.8 Correlations between phenotypic parameters and factor analysis

The correlations between the phenotypic parameters were calculated with the Pearson correlation test in R (Tab. S.2): The seed weight correlated highly significantly positively with the germination percentage (*cor* = 0.224) and slightly but significantly with the genome size (*cor* = 0.051). The germination correlated highly significantly negatively with the onset of flowering (*cor* = -0.143) and slightly but significantly with the plant height (*cor* = 0.069), number of branches (*cor* = 0.077) and genome size (*cor* = -0125). The onset of flowering correlated highly significantly with the number of branches (*cor* = -0.188) and the genome size (*cor* = 0.526) and slightly but significantly with the final plant height (*cor* = 0.048). The plant height also correlated highy significantly with the number of branches (*cor* = 0.231) and significantly with the genome size (*cor* = 0.104). The number of branches also correlated significantly with the genome size (*cor* = -0.143).

Table S.2: Correlation matrix.
SW: seed weight. GER: germination percentage. FLW: flowering day. HGT: plant height. BR: number of branches. GS: genome size. **: *p* < 5.408e-7. *: 5.408e-7 < *p* < 0.05. n.s.: *p* > 0.05.

Pearson‘s	*cor*						
p		SW	GER	FLW	HGT	BR	GS
	SW	/	0.224	n.s.	n.s.	n.s.	0.051
	GER	**	/	-0.143	0.069	0.077	-0.125
	FLW	n.s.	**	/	0.048	-0.188	0.526
	HGT	n.s.	*	*	/	0.231	0.104
	BR	n.s	*	**	**	/	-0.143
	GS	*	*	**	*	*	/

Principal component analyis (PCA) was used to reduce dimensions and illustrate the parameters seed weight, germination, flowering, height, number branches and genome size (Fig. S.10a). Since the leaf type is not a numerical but a categorical variable, additionally a factor analysis of mixed data (FAMD) was performed to include information from this variable (Fig. S.10b). In the PCA plot, clustering of individuals is not obvious at first glance (Fig. S.10a), whereas samples are evidently structured as soon as information from the leaf type is added in a FAMD plot (Fig. S.10b): Plants with *simplex* (blue), *rhomboidea* (green) and *heteris* (red) shaped leaves show little overlap. However, the *tenuis* type is more widely distributed and does not seem to contribute to population structure. Yet, when studying population structure, the main focus should be put on the results of *II. Geographical structure of genetic diversity in Shepherd's Purse, Capsella bursa-pastoris – a global perspective* and *III. Taking the long way around – Worldwide geographical structure of the cosmopolitan weed Capsella bursa-pastoris (Brassicaceae)*, where genetic markers have been included. These plots here only serve to see whether the phenotypic parameters alone are sufficient to recognize a certain structuring, which is only very conditionally the case.

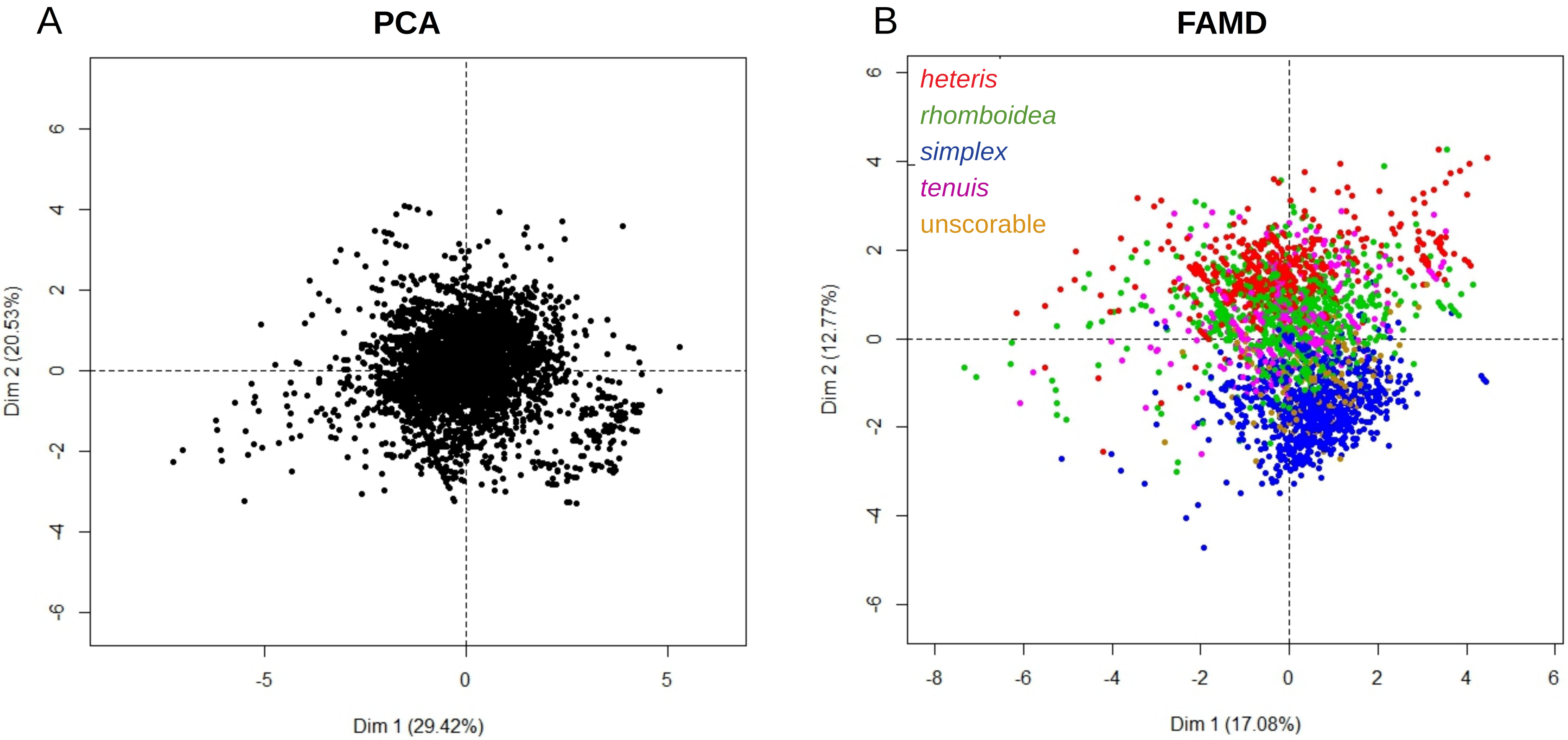

Figure S.10: Multiple factor analysis of the imputed phenotypic dataset.
A: PCA of numerical data. B: FAMD of numerical and categorical data. Colors refer to leaf types.

5.2 Genome sequencing and analysis

A few samples, that were not used in the common garden experiment but had tissue left from former studies, were also used for sequencing and added to the dataset.

DNA extraction, library preparation and sequencing were performed in the Max Planck Institute for Developmental Biology in the Department of Molecular Biology of Detlef Weigel and in the Genome Center in Tübingen, Germany. DNA extracts and library backup are stored in the Weigel Lab. The raw sequence data have also been deposited on a local harddrive in the lab of Barbara Neuffer. The sequence data was further processed and the resulting data used for population structure analyses.

5.2.1 DNA extraction

Extraction of genomic DNA from *C. bursa-pastoris* rosette leaves was performed with the CTAB method: The tissue was disrupted to a fine powder by shaking with a bead homogenizer for two minutes and then mixed with 300 µl CTAB DNA extraction buffer (2 % CTAB, 1.4 M NaCl, 100 mM Tris-HCl pH 8.0, 20 mM EDTA pH 8.0). After incubation in a water bath at 65 °C for 30 min, 300 µl chloroform was added. After centrifugation for 15 min at 3.000 rcf the DNA was precipitated with 200 µl isopropanol, washed in 80 % ethanol, dried and resuspended in *Buffer EB* (Qiagen, Germany). DNA quality and concentration were determined on a 1 % agarose gel and with the Tecan Fluorescence Microplate Reader (Teacan Group AG, Swiss).

CTAB extraction buffer

Reagent	**C**
CTAB (Cetrimonium bromide)	2 %
NaCl	1.4 M
Tris-HCl, 1M pH 8	100 mM
EDTA, 0.5 M pH 8	20 mM

The more detailed protocol for DNA extraction can be found in table S.3.

Table S.3: DNA extraction protocol

Step	**Instruction**
1.	Disrupt tissue by shaking with bead homogenizer for 2 min
2.	Centrifuge briefly, then add 300 µl CTAB and vortex vigorously to mix tissue with buffer
3.	Incubation at 65 °C in water bath for 30 min
4.	Add 300 µl chloroformand and vortex vigorously
5.	Centrifuge at 3.000 rcf for 15 min
6.	Transfer 200 µl of the chloroform-extracted supernatant to 200 µl isopropanol
7.	Centrifuge at 3.000 rcf for 15 min
8.	Pour off the liquid; the DNA pellet should stay behind
9.	Wash the pellet with 200 µl of 70 % ethanol
10.	Centrifuge at 3.000 rcf for 10 min
11.	Pour off the liquid; the DNA pellet should stay behind. Dry the pellet
12.	Resuspend the pellets in 100 µl (+1 µl RNAse) *Buffer EB* (Qiagen, Germany)

5.2.2 Library preparation

	Step	**Instruction**
1.	*KpnI* restriction	• Add 3 µl *10X FastDigest Buffer* and 1 µl *FastDigest KpnI* to 26 µl DNA (200 ng) • Incubate in a thermal cycler at 37 °C for 30 min
2.	*AMPure XP* clean-up	• Add 54 µl of *AMPure® XP SPRI® beads* (1.8:1 ratio) • Incubate at room temperature for 10 min • Keep sample on magnet, after 5 min remove 75 µl of the supernatant • Add 200 µl of 80 % ethanol (freshly prepared) to the sample still on the magnet. Wait for 30 s, then remove all the supernatant. Repeat ethanol wash • Dry samples on magnet for 10 – 20 min until rest ethanol has evaporated. Once pellet is dry, remove from the magnet • Elute sample in 11.5 µl *Buffer EB*. Incubate at room temperature for 2 min • Return sample to magnet for 5 min. Pipet 10 µl of the eluate into a new plate
3.	*KpnI* adapter ligation using T4 Ligase	• Add 3 µl *10X ligase buffer*, 3 µl *PEG 4000*, 1 µl 0.05 µl mixed *KpnI* adapter, 1 µl *T4 DNA ligase* (5U/µl) and 12 µl water to 10 µl digested DNA • Incubate sample at room temperature for 30 min, then place sample on ice for 15 min
4.	*AMPure XP* clean-up (to remove fragments < 300 bp)	• Add 70 µl of *Buffer EB* to the 30 µl ligation reaction and then *AMPure® XP SPRI® beads* (1.8:1 ratio) • Incubate at room temperature for 10 min • Keep sample on magnet, after 5 min remove 170 µl of the supernatant • wash with ethanol and dry sample as described in 2. • Elute sample in 31.5 µl *Buffer EB*. Incubate at room temperature for 2 min • Return sample to magnet for 5 min. Pipet 30 µl of the eluate into a new plate
5.	Multiplex and *AMPure XP* clean-up (to concentrate multiplexed samples)	• Combine up to 96 barcoded samples by pipetting 10 µl of each sample into a 2 ml tube. The directions below are for 4 pools of 96 samples each • Add 960 µl of *AMPure® XP SPRI® beads* (1:1 ratio) to each of your pools • Incubate samples at room temperature for 10 min • Keep sample on magnet, after 5 min remove 1800 µl of the supernatant • Add 1000 µl of 80 % ethanol (freshly prepared) to the sample still on the magnet. Wait for 30 s, then remove all the supernatant. Repeat ethanol wash • Dry samples on magnet for 20 – 30 min until rest ethanol has

		evaporated. Once pellet is dry, remove from the magnet • Elute samples serially. All 4 pools must be combined in a total volume of 55 µl of *Buffer EB* at the final elution: ○ Elute pool 1 in 62 µl Buffer EB. Incubate for 2 min at room temperature ○ return sample to magnet for 5 min ○ pipet 60 µl of the eluate into pool 2 ○ incubate sample for 2 min at room temperature ○ return sample to magnet for 5 min ○ pipet 58 µl of the eluate into pool 3 ○ incubate sample for 2 min at room temperature ○ return sample to magnet for 5 min ○ pipet 56 µl of the eluate into pool 4 ○ incubate sample for 2 min at room temperature ○ return sample to magnet for 5 min ○ pipet 55 µl of the eluate (containing 384 barcoded samples) into a new *microTUBE* (Covaris, United States)
6.1	Sample fragmentation via *Covaris* shearing	• Shear the samples with focused-ultrasonicator. To target 500 bp fragments, use the following settings: Duty cycle 10 %, Intensity 5, Cycles per Burst 200, Time 40 s. Transfer sheared sample to new tube/plate.
7.	*AMPure XP* clean-up (to remove fragments < 300 bp)	• Add 45 µl of *Buffer EB* to 55 µl sheared sample and then 80 µl of *AMPure® XP SPRI® beads* (0.8:1 ratio) • Incubate samples, wash and dry as described in 4. • Elute sample in 21.5 µl *Buffer EB*. Incubate at room temperature for 2 min • Return sample to magnet for 5 min. Pipet 20 µl of eluate into new tube/plate
8.	End Repair (NEW Next DNA Sample Prep MMS1)	• Add 10 µl 1*0X NEBNext® End Repair Reaction Buffer*, 5 µl *NEBNext End Repair Enzyme Mix* and 65 µl water to 20 µl fragmented DNA • Incubate in thermal cycler for 30 min at 20 °C
9.	*AMPure XP* clean-up (to remove fragments < 300 bp)	• Add 80 µl of *AMPure® XP SPRI® beads* to your 100 µl sample (0.8:1 ratio) • Incubate samples, wash and dry as described in 4. • Elute sample in 31.5 µl *Buffer EB*. Incubate at room temperature for 2 min • Return sample to magnet for 5 min. Pipet 30 µl of eluate into new tube/plate
10.	DA-Tailing (NEBNext DNA Sample Prep MMS1)	• Add 5 µl 10X NEBNext dA-Tailing Reaction Buffer, 3 µl Klenow Fragment (3'→5' exo-) and 12 µl water to 30 µl end repaired DNA • Incubate the sample in a thermal cycler for 30 min at 37 °C
11.	*AMPure XP* clean-up (to remove	• Add 50 µl of *Buffer EB* to 50 µl A-tailed sample and then 80 µl of *AMPure® XP SPRI® beads* (0.8:1 ratio)

	fragments < 300 bp)	• Incubate samples, wash, dry and elute as described in 9.
12.	Universal Adapter Ligation	• Add 10 µl *5X NEBNext Quick Ligation Reaction Buffer* to 5 µl *Quick T4 DNA ligase*, 1 µl 10 µM Universal Adapter (G-34024 and G-34025) and 4 µl water to 30 µl dA-tailed DNA • Incubate the sample in a thermal cycler for 15 min at 20 °C
13.	*AMPure XP* clean-up (to remove fragments < 300 bp)	• Add 50 µl of *Buffer EB* to 50 µl A-tailed sample and then 80 µl of *AMPure® XP SPRI® beads* (0.8:1 ratio) • Incubate samples, wash and dry as described in 4. • Elute sample in 26.5 µl *Buffer EB*. Incubate at room temperature for 2 min • Return sample to magnet for 5 min. Pipet 25 µl of eluate into new tube/plate
14.	PCR enrichment	• Add 15 µl 2X Phusion HF Master Mix, 1 µl 10 µm Primer G-26878 and G-33106, 1 µl 10 µM Primer G-34025 and 8 µl water to 5µl ligated DNA (PCR primers have to be HPLC purified) • Amplify using the following protocol: Initial denature 98 °C 0:30 14 cycles of 98 °C 0:10 65 °C 0:30 72 °C 0:30 Final extension 72 °C 0:30 Hold 10 °C ∞
15.	*AMPure XP* clean-up (to remove fragments < 300 bp)	• Add 70 µl of *Buffer EB* to 30 µl PCR product and then 80 µl of *AMPure® XP SPRI® beads* (0.8:1 ratio) • Incubate samples, wash, dry and elute as described in 13.
16.1	Library validation	• Measure final sample with the *Qubit Fluorometer* (Thermo Fisher Scientific, United States) and validate concentration and size distribution on the *Agilent Bioanalyzer* with a *DNA1000 chip* (Agilent Technologies, United States). • Also quantify sample on fluorometer using BR assay • For sequencing, dilute final sample to 10 nM with 0.1 % Tween-EB and validate 10nM dilution on the Qubit using BR assay. Use conentration (ng/µl) and mean fragment size obtained from Agilent Bioanalyzer to calculate nanomolarity (nM).

Libraries were sequenced on Illumina HiSeq Analyzer in a total throughput of six lanes.

5.2.3 Barcode sequences

(1/2)

Nr.	Barcode	Nr.	Barcode	Nr.	Barcode	Nr.	Barcode	Nr.	Barcode	Nr.	Barcode	Nr.	Barcode	Nr.	Barcode
1	TGTTT	35	GGGTG	69	GGAGC	103	ATCCCT	137	CCTGTT	171	AATGGC	205	CATGTAT	239	AGCGAAG
2	TTAGT	36	CAACG	70	CAGAC	104	AGAGCG	138	ATACCG	172	GAAAGC	206	CCTATTA	240	CAAGTGA
3	GGAAT	37	TTTGC	71	CTGGA	105	GCAGTA	139	ACGGAC	173	GCTTAT	207	AGTACGC	241	CTCAAGC
4	GATCG	38	ACTAC	72	AAACA	106	CGCGAT	140	TTCACT	174	CCCAAT	208	TTACAAG	242	CTACTTT
5	AAGTC	39	AGCTA	73	CCGGT	107	GCGCAA	141	CTAGTA	175	TCGTAA	209	AAGGCCG	243	GTATACC
6	CTGCC	40	ATCCA	74	GCTAT	108	TTAGGG	142	CTGCTG	176	AATCTC	210	GATTGTC	244	CGACTGG
7	TGAAC	41	AACTT	75	CTCGG	109	GCATCC	143	CTTTAC	177	GTTCAT	211	CACGGAA	245	GTTCTTA
8	TCCGA	42	GTCCT	76	ACCAG	110	CTACAA	144	GAATCT	178	GCAAGA	212	TTAATGG	246	TAGCGAG
9	GATTT	43	AGATG	77	CCAGC	111	GGAAAG	145	CTGAGA	179	TTTCTC	213	CTCGGTC	247	TATGCTT
10	CGAGT	44	CTTAG	78	CTCAC	112	TGGCGT	146	GTTACT	180	TTCTGG	214	ACAATGT	248	AGAGTCA
11	TCAAT	45	ATGGC	79	AGGGA	113	TAGACT	147	ATAGTG	181	AGCACG	215	CTTTCTT	249	CATTCCA
12	GTGCG	46	TATAC	80	AGTAA	114	TCTGAG	148	TCACAC	182	AGAGAA	216	CGTTGTA	250	CTAACCC
13	TGCTC	47	CTATA	81	TGCGT	115	AGGTTA	149	AACGAA	183	CTTACG	217	AGGTGGT	251	CAGAAGG
14	TTCCC	48	GACCA	82	GTGAT	116	GATGGT	150	TACCTT	184	GAAGTC	218	GTGGCTT	252	AGCAGCG
15	GAAAC	49	GCATT	83	ATTCG	117	GGGTCA	151	CAAAGA	185	TCGCCT	219	GCATGCT	253	GAGCACG
16	CTTCA	50	TGACT	84	CCAAG	118	AACTGC	152	AAGGAT	186	GTGTCG	220	TCGCTTC	254	ACCTTTA
17	ATGTT	51	GAATG	85	GTTCC	119	ATTCAG	153	TCAGCC	187	ATCGGG	221	TACTGTA	255	GCAATCG
18	ACAGT	52	TCGAG	86	TCCAC	120	CATCTT	154	TCTCTA	188	CGAAGC	222	AACCCGC	256	GGTCATC
19	TTTTG	53	TAGGC	87	GTCGA	121	AAGAGT	155	CTCTAT	189	ACTCAT	223	TCATTCC	257	AAACAGG
20	ACGCG	54	TTGAC	88	CCTAA	122	GGACGT	156	TAAGAC	190	ACCATC	224	CTTCGGG	258	CTTAGAC
21	GTATC	55	GGATA	89	CACGT	123	CGGATA	157	GCACTC	191	TGTGAT	225	TCTGCCC	259	ACGTGAA
22	ACACC	56	CGACA	90	TACAT	124	AGAGTT	158	GTGCGA	192	CTAGCC	226	GCACTAT	260	GGGTTAT
23	TCTTA	57	AGTGT	91	CCTCG	125	ACCAGG	159	TTCCAA	193	TGGTACT	227	AAGGGTA	261	GAGTCGA
24	GGTCA	58	CCACT	92	GCTTC	126	CTGTCA	160	AGTTCT	194	GAAGGTT	228	ATGGTGC	262	TGCCATA
25	CTCTT	59	TGGGG	93	CATCC	127	TGTTTC	161	TGCCTC	195	CCACGAG	229	GGTTAGT	263	AGTGACT
26	TCGCT	60	GAGAG	94	ATAAC	128	ACCCAC	162	TACGTG	196	TCAGTAA	230	TTGCTCA	264	CGATCCT
27	AATTG	61	AACGC	95	CGCGA	129	ACCCGA	163	ACTACC	197	CACCTTC	231	GTGCAAC	265	TCCACAT
28	AACCG	62	AGGAC	96	GGGAA	130	TTCGCG	164	TCTACG	198	TACGTAC	232	GCAGATA	266	GTCTGAG
29	TCATC	63	AATGA	97	ATGACC	131	GATAGG	165	AGTTGC	199	TGGCGGA	233	AGGATTG	267	GGAGTTC
30	GGTAC	64	GCACA	98	AGACGA	132	AGTCTT	166	AGTAGG	200	CCCATAC	234	ATATCGG	268	GTAGCCG
31	CAGTA	65	GGGGT	99	CGACCT	133	GATCAA	167	CTAACT	201	GCTAGAG	235	GCGAGCA	269	CCGTACA
32	TATCA	66	CGTAT	100	GCTCAG	134	TGGGTG	168	TTTGGT	202	ACTGGTG	236	GACAACA	270	GGACCAA
33	GGCTT	67	AAGGG	101	AAATGG	135	ACGTAG	169	TCCCAT	203	TCAGGTC	237	GAGGTAG	271	ACCCGGA
34	AAGCT	68	GGCAG	102	AATACG	136	TCATGA	170	TGAACT	204	TATCCCG	238	CAACACT	272	GGACGTG

(2/2)

Nr.	Barcode	Nr.	Barcode	Nr.	Barcode	Nr.	Barcode
273	AATACTG	307	GTCCGGAC	341	GTCAGCGA	375	ATCTTGCA
274	CACTCGT	308	GAAGAGCG	342	AAGACGTT	376	ACTTCTGT
275	TACCGCC	309	GCGGGCAA	343	GGAGGTCA	377	GTAAGGCT
276	GAACTCC	310	GTTTGTTA	344	CCGTTGAT	378	TGTGATGG
277	CGGTGAG	311	CTTACGCG	345	ATTGGTCG	379	TAACATCT
278	GGTCGCA	312	CCACCACC	346	CATGAATA	380	ACCGACCC
279	GTTAATG	313	TTAAAGAC	347	AACATGAC	381	GTGGTGGT
280	CCTGAGT	314	TCGGGTCT	348	CGTCCATG	382	ACTTTATC
281	ATTTGGA	315	TGCCTACC	349	GCACTCGC	383	TAGTCCGT
282	TTCTAGT	316	TATGCTTC	350	GGGTTACA	384	CAACTTGA
283	ACGCTCG	317	GACTCGAT	351	CTAACCAT		
284	ATAAGCT	318	AGAATGGA	352	AGCACAGT		
285	ATCGCTA	319	GCGATCTG	353	CTGCTATC		
286	CTGCATG	320	GCCAAGAA	354	CTTTAACT		
287	TGGAAAG	321	ATTAATGC	355	TCACGTTC		
288	GATATGC	322	GAGCGCTC	356	AAACCTTG		
289	GACGCTGG	323	GGCCATGT	357	TCCTGCAT		
290	TGCGCGCA	324	GATCGAAG	358	CTCTACTC		
291	CCTGTTAA	325	TTCGTGTC	359	GGATAATT		
292	GCTTAGCC	326	AATGCCGA	360	ATGTTAGG		
293	GAGGAAGC	327	CCAATCCA	361	ATCCTCAG		
294	GGTATAGG	328	TCGGCGTG	362	CCTTGGTG		
295	ACGAGTAC	329	ATCAGATG	363	GACTACCA		
296	CGTTGCGA	330	TACTGGTA	364	CTCGATAT		
297	TTCATTCT	331	CGCTTGGG	365	TGCGACTT		
298	AGGATCCT	332	GCGCCTTA	366	ACCCTAGA		
299	GTTCTCCA	333	GCCGTAAG	367	AAGTCTCC		
300	GTTGACTG	334	TGAGGGAT	368	AGGGCTAT		
301	ACGCGGGT	335	AACCAAAT	369	TCTACAAG		
302	CTCCCAAA	336	TAATTGAG	370	TGCTATAC		
303	CACGGACG	337	CAATGTAT	371	ATTCAGAA		
304	TCTCACGT	338	CCAGGTGG	372	TTACCGTA		
305	CGGGAGCT	339	TGAATCTC	373	CGTTCGAC		
306	GTATTTCC	340	ATGGCACA	374	AGAGACAA		

5.2.4 Description of raw sequencing data

The single-end (“R1“) sequencing lanes were generated during 6 sessions from November 2014 to December 2016 with up to 384 individual adaptor barcodes each multiplex. Sequencing files are on the HPCC biocluster of the UCR (biocluster.ucr.edu) in: ‘koeniglab/shared/OUTSIDE_SEQUENCE_DATA/CHRISTINA_CBP_RAD’ and on hard drive “OS-Botanik-Neuffer“ in the Neuffer Lab.

For better overview, each lane will be named referring to their sample composition:

"Nov14":
illumina_SN7001143flowcellB_SampleIdNA_RunId0213_LaneId5/
lane5_NoIndex_L005_R1_001.fastq.gz
File size: 17,1 GB
Total number of reads: 176478793
Barcodes used: 384

"Am1" ("America1"):
illumina_ST-J00101flowcellA_SampleIdNA_RunId0008_LaneId7/
Undetermined_S0_L007_R1_001.fastq.gz
File size: 19,1 GB
Total number of reads: 323376947
Barcodes used: 232

"Am2" ("America2"):
illumina_ST-J00101flowcellA_SampleIdNA_RunId0008_LaneId8/
Undetermined_S0_L008_R1_001.fastq.gz
File size: 18,3 GB
Total number of reads: 313280012
Barcodes used: 192

"Nov15":
illumina_ST-J00101flowcellA_SampleIdNA_RunId0016_LaneId2/
Undetermined_S0_L002_R1_001.fastq.gz
File size: 14,9 GB
Total number of reads: 166348949
Barcodes used: 192

"Eu1" ("Eurasia1"):
illumina_ST-J00101flowcellA_SampleIdNA_RunId0049_LaneId6/
Undetermined_S0_L006_R1_001.fastq.gz
File size: 17,5 GB
Total number of reads: 200247739
Barcodes used: 384

"Eu2" ("Eurasia2"):
illumina_ST-J00101flowcellA_SampleIdNA_RunId0049_LaneId7/
Undetermined_S0_L007_R1_001.fastq.gz
File size: 27,8 GB
Total number of reads: 319991542
Barcodes used: 96

5.2.5 Demultiplex-CW-7-4-17.py

```
#!/usr/bin/python

# Demultiplex-CW Python Script. Author: Christina Wesse (7.4.2017)
# USAGE: call python script with 2 arguments: 1: path/to/raw.fastq.gz 2:
path/to/barcodesfile.txt
# Barcode input file must contain barcode sequences in column 1 and sample names in
column 2 separated by a space
# Default restriction enzyme is KpnI ("Resite")

from __future__ import division
import gzip, sys, time, os.path

def isAscii(s):
      return all(ord(c) < 127 for c in s) and all(ord(c) > 32 for c in s)

def isDna (dna):
      no_bases = dna.count("A") + dna.count("G") + dna.count("C") + dna.count("T") +
dna.count("N")
      if dna.count("\n")==1:
      if ((len(dna))-1) == no_bases:
      return 1
      else:
      if no_bases == (len(dna)):
      return 1

def checkEntry (entry):
      format_ok = 0
      if entry[0].startswith("@"):
      format_ok = format_ok + 1
      if isDna(entry[1]) == 1:
      format_ok = format_ok + 1
      if entry[2]==("+\n"):
      format_ok = format_ok + 1
      if isAscii(entry[3].rstrip("\n")) is True:
      format_ok = format_ok + 1
      return format_ok

def getBarcodedict (barcodefile_input):
      dictionary = {}
      with open(barcodefile_input,'r') as f:
            for line in f:
                  k, v  = line.split()
                  dictionary[k]=v
      return dictionary

singleRead = []


def readFourLines ():
      global singleRead
      singleRead = []
      for i in range(4):
            line = f.readline()
            if not line:
                  return False
            else:
                  singleRead.append(line)
      return True
readDictionary = {}
readCount=0

def logReads ():
    logfileName=os.path.basename(barcodefile)+".log"
    f = open(logfileName, 'w')
    for key in readDictionary:
        f.write(key+":"+str(readDictionary[key])+"\n")
    f.write( "Total reads:"+str(readCount)+"\n")
    duration = time.time() - ticksStart
    f.write( "Duration:" + str(duration)+"\n")
    f.close()

######### MAIN ##########

filename = sys.argv[1]
barcodefile = sys.argv[2]
REsite = "GTACC"
barcodesdict = getBarcodedict(barcodefile)

with gzip.open(filename, 'rb') as f:
    ticksStart = time.time()
    while readFourLines():
        readCount=readCount+1
        if checkEntry(singleRead)!=4:
            logfile = open("Log-Barcodes.txt", "a")
            logfile.write("Input file is corrupted")
            continue
        dna = singleRead[1]
        if dna.find(REsite,4)<0:
            outputFilename="trash.fastq.gz"
            with gzip.open(outputFilename, "a") as myfile:
                read = str(singleRead[0]) + str(singleRead[1]) +
str(singleRead[2]) + str(singleRead[3])
                myfile.write(str(read))
                readDictionary[outputFilename] =
readDictionary.get(outputFilename,0) +1
                continue

        start = dna.find(REsite,4)
        brcd = dna[:start]
        if brcd not in barcodesdict:
            outputFilename="no_barcodes.fastq.gz"
            with gzip.open(outputFilename, "a") as myfile:
                read = str(singleRead[0]) + str(singleRead[1]) +
str(singleRead[2]) + str(singleRead[3])
                myfile.write(str(read))

                readDictionary[outputFilename] =
readDictionary.get(outputFilename,0) +1
                continue
        samplename = barcodesdict.get(brcd)
    outputFilename="sample_" + str(samplename) + ".fastq.gz"
        with gzip.open(outputFilename, "a") as myfile:
            read = str(singleRead[0]) + str((singleRead[1])[(len(brcd)):]) +
str(singleRead[2]) + str((singleRead[3])[(len(brcd)):])
            myfile.write(str(read))
        readDictionary[outputFilename] = readDictionary.get(outputFilename,0) +1
        if readCount % 100000 == 0:
        logReads()
    logReads()
```

5.2.6 Number of reads in demultiplexed raw data

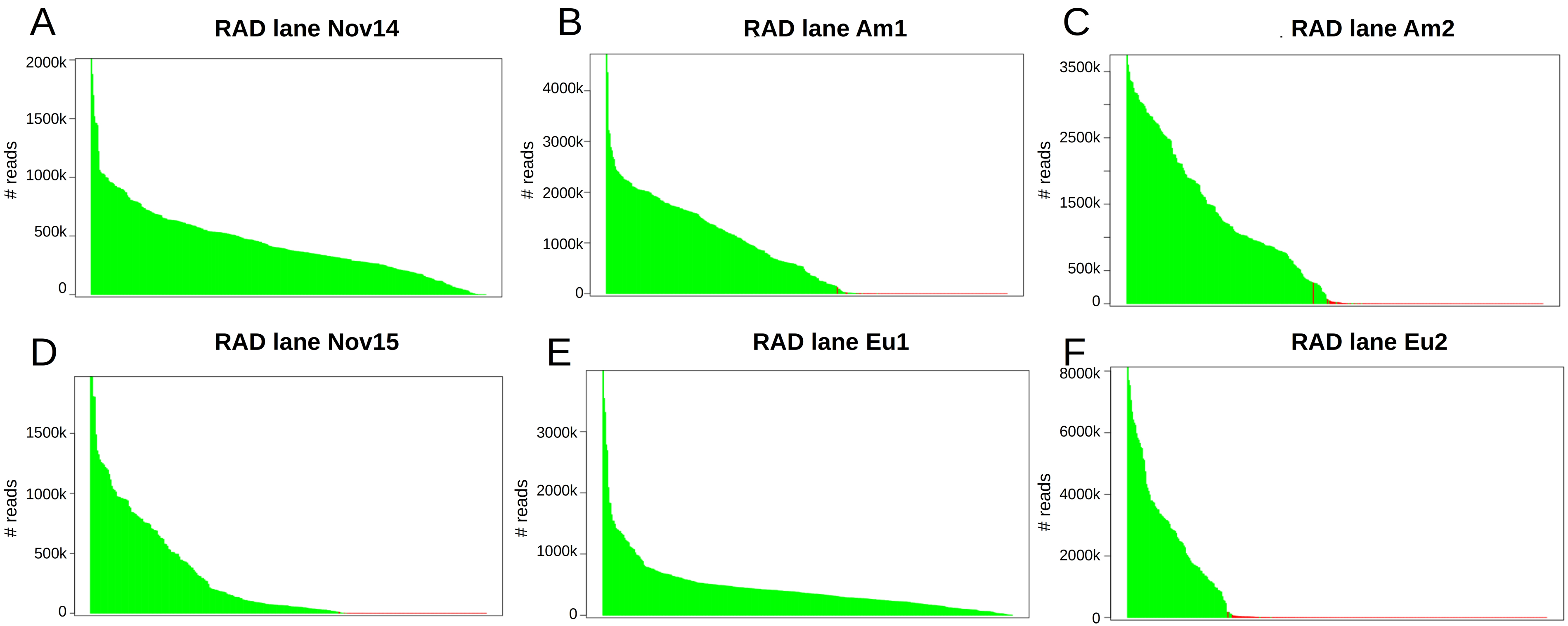

Figure S.11: Visualization of demultiplexed reads per sequencing lane. Green: barcoded sample. Red: Sequencing artifact. A: Nov14. B: Am1. C: Am2. D: Nov15. E: Eu1. F: Eu2. For data explaination see *5.2.4 Description of raw sequencing data.*

5.2.7 Read mapping and SNP discovery

Demultiplexing of the sequencing reads and removal of the barcode sequences was done with a custom python script (see *5.2.5 `Demultiplex-CW-7-4-17.py`*). A visualization of the number of reads per sequencing lane can be found in the supplement (see *5.2.6 Number of reads in demultiplexed raw data*, S.Fig. 11).

Usually, the quality of the reads declines towards the end of the sequences, therefore it is recommended to trim the reads at a certain point (e.g. Buffalo, 2015). After demultiplexing, all reads were trimmed and then aligned to a reference genome with a custom made pipeline script: Twimming was performed with *Trimmomatic* version *0.36* (Bolger et al., 2014) with a sliding window to eliminate bad quality reads, removal of the illumina adaptor sequences and verified for a minimum length of 75 bases. The trimmed reads were then mapped either to the *C. bursa-pastoris* reference genome (*`Cbp-2-2_contigs.fasta`*) or a pseudoreference (*`CbpPseudoRef.fasta`*) generated *in silico* from the concatenated genomes of *C. orientalis* and *C. rubella* using *BWA* version *0.7.12* (Li and Durbin, 2009) with default parameters but ignoring indels. *BWA* is a software package for mapping sequences against a reference genome. The algorithm BWA-MEM (Max Exact Matches) is designed for Illumina sequence reads from 70 bp – 1 Mbp. Since reads were trimmed with a minimum length of 75 bp, BWA-MEM was chosen. It has the best performace for 70 – 100 bp Illumina reads (Li and Durbin, 2009).

Sorting of the resulting alignment files then was performed with *SAMtools* version *1.4.1* (Li et al., 2009). The program *freebayes* version *1.1.0* (Garrison and Marth, 2012) was used for initial SNP calling with default parameters using the freebayes-parallel script. *Freebayes* is a variant detection tool to find SNPs, indels and MNPs (multiple nucleotide polymorphisms). The program is based on hyplotypes and SNPs are recorded in a .vcf file.

5.2.8 checkGL-17-1-2018.py

```
#!/usr/bin/python
# Check GL in a vcf file. Author: Christina Wesse (17.1.2018)
# USAGE:
# call python script with argument: path/to/chunk.txt <logfilename>

from __future__ import division
import sys

######### MAIN ##########

filename = sys.argv[1]

chunkname= str(sys.argv[2]) + ".log"

with open(filename) as f:
        for line in f:
                SNPs = line.split("\t")
                snps = SNPs[9:]
                gtCount0 = 0
                gtCount1 = 0
                gtCount2 = 0
                naCount = 0

                if snps[0] != "\n":
                        for snp in snps:
                                sample = snp.split(":")
                                if sample[0] != ".":
                                        GT = sample[0]
                                        gl = sample[7]
                                        GL=gl.split(",")
                                        if GT == "0/0" and GL[0] == "0":
                                                gtCount0 = gtCount0 + 1
                                        elif GT == "0/1" and GL[1] == "0":
                                                gtCount1 = gtCount1 + 1
                                        elif GT == "1/1" and GL[2] == "0" or GL[2] == "0\n":
                                                gtCount2 = gtCount2 + 1
                                        else:
                                                print "odd NP:" + str(snp) + " GT:" +
                                        str(GT) + " GL:" + str(GL[2])
                                else:
                                        naCount = naCount + 1

                        resultSnpsLength = naCount + gtCount0 + gtCount1 + gtCount2
                        if resultSnpsLength != len(snps):
                                numberOfErrors = len(snps) - resultSnpsLength
                                logfile = open(chunkname, "a")
                                logfile.write("Error at position " + str(SNPs[0:2]) + ":
                        Number of strange genotypes:" + str(numberOfErrors) + "\n")
                        else:
                                logfile = open(chunkname, "a")
                                logfile.write(str(SNPs[0:2]) + " ok\n")

                else:
                        logfile = open(chunkname, "a")
                        logfile.write("No SNPs in " + str(SNPs[0:2]))

        logfile = open(chunkname, "a")
        logfile.write("Python script completed\n")
```

5.2.9 `Job-genomeBias.sh`

```
# pilon is Cor, scaffold is Cru
module load samtools
outDate=$(date +"%d-%m-%y")
outTime=$(date +"%T")
outputcsv="alignBias-${outDate}.csv"
files=$(ls ../align/pseudoref*/*.sorted.bam)
echo "Sample; Cor; Cru" > ${outputcsv}
for file in ${files}; do
      sampleName=$(echo ${file} | cut -d '/' -f4 | cut -d '.' -f1 )
      echo -n "${sampleName};" >> ${outputcsv}
      samtools idxstats ${file} | head -21 | awk '{ sum += $3; }
END { print sum; }' "$@" | tr '\n' ';' >> ${outputcsv}
samtools idxstats ${file} | grep -w "scaffold_[1-8]" | awk '{ sum += $3; }
END { print sum; }' "$@" >> ${outputcsv}
done
```

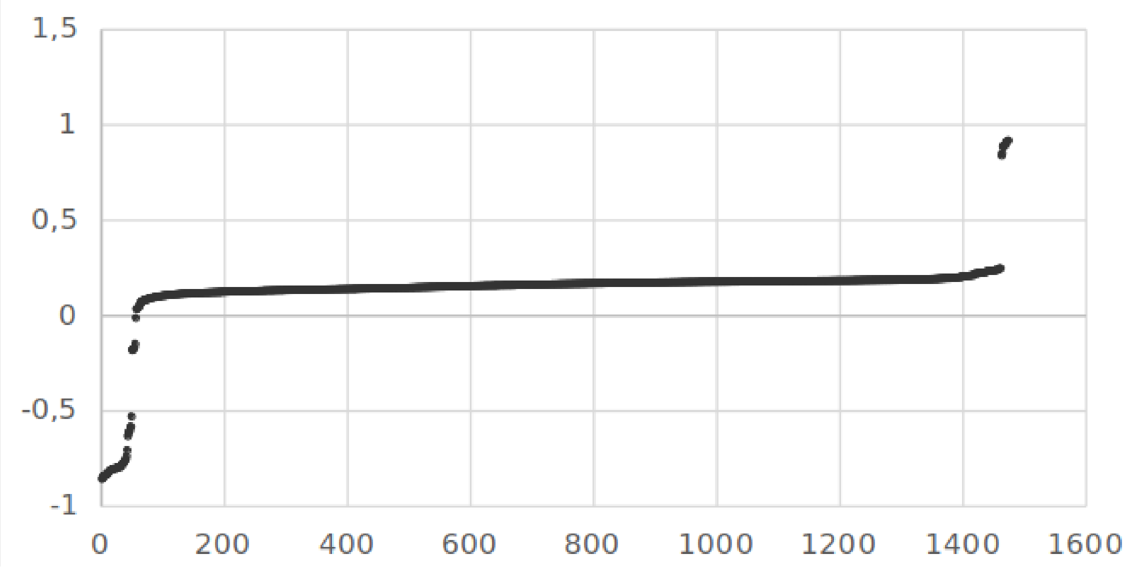

Figure S.12: Detection of diploid samples within the dataset.
Diploids showed biased mapping to either one of the reference genomes from the pseudo reference. Samples are ordered by number of mapping to either one of the genomes: Samples on the far left mapped preferably to *C. rubella*, and samples on the far right preferably to *C. orientalis*. The plateau of the graph shows that the majority of the samples are *C. bursa-pastoris* with all probablility. Only the calculated bias from mappings from the beginnings to 000020F_pilon and scaffold_8 respectively of the two reference genomes have been used to detect diploids.

5.2.10 hetCalls.sh

```
vcffile="hardfiltered-pseudo-22-2-18_2.recode.vcf"
module load plink
cat ${vcffile} | grep "^#" > header.vcf
cat ${vcffile} | grep -v "^#" | grep "pilon" > pilon.vcf
cat header.vcf pilon.vcf > Corsubset.vcf
rm pilon.vcf
plink --vcf Corsubset.vcf --het –allow-extra-chr
rm Corsubset.vcf
mv plink.het cor.het
cat ${vcffile} | grep -v "^#" | grep "scaffold" > scaffold.vcf
cat header.vcf scaffold.vcf > Crusubset.vcf
rm scaffold.vcf
plink --vcf Crusubset.vcf --het –allow-extra-chr
rm Crusubset.vcf
mv plink.het cru.het
rm header.vcf
```

5.2.11 `R-Bioclim.R`

```
install.packages("raster")
install.packages("sp")
install.packages("data.table")
library(raster)
library(sp)
library(data.table)

climate <- getData('worldclim', var='bio', res=2.5)

samples <-fread("RAD-Data-28-8-2017.csv", header=T)
coordinates <- samples[, list(Long,Lat)]

climate <- climate[[c(1:19)]]
names(climate) <-
c("BIO1","BIO2","BIO3","BIO4","BIO5","BIO6","BIO7","BIO8","BIO9","BIO10","BIO11","BIO12",
"BIO13","BIO14","BIO15","BIO16","BIO17","BIO18","BIO19")
values <- extract(climate,coordinates)
df <- cbind.data.frame(coordinates(coordinates),values)
write.table(df, "bioclimdata-all.txt", sep="\t")
```

The data is coded as follows (http://www.worldclim.org/formats1):

BIO1 = Annual Mean Temperature (in °C x 10)
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3 = Isothermality (BIO2/BIO7) (* 100)
BIO4 = Temperature Seasonality (standard deviation *100)
BIO5 = Max Temperature of Warmest Month
BIO6 = Min Temperature of Coldest Month
BIO7 = Temperature Annual Range (BIO5-BIO6)
BIO8 = Mean Temperature of Wettest Quarter
BIO9 = Mean Temperature of Driest Quarter
BIO10 = Mean Temperature of Warmest Quarter
BIO11 = Mean Temperature of Coldest Quarter
BIO12 = Annual Precipitation
BIO13 = Precipitation of Wettest Month
BIO14 = Precipitation of Driest Month
BIO15 = Precipitation Seasonality (Coefficient of Variation)
BIO16 = Precipitation of Wettest Quarter
BIO17 = Precipitation of Driest Quarter
BIO18 = Precipitation of Warmest Quarter
BIO19 = Precipitation of Coldest Quarter

5.2.12 Quality calculations and filtering

The custom made demultiplexing script (*see 5.2.5* `Demultiplex-CW-7-4-17.py`) also logs the number of sequencing reads, and samples with overall low read numbers (<50.000; *see 5.2.6 Number of reads in demultiplexed raw data*, S.Fig. 11) were removed from the dataset to improve the quality of the rest of the data.

The programm *VCFtools* version *0.1.13* (Danecek et al., 2011) was used to estimate the amount of missing data within samples of the whole dataset. It depends on the dataset itself and the researchers intuition where to put the threshold for the amount of acceptable missing data. In this case, most samples had less than 50 % missing data (Fig. S.13). To keep the amount of excluded samples rather low but also remove enough bad data to improve the quality of the rest, it was decided to exclude samples with more than 70 % missing data.

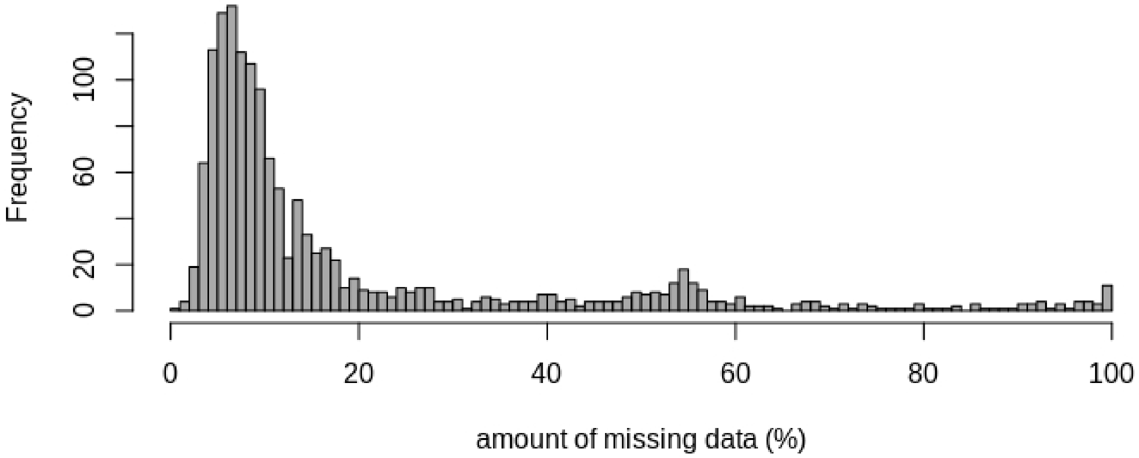

Figure S.13: Histogram showing the amount of missing sequence data for each sample.

In the field, *C. bursa-pastoris* is easily to be confused with other species from the sames genus, so it is possible that some samples are *C. orientalis, C. grandiflora* or *C. rubella* instead. Diploid samples are easily to be identified via the estimated genome size obtained from FCM. However, since not every sequenced sample had been used for FCM, it was expected that there were still other accidentally sequenced individuals within the dataset. To find these possible false samples, all sequences were aligned to an artificially created “pseudoreference genome“ concatenated from the reference genomes of *C. orientalis* and *C. rubella* (see *5.2.7 Read mapping and SNP discovery*). A custom made script was used to identify the individuals that mapped preferably to either one of the two species (see *5.2.9* `Job-genomeBias.sh`). Samples deviating more than three times the standard deviation from the mean were defined as outliers and removed from the dataset.

A custom made shell script was used to (i) seperate the mappings on the *C. rubella* genome from the ones on the *C. orientalis* genome and (ii) output the number of the heterozygosity for each locus within each sample with the program *PLINK* version *1.90b3.38* (Purcell et al., 2007) (see *5.2.10* *`hetCalls.sh`*). Then the percentages of homozygous loci were calculated and the differences between the two reference genomes compared. If the degree of heterozygosity deviated more than three times the standard deviation from the mean, those samples were defined as outliers and removed from the dataset.

A custom python script has been used to check if the determined genotypes have been assigned with a reasonable likelihood (see *5.2.8* *`checkGL-17-1-2018.py`*).

Summary

A total number of 709,542 raw SNPs were called. The resulting vcf was filtered with *VCFtools* version *0.1.13* (Danecek et al., 2011) for minimum quality of 30 and maximal fraction of missing data 70 % and for minor allele frequency (MAF) = 0.05. Samples with reads < 50,000 were removed from the vcf, as also samples which seemed to be diploid or triploid according to preferred mapping on either one of the diploid genomes within the aforementioned artificially created pseudoreference. Samples with unusually high heterozygosity were also removed. This resulted in a final vcf containing 1,273 different sequenced *C. bursa-pastoris* individuals with 13,006 high quality SNPs eventually used in the genetic diversity analysis.

5.2.13 Cluster climate adaptation

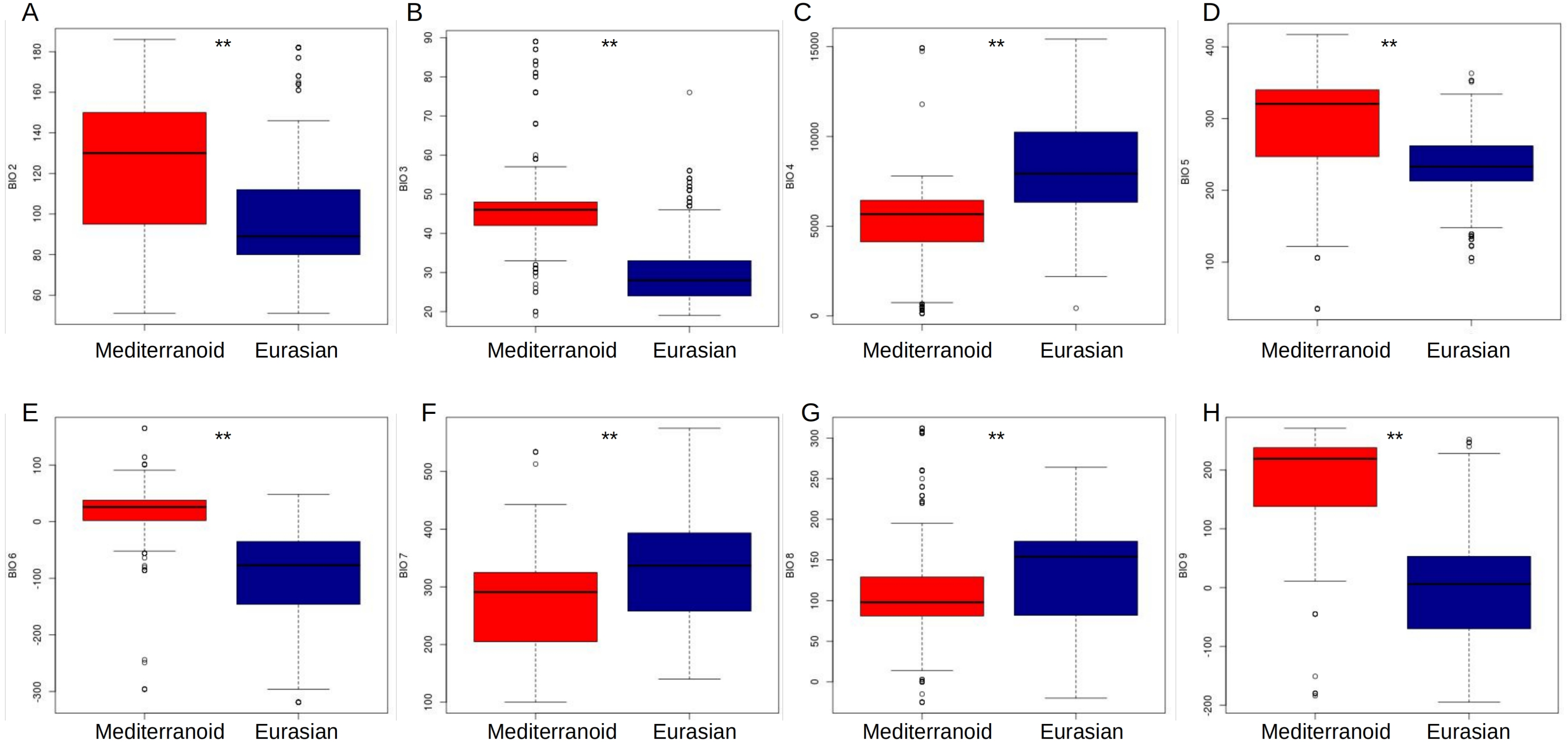

Figure S.14 (1/2): Differences in climate adaptation between the two clusters. A: BIO2 = Mean Diurnal Range. B: BIO3 = Isothermality.C: BIO4 = Temperature Seasonality. D: BIO5 = Max Temperature of Warmest Month. E: BIO6 = Min Temperature of Coldest Month.F: BIO7 = Temperature Annual Range. G: BIO8 = Mean Temperature of Wettest Quarter. H: BIO9 = Mean Temperature of Driest Quarter. **: $p < 0.003$. *: $0.003 < p < 0.05$.

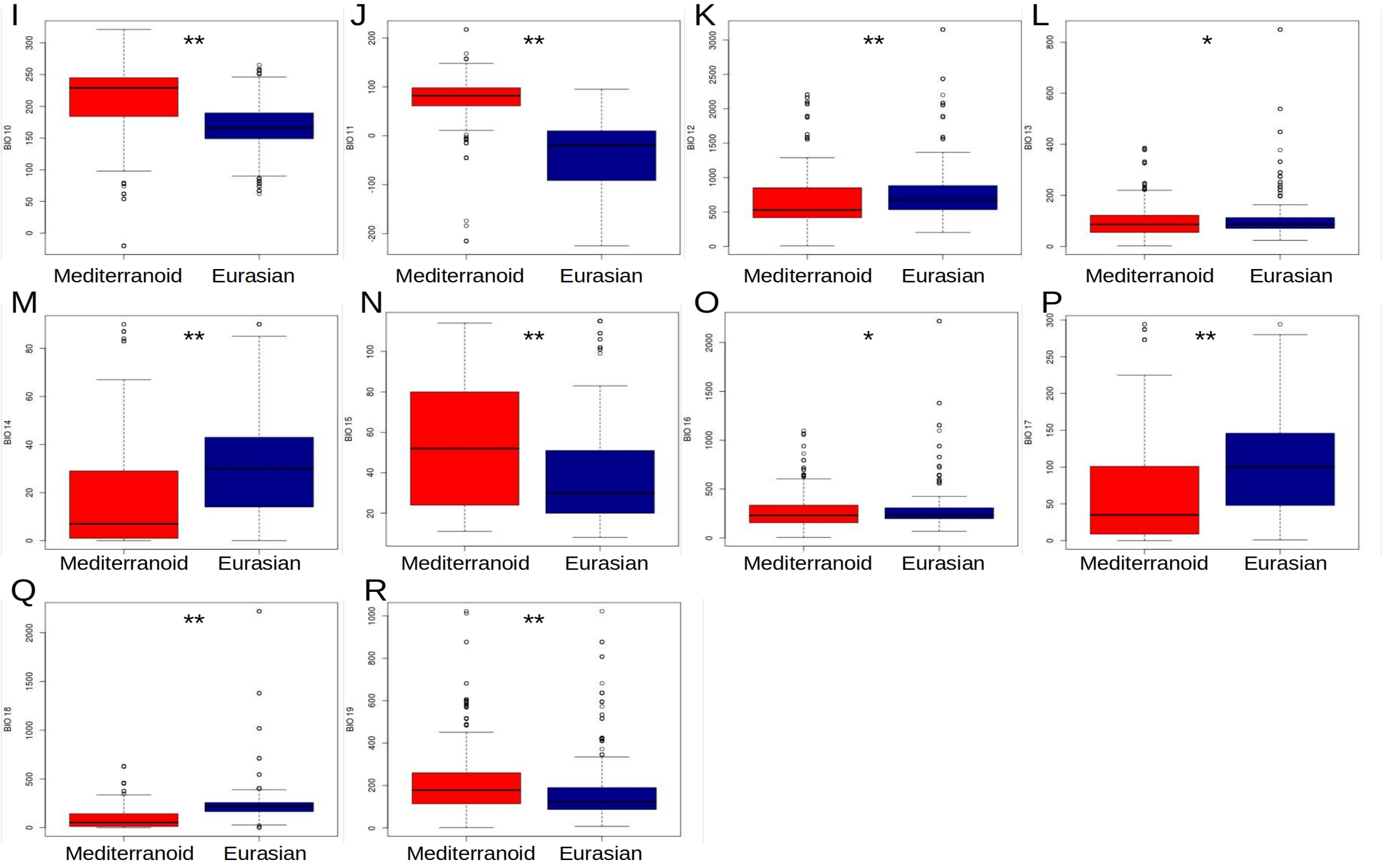

Figure S.14 (2/2): I: BIO10 = Mean Temperature of Warmest Quarter. J: BIO11 = Mean Temperature of Coldest Quarter. K: BIO12 = Annual Precipitation. L: BIO13 = Precipitation of Wettest Month. M: BIO14 = Precipitation of Driest Month. N: BIO15 = Precipitation Seasonality. O: BIO16 = Precipitation of Wettest Quarter. P: BIO17 = Precipitation of Driest Quarter. Q: BIO18 = Precipitation of Warmest Quarter. R: BIO19 = Precipitation of Coldest Quarter. **: $p < 0.003$. *: $0.003 < p < 0.05$.

Danksagung

Diese Arbeit wurde im Zeitraum von Januar 2015 bis Mai 2018 im Institut für Biologie und Chemie an der Universität Osnabrück unter Leitung von Frau apl. Prof. Dr. Barbara Neuffer durchgeführt. Ihr gilt mein besonderer Dank für die Durchführung dieser Arbeit sowie ihre ständige Diskussions- und Hilfsbereitschaft.

Dieses Projekt wurde finanziell von der Deutschen Forschungsgemeinschaft (DFG) unterstützt (NE 314/11-2) und war Teil des DFG Schwerpunktprogramms 1529 „Adaptomics“. Ich danke der DFG und allen Kollegen des Schwerpunktprogramms, die ich während zahlreicher Tagungen und Workshops kennen lernen durfte für den regen Wissensaustausch.

Ferner danke ich Herrn Prof. Dr. Detlef Weigel für die freundliche Unterstützung des Projekts und vor allem die Möglichkeit am Max-Planck-Institut für Entwicklungsbiologie in Tübingen die DNA-Sequenzierungen durchzuführen. Insbesondere danke ich Dr. Danelle Seymour für die Einarbeitung in die RAD Library Prep und Dr. Daniel Koenig für die Unterstützung bei der Datenauswertung und der Gelegenheit zweier gastwissenschaftlicher Aufenthalte an der University of California Riverside. Meine Zeit in Kalifornien habe ich dank der freundlichen Arbeitsgruppe vor Ort sehr genossen.

Natürlich danke ich ebenso den jetzigen und ehemaligen Mitarbeiterinnen und Mitarbeitern der Arbeitsgruppe Botanik vor Ort für die freundliche Zusammenarbeit und insbesondere Herrn Rudolf Hungerland-Grupe für die vielen Diskussionen und Anregungen. Vielen Dank auch an Herrn Martin Rebilas, Frau Denise Pietzka, Frau Annica Härig, Frau Johanne Tietmeyer und Frau Madita Knieper, die mit mir gemeinsam die Versuchspflanzen im Botanischen Garten betreut haben. Ganz besonders danke ich in diesem Zusammenhang auch den Mitarbeiterinnen und Mitarbeitern des Botanischen Gartens, vor allem Frau Birgit Ilgener, Frau Alexandra Lohstroh und dem Freilandmeister Herrn Hermann Brüggemann. Ferner danke ich Frau Lucille Schmieding und Frau Petra Köhne für Unterstützung bei Verwaltung und Bewätigung von Anträgen und Formularen.

Außerdem bin ich dankbar für die Liebe und Geduld meines Lebensgefährten Bastian Breit und die persönliche Unterstützung meiner Familie, die zum Gelingen dieser Arbeit in hohem Maße beigetragen haben.

Erklärung über die Eigenständigkeit der erbrachten wissenschaftlichen Leistung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise ~~entgeltlich~~ / unentgeltlich geholfen.

1. ....Die Co-Autoren der Manuskripte werden an den …...entsprechenden Stellen in der Dissertation namentlich erwähnt.
2. ....…
   ....
3. ....…
   …..…

Weitere Personen waren an der inhaltlichen materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

……………………….	……………………….
(Ort, Datum)	(Unterschrift)

Curriculum vitae

Christina Wesse, geb. 24. September 1987 in Rendsburg, Staatsangehörigkeit deutsch

Schulausbildung

1994 – 1998	Grundschule Dellstedt-Wrohm
1998 – 2007	Gymnasium Heide-Ost in Heide (Abitur)

Hochschulausbildung

10/2008 – 05/2014	Studium der Biologie an der Carl-von-Ossietzky-Universität in Oldenburg
12/2011	Bachelor of Science
05/2014	Master of Science
09/2014 – 12/2014	Wissenschaftliche Mitarbeiterin in der AG Botanik der Universität Osnabrück bei apl. Prof. Dr. rer. nat. Barbara Neuffer
Seit 01/2015	Doktorandin in der Biologie bei apl. Prof. Dr. rer. nat. Barbara Neuffer

Berufserfahrung

01/2019 - 09/2019	Mitarbeiterin im Qualitätsmanagement, AniCon Labor GmbH, Höltinghausen
Seit 11/2019	Data Scientist, Cordes und Graefe, Stuhr

Auszeichnungen

09/2017	Best Poster Prize in "Biodiversity and Ecosystem", Botanikertagung 2017 in Kiel, DBG (Deutsche Botanische Gesellschaft)
09/2017	Talk Prize, Eduard Strasburger Workshop 2017 in Bremen, DBG

Publikationen

König, E., Wesse, C., Murphy, A. C., Zhou, M., Wang, L., Chen, T., ... & Bininda-Emonds, O. R. (2013). Molecular cloning of the trypsin inhibitor from the skin secretion of the Madagascan Tomato Frog, Dyscophus guineti (Microhylidae), and insights into its potential defensive role. *Organisms Diversity & Evolution*, *13*(3), 453-461.

Raupach, M. J., Barco, A., Steinke, D., Beermann, J., Laakmann, S., Mohrbeck, I., ... & Segelken-Voigt, A. (2015). The application of DNA barcodes for the identification of marine crustaceans from the North Sea and adjacent regions. *PloS one*, *10*(9), e0139421.

Neuffer, B., Wesse, C., Voss, I., & Scheibe, R. (2018). The role of ecotypic variation in driving worldwide colonization by a cosmopolitan plant. *AoB Plants*, *10*(1), ply005.

Ganderkesee, 30. Juni 2020