Genomic Signals of Adaptation in the Allotetraploid Weed

*Capsella bursa-pastoris*

by

Adriana Salcedo

A thesis submitted in conformity with the requirements for the degree of Master of Science
Department of Ecology and Evolutionary Biology
University of Toronto

© Copyright by Adriana Salcedo 2015
Genomic signals of adaptation in the allotetraploid weed *Capsella bursa-pastoris*

Adriana Salcedo

Master of Science

Department of Ecology and Evolutionary Biology
University of Toronto

2015

Abstract

Polyploids are often exceptionally evolutionarily successful, and increased adaptive potential may contribute to their prevalence. However, adaptation is not well characterized in polyploids, and the extent that it contributes to polyploid success is currently unclear. Here we investigate the genomic basis of adaptation in one of the world’s most successful weeds. We obtained genotype-by-sequencing (GBS) sequences from 261 *C. bursa-pastoris* accessions and 24 whole genome sequences spanning Eurasia. We identified climate associations and found adaptive signatures consistent with contemporary adaptation in recently colonized Asia, and longer adaptive divergence in the ancestral European range. Signatures of adaptation through selective sweep scans suggest preadapted parental genomes may also contribute to local adaptation in similar habitats. However, native climate did not predict fitness in a common garden. Local environmental adaptation is then likely a subtle factor in *C. bursa-pastoris*’ success, although the features of its diploid progenitor genomes notably influenced its adaptation.
Acknowledgments

I would like to thank the many people who helped me during my Masters studies at the University of Toronto and who contributed to this thesis. I cannot sufficiently thank my supervisors John Stinchcombe and Stephen Wright for their support, guidance, and patience throughout the completion of this thesis. You gave me an exciting project that offered valuable opportunities and challenged me to develop my skills in many ways. Working with you helped shape me into a better scientist and leader and I will use these insights throughout the rest of my career. I enjoyed all of our interactions and I could not have asked for better mentors. Many thanks to my committee members Marc Johnson and Asher Cutter, as well as Martin Lascoux and Amandine Cornille for their contributions to this work. Thank you to Bruce Hall, Niroshini, Najma, and Viktor for your hard work on the common garden experiment. I am also very thankful to the EEB department for allowing me the resources and flexibility to complete this work and providing a stimulating academic environment.

I would also like to thank my labmates in the Wright and Stinchcombe labs for their support and friendship, especially Wei Wang, Emily Josephs, Stuart Campbell, Brechann McGoey, and Robert Williamson. Special thanks to Billie Gould and Michelle Afkhami for being excellent mentors and friends, I really learnt so much from you both. I am also extremely thankful for the friendship, kindness, and support I received from Tia Harrison and Alison Wardlaw, sharing office with you was wonderful and I cannot imagine the last two years without you. Many thanks to my other EEB friends who made this department a welcoming place and helped me create some great memories outside of the lab. Finally, special thanks to my parents, Caro, and Shash. Your love and support mean the world to me and I could not have done this without you.
Table of Contents

Table of Contents .................................................................................................................. iv
List of Tables ....................................................................................................................... vi
List of Figures ....................................................................................................................... vii
List of Appendices ............................................................................................................ ix

1 Literature Review .......................................................................................................... 1
  1.1 The evolutionary processes of adaptation ................................................................. 2
  1.2 The genomic basis of adaptation .............................................................................. 4
  1.3 Polyploidy and adaptation ........................................................................................ 7
  1.4 Study system ............................................................................................................. 11

2 Study ............................................................................................................................... 13
  2.1 Introduction ............................................................................................................... 13
  2.2 Methods .................................................................................................................... 16
    2.2.1 Genomic Dataset ............................................................................................... 16
    2.2.2 Phasing ............................................................................................................. 17
    2.2.3 Demographic analysis ......................................................................................... 18
    2.2.4 Environmental Associations .............................................................................. 20
    2.2.5 Enrichment analyses .......................................................................................... 23
    2.2.6 SNP origins ....................................................................................................... 24
    2.2.7 Sweep scans and proximity analysis .................................................................. 24
    2.2.8 Linkage disequilibrium and effect size .............................................................. 26
    2.2.9 Common garden GWAS .................................................................................... 27
  2.3 Results ......................................................................................................................... 28
    2.3.1 Population genetics ............................................................................................ 28
2.3.2 Environmental Associations ................................................................. 33
2.3.3 Enrichment analyses ............................................................................. 35
2.3.4 SNP segregation patterns ....................................................................... 37
2.3.5 Selection scans ...................................................................................... 38
2.3.6 Linkage and effect sizes in putatively adaptive SNPs ......................... 41
2.3.7 Phenotypic experiment .......................................................................... 41

2.4 Discussion ................................................................................................. 43
2.4.1 Local adaptation in Capsella bursa-pastoris ........................................ 43
2.4.2 Demographic and adaptive history ...................................................... 45
2.4.3 Polyploidy and adaptation ..................................................................... 50
2.4.4 Methodological considerations and future directions ......................... 56

3 Conclusions .................................................................................................. 58

4 Figures and Tables ....................................................................................... 60
References ....................................................................................................... 101

Appendix ......................................................................................................... 128
List of Tables

Table 1..................................................................................................................72
Table 2..................................................................................................................84
Table 3..................................................................................................................85
Table 4..................................................................................................................87
Table 5..................................................................................................................88
Table 6..................................................................................................................89
Table 7..................................................................................................................91
Table 8..................................................................................................................99
Table 9...............................................................................................................100
List of Figures

Figure 1 ........................................................................................................60
Figure 2 ........................................................................................................61
Figure 3 ........................................................................................................62
Figure 4 ........................................................................................................63
Figure 5 ........................................................................................................64
Figure 6 ........................................................................................................65
Figure 7 ........................................................................................................66
Figure 8 ........................................................................................................67
Figure 9 ........................................................................................................68
Figure 10 ......................................................................................................69
Figure 11 ......................................................................................................70
Figure 12 ......................................................................................................71
Figure 13 ......................................................................................................73
Figure 14 ......................................................................................................74
Figure 15 ......................................................................................................76
Figure 16 ......................................................................................................77
Figure 17 ......................................................................................................78
Figure 18 ......................................................................................................79
Figure 19 ......................................................................................................80
Figure 20 ......................................................................................................81
Figure 21 ......................................................................................................82
Figure 22 ......................................................................................................90
Figure 23 ......................................................................................................93
Figure 24 ......................................................................................................97
Figure 25 ......................................................................................................98
1 Literature Review

Understanding the evolutionary processes underlying local adaptation remains a fundamental question in biology as adaptation has important ecological consequences for populations and species (Hendry 2013, Lee et al. 2014). Rapid adaptation can be especially important for shaping species ranges, and determining population success in new environments (Lee 2002; Sexton et al. 2009). While the causes of range limits vary widely by species, both theoretical and empirical studies show a lack of adaptation can restrict species’ ranges (Griffith and Watson 2006; Bridle and Vines 2007; Gaston 2009; Colautti et al. 2010). Conversely, rapid adaptation can facilitate range expansion, promote invasiveness, and allow populations to persist in changing climate (Sakai et al. 2001; Maron et al. 2004; Leimu and Fischer 2008; Prentis et al. 2008; Allan and Pannell 2009; Franks 2011; Hoffmann and Sgrò 2011; Franks and Hoffmann 2012; Anderson et al. 2012; Colautti and Barrett 2013). Local adaptation throughout a species range also often contributes to the maintenance and distribution of its genetic variance (Yeaman and Jarvis 2006; Mitchell-Olds et al. 2007a; Lee and Mitchell-Olds 2011), and ecological divergence through adaptation can also contribute to speciation (Schluter 2009). However, as adaptation is a complex process shaped by diverse evolutionary, ecological, and genetic factors, the evolutionary mechanisms of adaptation are not fully resolved (Olson-Manning et al. 2012).

While empirical and theoretical work have advanced our understanding of the ecological and evolutionary dynamics of adaptation, researchers have only begun to link these processes to the genome (Kawecki and Ebert 2004; Stapley et al. 2010; Anderson et al. 2011; Savolainen et al. 2013). Data on the genetic basis of adaptation in species with diverse genomes and evolutionary history is especially lacking. Resolving the genetic basis of adaptation is essential for fully understanding its evolutionary mechanisms as the complex interactions between evolutionary forces and the genome shape the adaptive process and thus its ecological consequences (Savolainen et al. 2013).
1.1 The evolutionary processes of adaptation

While diverse evolutionary factors influence adaptation, genetic variation is one of the most important factors determining the rate and extent of adaptation (Kawecki and Ebert 2004). The classic breeder’s equation reflects the essential role of genetic variation for evolutionary change, showing that the amount of phenotypic change in a trait with directional selection is directly proportional to the additive genetic variance in that trait (Falconer and Mackay 1996). Modern theoretical (Blows & Hoffman, 2005; Kawecki & Ebert, 2004; Kirkpatrick & Barton, 1997a) and empirical work (Hoffmann et al. 2003; Eckert et al. 2008; Pujol and Pannell 2008; Gaston 2009; Colautti et al. 2010; Kelly et al. 2012) confirm this principle, demonstrating that a lack of genetic variation slows adaptation even under strong selection, and consequently limits range size and increases the probability of extinction with environmental change (Burger et al. 1995). A lack of genetic variance in trait combinations, which can arise through pleiotropy, linkage or trade-offs among traits, is also important in limiting the evolution of adaptive trait combinations (Etterson and Shaw 2001; Blows and Hoffman 2005; Chen and Lübberstedt 2010). Demographic factors that reduce genetic variation, such as genetic bottlenecks and small population size can then limit adaptive potential (Willi et al. 2006). Accordingly, invasive species that experience severe founder events often show slower or reduced adaptation, although the extent that these processes limit adaptation can vary with the severity of the bottleneck (Lambrinos 2004). Subsequent introductions and the conversion of epistatic variance to additive variance after a bottleneck can alleviate the effects of founder events and bottlenecks (Lambrinos 2004; Turelli and Barton 2006; Dlugosch and Parker 2008). However, the joint effects of selection, drift, and migration will also influence population diversity and adaptive success.

Selection, drift, migration, and diversity have complex interactions that shape how adaptation proceeds in a population (Kawecki and Ebert 2004). The balance between drift and selection determine selection efficacy and thus how efficiently adaptation can proceed (Charlesworth 2009; Akashi et al. 2012; Hough et al. 2013). Many ecologically important factors increase the influence of drift relative to selection, decreasing effective population size (Ne) and selection efficacy. Lower recombination rates decrease Ne and can lead to diminished selection efficacy in inbreeding populations or around selected variants through linked positive selection or background negative selection (Charlesworth 2003, 2009; Hough et al. 2013; Slotte 2014; Barrett et al. 2014). Similarly, population subdivision also generally decreases Ne because
variable demes size and success will lead to ongoing colonization and extinction, strengthening linkage and drift within demes (Barton 1993; Whitlock and Barton 1997; Pannell and Charlesworth 2000; Pannell and Fields 2014). Moderate gene flow among divergent populations can increase diversity and maintain adaptive potential within demes (Whitlock and Barton 1997; Lenormand 2002; Griswold and Whitlock 2003; Blanquart et al. 2012), although strong gene flow can swamp populations with maladaptive genes that impede local adaptation (Débarre, Ronce, & Gandon, 2013; Kirkpatrick & Barton 1997b; Lenormand, 2002; Yeaman & Otto, 2011). While selection against migrants in divergent demes reduces effective gene flow (Feder et al. 2012b), when migration is sufficiently high it will overwhelm both selection and drift (Blanquart et al. 2012). Similarly, when migration and selection have similar strengths, drift can dominate (Yeaman and Otto 2011). Selection must then be strong enough to overcome the both homogenizing effects of migration and the stochastic effects of drift to produce adaptation.

Empirical studies support the importance of strong selection for adaptation, as environmental divergence strongly predicts population divergence and adaptation often requires either a steep environmental gradient or large geographic distance (Galloway and Fenster 2000; Smith et al. 2005; Bridle and Vines 2007; Leimu and Fischer 2008; Hereford 2009; Gonzalo-Turpin and Hazard 2009; Lee and Mitchell-Olts 2011; Sexton et al. 2014).

Other factors can also influence the way that local adaptation proceeds when populations face a new environmental pressure. When colonizing a new habitat, preadapted individuals can sort themselves into their appropriate habitats, reducing the need for further local genetic adaptation (Quinn et al. 2001; Lambrinos 2004; Leger and Rice 2007). Alternately, highly plastic, generalist genotypes may spread throughout the species range instead of allowing genetic differentiation among populations (Parker et al. 2003; Ross et al. 2008; Latta 2009). In this case, plasticity may impede adaptation if it closely matches the optimal genotype for a given environment, or if it promotes gene flow among divergent populations (Ghalambor et al. 2007; Crispo 2008; Chevin et al. 2012). However, plasticity may also promote adaptation if it does not produce the optimal phenotype for an environment but allows populations to persist in a novel environment long enough to adapt (Price et al. 2003). Plasticity can then be important for rapid responses to novel selection pressures, such as anthropogenic climate change (Anderson et al. 2012), but may then give way to genetic adaptation through genetic assimilation (Price et al. 2003; Pigliucci et al. 2006; Nicotra et al. 2010; Chevin et al. 2012). Plasticity and genetic sorting
are clearly both important to population persistence but their effect on adaptation are then likely to depend on both organismal and evolutionary factors.

1.2 The genomic basis of adaptation

Although theoretical and empirical work have improved our understanding of how broad evolutionary processes contribute to adaptation, the interactions among evolutionary forces, phenotype, and genotype are still poorly understood (Ehrenreich and Purugganan 2006). Uncovering the genomic basis of adaptation is essential to understanding adaptation as genomic features such as ploidy, pleiotropy, and past adaptation will affect how adaptation proceeds, while evolutionary mechanisms will also shape the genetic architecture of adapted traits (Savolainen et al. 2013; Lee et al. 2014). The interaction of selection and migration with the genome can affect the physical distribution of adaptive variants in the genome as well as their effect size distribution depending on their relative strength and genomic features such as recombination rate (Yeaman and Whitlock 2011; Feder et al. 2012a; Nosil and Feder 2012). Current genomic diversity, mutation rate, and past evolutionary history will also determine if adaptation is more likely to proceed from standing variation or new mutation (Barrett and Schluter 2008), and subsequently influence the probability of extinction, the adaptive process, and genetic trait architecture in different ways. Pleiotropic constraints can also determine the type of loci recruited in adaptation and produce convergent evolution among populations (Stern and Orgogozo 2008). Recent theoretical work and a growing number of genomic studies have progressed towards understanding these processes, but more diverse empirical data is needed to fully resolve them.

Selection, migration, and adaptive and demographic history all contribute to shaping the genetic architecture of adaptive traits. Ecological divergence models posit that genetic architecture can vary with the time since divergence and the relative strength of selection and migration (Feder et al. 2012b). Early in divergence with gene flow, locally adaptive alleles may be more likely to persist if they occur around large effect variants under strong selection and lead to concentrated clusters of adaptive variants across the genome known as “divergence islands”, a process known as “divergence hitchhiking” (Nosil et al. 2009; Yeaman and Whitlock 2011; Feder et al. 2012a; b). Loci under divergent selection may then form strongly linked clusters in
the genome as strongly selected variants will impede the introgression of variants linked to immigrant alleles in their genomic region. Divergence islands are especially likely to form with low local recombination, which may occur through chromosomal inversions or rearrangements (Noor et al. 2001; Kirkpatrick and Barton 2006; Yeaman and Whitlock 2011; Yeaman 2013). Selection can further maintain clusters of adaptive variants to preserve favourable allele combinations (Via and West 2008; Nosil et al. 2009; Yeaman and Whitlock 2011; Feder et al. 2012b). Higher migration, and weak selection for new mutations near divergent variants are also likely to promote the formation and maintenance of divergence islands, as the relative advantage of clustering diminishes with lower migration rates (Noor et al. 2001; Kirkpatrick and Barton 2006; Yeaman and Whitlock 2011; Yeaman 2013). Strong selection at many loci reduces the effective migration rate by eliminating migrants from divergent demes, and produces more diffuse, genome wide differentiation that involves many loci of small effect (“genomic hitchhiking”) (Nosil et al. 2009; Feder et al. 2012a; Nosil and Feder 2012). Linkage of small effect alleles in genomes can also promote genome-wide selection if, once a population has built enough variation that locally fit genomes emerge, selection on local allele combinations enhances within-population linkage and further differentiates populations (Flaxman et al. 2014). Divergence hitchhiking may be more important in differentiating populations early in divergence, while genomic hitchhiking may become dominate divergence later in the process (Feder et al. 2012a). However, the relevance of these models of divergence still remains unclear and further empirical study that clarifies their role in different adaptive situations is necessary (Nosil and Feder 2012).

The adaptive process and the resulting trait architecture will also depend on whether adaptation occurs from standing variation or new mutation. Adaptation from standing variation will often occur more quickly than adaptation from new mutation and is likely essential for population survival after a bottleneck or sudden environmental change when new mutations are rare (Hermisson and Pennings 2005; Prentis et al. 2008; Orr and Unckless 2008; Messer and Petrov 2013). It is also likely to favour different types of variants than adaptation from new mutation. During adaptation from standing variation, the success of alleles largely depends on their frequency prior to selection (Innan and Kim 2004; Hermisson and Pennings 2005). Previously neutral or weakly beneficial alleles have a greater chance of fixing with selection and previously deleterious alleles are much less likely to fix (Hermisson and Pennings 2005; Barrett
and Schluter 2008; Pritchard et al. 2010). Consequently, weakly selected loci segregating in the population are more likely to contribute to adaptation than small effect new mutations as they are less likely to be lost by drift (Hermisson and Pennings 2005). Adaptation from standing variation is also likely to involve more small effect loci as it often occurs for polygenic traits which are less likely to fix alleles than single locus traits and tend to conserve more variation (Willi et al. 2006; Chevin and Hospital 2008; Coop et al. 2009; Pritchard et al. 2010). Polygenic adaptation can further reduce the prevalence of large effect adaptive variants as selection on polygenic traits reduces the selection coefficient at each locus resulting in fewer divergence islands and weaker signals of selection (Gavrilets and Vose 2005). Adaptation from standing variation further reduces molecular signatures of divergence by diminishing hitchhiking around adaptive variants as selected variants will already occur on various backgrounds before increasing in frequency, producing subtler “soft sweep” signatures as they increase in frequency than traditional “hard sweeps” seen in adaptation from new mutation (Przeworski et al. 2005; Caruso et al. 2006; Chevin and Hospital 2008; Strasburg et al. 2012; Messer and Petrov 2013). As these properties make soft sweeps difficult to detect (Teshima et al. 2006), few clear empirical examples of sweeps from standing variation exist and it is not well characterized in nature (Barrett and Schluter 2008). Nevertheless, the rarity of hard sweeps and indirect evidence of adaptation from standing variation using allele frequency shifts, environmental correlations, or comparisons to ancestral populations suggest adaptation from standing variation through polygenic traits is widespread, and may dominate adaptation for ecologically relevant traits in a variety of species (Barrett and Schluter 2008; Coop et al. 2009; Hancock et al. 2010a; Pritchard et al. 2010; Fournier-Level et al. 2011; Deagle et al. 2012; Fu and Akey 2013; Carneiro 2014).

Finally, pleiotropic constraints can also affect how traits respond to selection and limit the types of loci involved in adaptation. Pleiotropy can restrict adaptation by contributing to trait correlations that can limit a trait’s evolutionary response (Etterson and Shaw 2001; Otto 2004; Blows and Hoffman 2005; Chen and Lübberstedt 2010). On average, pleiotropy reduces net selection on a beneficial mutation on a single trait by half, dramatically reducing the adaptive response (Otto 2004). Consequently, selection may favour mutations with fewer pleiotropic consequences (Stern 2000, 2013; Carroll 2008). Proteins with multiple functions that are expressed in multiple tissues, such as developmental genes, may then be more likely to adapt through cis-regulatory mutation than coding sequence changes (Gompel et al. 2005; Jeong et al.
Cis-regulatory mutations are thought to have fewer pleiotropic effects because their modular nature allows precise gene expression control and they are less likely to disturb upstream and downstream network interactions (Prud’homme et al. 2006; Carroll 2008; Stern and Orgogozo 2008). Cis-regulatory mutations may then facilitate adaptation at shorter timescales and for more dynamic traits (Wray 2007), although coding sequence changes also contribute to adaptation, especially when they have few pleiotropic consequences (Hoekstra 2006). Regardless of the type of site involved, selection may also repeatedly favour adaptation at preferred loci when a trait has strong pleiotropic constraints, resulting in convergent evolution (Stern 2013; Lee et al. 2014). Convergent evolution is commonly seen in nature across taxa. Although other factors also influence the probability of convergence (Ralph and Coop 2010; Stern 2013), repeated use of the same loci suggest that pleiotropic effects often limit which sites change adaptively (Hoekstra 2006; Hohenlohe et al. 2010; Deagle et al. 2012; Keller et al. 2013; Lee et al. 2014; Tiffin and Ross-Ibarra 2014). However, recent work suggests organisms may escape pleiotropic constraints through gene duplications, and allow coding mutations to also contribute strongly to adaptation (Hoekstra and Coyne 2007; Des Marais and Juenger 2010). The relative importance of these processes for adaptation remains contentious and requires further empirical support.

### 1.3 Polyploidy and adaptation

Polyploids are unusually evolutionarily successful and their unique genomic features may enhance their adaptive potential and contribute to their prevalence (Petit and Thompson 1999; Otto and Whitton 2000; Otto 2007; Soltis et al. 2010, 2014b; Husband et al. 2013). Polyploids often occupy larger ranges than their diploid progenitors (Levin 2002) and an excess of invasive and crop plants are polyploid (Pandit et al. 2006, 2011; Prentis et al. 2008; Treier et al. 2009; Beest et al. 2012). Broad environmental tolerances contribute to increased polyploidy, range size and productivity relative to diploids (Levin 2002; Brochmann et al. 2004; Ramsey 2011; Manzaneda et al. 2012) and as a result polyploids are also especially common in stressful environments such as at high latitudes (Brochmann et al. 2004; Suda et al. 2007; Popp et al. 2007; Manzaneda et al. 2012; Husband et al. 2013). As rapid adaptation is important to range expansion, niche differentiation, and invasiveness (Thompson 1992; Levin 2002; Prentis et al.
2008; Colautti and Barrett 2013), adaptation likely plays a key role in polyploid success. However, as the ecological consequences of polyploidy are not well understood, the importance of rapid adaptation to polyploid success remains unclear (Soltis and Soltis 2000; Hegarty and Hiscock 2008; Beest et al. 2012).

Duplicated polyploid genomes harbour increased genetic variation which may promote rapid adaptation in polyploids (Wendel 2000; Mcgrath and Lynch 2012). Gene duplication itself is important for adaptation as it generates novel genetic variation and releases pleiotropic constraints (Lynch and Conery 2000; Wendel 2000; Zhang 2003; Mcgrath and Lynch 2012). Whole genome duplications also generate variation through these processes but allow greater divergence and preserve more duplicated genes than individual gene duplications (Lynch and Conery 2000; Wendel 2000; Blanc and Wolfe 2004; Birchler 2012; Mcgrath and Lynch 2012). The retention of many gene duplicates following whole genome duplications is likely adaptive as gene loss following whole genome duplication often differs by functional category or subgenome, and is consistent among polyploidization events (Adams and Wendel 2005; Chang et al. 2010). Unlike individual gene duplications, whole genome duplication can also generate variation by promoting genomic rearrangements through transposable element activation or nonhomeologous recombination (Otto and Whitton 2000; Soltis and Soltis 2000; Otto 2007; Hegarty and Hiscock 2008; Pyhäjärvi et al. 2013), which may lead to adaptive phenotypic changes such as shifts in flowering time (Schranz and Osborn 2000; Pires et al. 2004). Whole genome duplication also produces regulatory variation that may further contribute to phenotypic variation (Wang et al. 2006; Finigan et al. 2012) as it allows for more possible expression levels in dosage-regulated genes (Osborn et al. 2003), and promotes epigenetic modifications that may shift gene expression from the mid-parent value leading to novel phenotypes (Otto 2003; Adams and Wendel 2005; Buggs et al. 2010). Through these evolutionary processes, whole genome duplications then increase phenotypic variation substantially beyond single gene duplications (Finigan et al. 2012), and may increase the likelihood of rapid adaptation.

Allopolyploids may incorporate even more variation, and thus adaptive potential, than autopolyploids, through their hybrid genomes (Weiss-Schneeweiss et al. 2013). Hybridization can facilitate geographical and ecological range expansion by introducing novel, or locally adapted alleles, as well as novel allele combinations (Ellstrand and Schierenbeck 2000; Rieseberg et al. 2007). Hybridization can also generate new variation through genomic
rearrangements and regulatory variation (Riddle et al. 2010). Multiple origins and introgression with diploid progenitors can further augment variation and help introduce preadapted alleles (Soltis and Soltis 1999, 2000; Meimberg et al. 2009). As a result, hybrids often show intermediate phenotypes, extreme transgressive phenotypes, or phenotypes with novel parental trait combinations (Rieseberg et al. 1999) and hybrids often show increased productivity and invasiveness (Ellstrand and Schierenbeck 2000; Paterson 2005). Many polyploids also show gene expression patterns that most closely resemble a single parental species (known as expression dominance), but can adjust the parental expression bias with environmental stress, suggesting polyploids may exploit pre-existing parental environmental adaptation (Bardil et al. 2011; Yoo et al. 2013). The variation that allopolyploids gain through hybridization and whole genome duplication then likely offsets any bottlenecks associated with speciation and may increase their potential for adaptation from standing variation.

In addition to their increased genetic variation, polyploid genomes may have other properties that facilitate adaptation. Dominant and partially dominant mutations can fix more quickly in polyploids than in diploids, especially in small populations (Otto and Whitton 2000; Anderson et al. 2004; Otto 2007; McGrath and Lynch 2012). Alleles may fix more easily in polyploids if whole genome duplication also releases pleiotropic constraints that may impede adaptation (Comai 2005; Otto 2007). Additionally, duplicated genomes harbour more variation as they have more alleles at a given locus and are more likely to acquire or have beneficial alleles when the environment changes (Otto and Whitton 2000). They may also conserve more slightly deleterious mutations that could later be beneficial, as they can buffer the effect of new deleterious mutations, especially at the haploid stage (Comai 2005). Thus, polyploidy may promote adaptation and spread by both maintaining and producing high levels of genetic variation and fixing adaptive variants more easily. However, the ecological relevance of these processes in polyploid adaptation lacks empirical support.

Polyploid genomes also face unique challenges that may limit their long term survival and adaptation. The doubling of alleles at a given locus reduces selection efficacy in polyploids allowing recessive deleterious mutations to accumulate and causing recessive beneficial mutations to fix more slowly. The shifts to selfing, and strong population bottlenecks that can often accompany polyploidization would further reduce selection efficacy (Arrigo and Barker 2012; Madlung 2013). While some deleterious mutations may become beneficial after an
environmental change, this advantage is transient and they will ultimately contribute to a long
term increase in polyploid genetic load (Otto and Whitton 2000). However, as polyploids may
generate new beneficial mutations more quickly than diploids, the rate of adaptation in
polyploids is not always lower than in diploids and depends on the dominance coefficients of
new mutations, mating system, and population size (Otto and Whitton 2000). A high rate of
adaptation may be important to the survival of new polyploids as they will often have to compete
with their progenitors (Arrigo and Barker 2012). Furthermore, new auto and allopolyploids have
to resolve complications that arise in meiosis to prevent to formation of aneuploidy gametes, and
combine with other unreduced gametes to prevent the formation of low fertility triploids and
pentaploids (Comai 2005). These challenges may contribute to low survival and higher rates of
extinction in neopolyploids than diploids, although the overall evolutionary success of polyploids
remains contentious (Soltis et al. 2014a; Tank et al. 2015). Better defining the role of adaptation
in polyploid evolution may then help clarify when polyploids are likely to be successful and
explain why some polyploid lineages are so successful while most are likely to go extinct.

While whole genome duplication may facilitate adaptation, it is unclear how strongly
novel adaptation contributes to polyploid spread. Immediate phenotypic changes associated with
whole genome duplication, rather than genetic adaptation, may be most important to polyploid
success. Changes in cell size, body size, water use efficiency, and growth rates following
duplication (Otto and Whitton 2000; Levin 2002; Otto 2007) as well as increased selfing rates, a
generalist life history, and increased plasticity in polyploids are likely important to polyploid
spread (Otto and Whitton 2000; Soltis and Soltis 2000; Barringer 2007; Husband et al. 2008;
Soltis et al. 2010; Hahn et al. 2012). Polyploid genomes may also elevate mean population
fitness and colonizing ability though masking recessive deleterious mutations, heterosis in
allopolyploids (Comai 2005; Prentis et al. 2008; Beest et al. 2012; Birchler 2012), and by
introgressing pre-adapted alleles (Treier et al. 2009; Beest et al. 2012). The few studies have
elucidated the immediate effects of polyploidization from adaptive evolution have conflicting
results and as these effects vary by species and environment, their general importance relative to
adaptation is not clear (Maherali et al. 2009; Soltis et al. 2010; Ramsey 2011).

Here we integrate genomic, phenotypic, and environmental data to characterize
the prevalence and genetic basis of adaptive evolution in one of the world’s most successful
polyploids. *Capsella bursa-pastoris* is an allotetraploid that originated in Eurasia and recently
spread worldwide with anthropogenic European colonization. *C. bursa-pastoris*’ global range contrasts with its diploid progenitors’ (*C. grandiflora* and *C. orientalis*) strongly limited European ranges (Hurka and Neuffer 1997; Hurka et al. 2003). Previous genetic and common garden studies suggest genetic adaptation, plasticity, and preadaptation all contribute to *C. bursa-pastoris*’ spread throughout Europe and North America (Hurka et al. 2003). I utilize *Capsella bursa-pastoris* samples from Europe, the United States, and Asia to address the following questions:

1. Is there evidence that post polyploidization adaptation contributes to *C. bursa-pastoris*’ spread? What is its adaptive history?

2. How do allopolyploid subgenomes with differing features and past evolutionary history influence polyploid adaptation?

3. What role do preadapted alleles, standing variation, new mutations and pleiotropy play in polyploid success?

### 1.4 Study system

*Capsella bursa-pastoris* is a self-fertilizing cosmopolitan weed found in diverse habitats throughout the world (Hurka et al. 2003). It tolerates a wide range of climates, occurring from the Egyptian desert to past the arctic-circle in Norway, and at altitudes ranging from sea level to 5900m in the Himalayas (Aksoy et al. 1998; Hurka et al. 2003). While it can withstand desert drought and moderate winters, it cannot survive in tropical climates or water-logged soil, and thus prefers temperate, steppe, Mediterranean, and subtropical climates (Hurka and Neuffer 1997; Aksoy et al. 1998). It also favours open habitats with shallow, rocky soils, and throughout its range it often occurs in ruderal or disturbed habitats such as wastelands or along path sides (Aksoy et al. 1998). It can also grow as a weed in crop fields and grasslands, although these plants will often outcompete it (Aksoy et al. 1998). Its selfing ability and short life cycle likely also helps its colonization success, allowing it to usually complete two life cycles per year with germination peaking in May and August, although winter annual varieties also exist in northern latitudes (Aksoy et al. 1998).
Despite its distinctly weedy lifestyle, both classic common garden and next generation sequencing studies suggest that genetic differentiation contributes to its broad environmental tolerance. It shows clear altitudinal and latitudinal clines in flowering time in Europe and China, which were recently linked to differences in flowering time candidate gene allele frequencies and expression patterns (Slotte et al. 2007, 2009; Huang et al. 2012). Common garden experiments in Europe also support genetic differentiation in flowering time, as well as in plant height, rosette size, and branching, although they do not identify the genetic basis of adaptation in these traits and also show plasticity for these features (Neuffer and Hurka 1986a; b; Neuffer et al. 1989; Neuffer 1990; Neuffer and Hoffrogge 2000). In contrast, North-American *C. bursa-pastoris* shows little novel ecotypic differentiation and instead suggest pre-adapted genotypes from Europe sorted themselves into appropriate environments (Neuffer and Hurka 1999). Further work on the genetic basis of adaptation in *C. bursa-pastoris* is then needed to clarify the importance and prevalence of local adaptation in this species.
2 Study

2.1 Introduction

Adaptation is often essential to range expansion and invasiveness (Lee, 2002; Sexton, McIntyre, Angert, & Rice, 2009; Fisher 1930). While the causes of range limits are not fully understood and vary widely by species (Gould et al. 2013), both theoretical and empirical studies show that a lack of genetic variation for key traits in peripheral populations that can arise through small population size, drift, inbreeding, and Allee effects can limit adaptation at range edges and species spread (Hoffman and Blows 1994; Sakai et al. 2001; Lee 2002; Griffith and Watson 2006; Bridle and Vines 2007; Gaston 2009, Eckert et al. 2008). Similarly, a lack of genetic variation for particular trait combinations and pleiotropic fitness tradeoffs can further constrain adaptation and expansion (Blows and Hoffman 2005; Colautti et al. 2010). However, multiple introductions and limited gene flow from divergent populations can increase genetic variation and facilitate species’ spread (Kirkpatrick et al. 1997; Weber and Schmid 1998; Sakai et al. 2001; Maron et al. 2004; Bridle and Vines 2007; Leger and Rice 2007; Kooyers and Olsen 2012). Standing genetic variation, rather than new mutation, is especially important to rapid adaptation during range expansion as species with standing variation are more likely to harbor a beneficial allele when they encounter a new environment (Barrett and Schluter 2008; Prentis et al. 2008). However, range spread can also occur without adaptation through preadapted genotypes and phenotypic plasticity, although these mechanisms do not necessarily exclude adaptation (Sakai et al. 2001; Ghalambor et al. 2007; Treier et al. 2009; Richards et al. 2009). As interactions among evolutionary forces, demographic history, environmental conditions, and the genome will shape the genetic architecture of adaptation in the genome, examining the genetic basis of adaptation may then clarify the evolutionary mechanisms of adaptation (Savolainen et al. 2013).

Adaptation also involves interactions between evolutionary forces and the genome and connections between the genotype, phenotype, and selective pressures are not well understood (Ehrenreich and Purugganan 2006). Theory predicts evolutionary process driving adaptation can shape the genetic architecture of adaptive traits (Savolainen et al. 2013). The magnitude of migration relative to selection can shift the effect distribution of adaptive loci with high
migration resulting in fewer loci of large effect and trade-offs in loci between demes rather than an exponential distribution of effect sizes, although the general distribution of effect sizes is not well resolved (Yeaman and Whitlock 2011; Olson-Manning et al. 2012; Savolainen et al. 2013; Hendry 2013). Adaptation can also select for clustering among adaptive loci which may occur through chromosomal inversions, translocation events, or selection for reduced recombination but the general importance of these events to adaptation is still not well understood (Nosil et al. 2009; Yeaman 2013; Savolainen et al. 2013). The preexisting genetic architecture of traits can also constrain adaptation through pleiotropy, often producing convergent adaptation in distinct populations (Stern and Orgogozo 2008; Stern 2013). Theory predicts adaptive alleles at protein coding genes with pleiotropic effects will have a drastically slower time to fixation and adaptation should occur more commonly through cis regulatory changes with fewer pleiotropic effects (Hansen 2006; Stern and Orgogozo 2008; Olson-Manning et al. 2012). Empirical studies challenge these predictions as most genomic adaptation studies find enrichment for protein coding genes, although few objective genomic studies address this question (Hoekstra and Coyne 2007; Hancock et al. 2011; Fournier-Level et al. 2011; Jones et al. 2012; Hendry 2013). Further empirical work investigating the genetic basis of adaptation is needed to test these theoretical predictions and clarify how selective forces interact with the genome to shape genetic architecture.

Most genomic studies of local adaptation to date use diploid organisms and little is known about adaptation in polyploids despite their unique genomic features, prevalence, and economical importance (Krasileva et al. 2013; Pyhäjärvi et al. 2013). Polyploidy is extremely common among angiosperms, crop plants, invasives, and some bony fish and likely affects adaptive processes (Levin 2002; Meyers and Levin 2006; Prentis et al. 2008; Pandit et al. 2011). Polyploid genomes can increase adaptive potential by creating and preserving duplicate genes, promoting large scale genomic rearrangements and changes in gene expression, and generating and conserving more variation (Otto and Whitton 2000; Soltis and Soltis 2000; Comai 2005; Otto 2007; Hegarty and Hiscock 2008; Finigan et al. 2012; Mcgrath and Lynch 2012). Genetic variance can also increase further through multiple origins and hybridization (in allopolyploids), likely offsetting any bottlenecks associated with speciation (Soltis and Soltis 2000; Weiss-Schneeweiss et al. 2013). Additionally, duplicated genomes may facilitate adaptation by reducing pleiotropic constraints and allowing dominant and partially dominant alleles to fix more
quickly, although these theoretical predictions are largely untested (Otto and Whitton 2000; Anderson et al. 2004; Otto 2007; Mcgrath and Lynch 2012). Polyploids may often rapidly adapt to persist alongside their diploid progenitors through niche differentiation, expand their range, and succeed in marginal habitats (Levin 2002; Brochmann et al. 2004; Ramsey 2011; Manzaneda et al. 2012; Husband et al. 2013). Polyploidy may thus have a large impact on adaptation and resolving adaptation in polyploids is essential to fully understand of the evolutionary mechanisms of adaptation.

*Capsella bursa-pastoris* is an ideal system to study the effects of polyploidy on adaptive evolution due to its allotetraploid genome, its recent known origins, and its tendency for ecotopic differentiation throughout its broad environmental and geographic range. Current evidence shows *C. bursa-pastoris* originated from an allopolyploidization event between diploid *C. grandiflora* and *C. orientalis* approximately 300 000 years ago (Douglas et al. 2014; Pannell 2015). It then spread throughout Eurasia with neolithic farming, and to North and South America, and now occupies a wide variety of environments within this range (Hurka and Neuffer 1997). Although its weedy features (eg. self-fertilization, plasticity, rapid annual life cycle) likely facilitated its spread, *C. bursa-pastoris* clearly shows genetic differentiation along latitudinal and altitudinal clines in Eurasia, suggesting local adaptation may have also contributed (Neuffer and Hurka 1986a; b; Neuffer et al. 1989; Slote et al. 2007, 2009; Neuffer 2011; Huang et al. 2012). Here, we investigate the genetic basis of local adaptation in *C. bursa-pastoris* throughout Eurasia. As demographic history influences adaptive patterns, we first examine population structure in our sample. We use these results to inform our analysis and interpretation of environment-genotype correlations that indicate likely adaptive events. We then further analyze patterns in putatively adaptive SNPs to investigate what features of *C. bursa-pastoris*’ polyploid genome may have influenced adaptation.
2.2 Methods

2.2.1 Genomic Dataset

We generated two main genetic datasets: 24 whole genome sequences and 261 reduced representation genotype-by-sequencing (GBS) genomes from throughout Eurasia (Figure 1). We sequenced 24 whole genomes from 10 countries with Illumina paired-end sequencing at Genome Quebec and aligned reads to the *C. rubella* reference genome using BWA and Stampy (Li and Durbin 2010; Lunter and Goodson 2011). We called SNPs using the GATK Unified Genotyper (Mckenna *et al.* 2010) as both diploid and tetraploid, but as the diploid SNP calls were generally higher quality and because there is no strong evidence of outcrossing (see below), we used the diploid SNP set for the remainder of the analysis. We only included SNPs with a quality score greater than 30 and per sample depth above 20 and below 149.62 (the 99th percentile). The final SNP set only included 21 samples, as mapping the remaining three samples was problematic. Our final SNP set is comprised of 7250911 variants at an average of 53.12 SNPs/kB of which 49.0% are fixed heterozygous sites representing fixed differences between the two homeologous genomes. We looked for evidence of outcrossing by computing cumulative heterozygosity across each chromosome with a custom Perl script, and did not identify any likely regions (Figure S1).

We sequenced the 261 samples from 65 populations and 18 countries using GBS sequencing with PstI at Cornell University. The 261 samples include 18 of those with whole genome sequences and 21 replicates within the GBS dataset. Populations have 1-6 individuals each, and each country has 1-31 populations. We filtered, mapped, and genotyped the sequences using the same pipeline as the whole genome data. We also compared our SNPs to genotypes generated through STACKS and found 77% agreement, but when we compared genotype calls within replicated samples, GATK SNPs showed lower within-replicate error (4% vs 9%). We estimated error by comparing GBS genotypes with the whole genome genotypes at common samples using a custom made Perl script, and subsequently removed SNPs with a depth below 35 and above 246 to optimize SNP number while controlling the genotyping error rate at 8.6%.
To account for selfing and allopolyploidy, we also removed fixed heterozygous sites (which we defined as sites that in the whole genome sequences were heterozygous at more than 75% of samples) as they are caused by fixed differences between the subgenomes and are not real polymorphisms. We also coded all unfixed heterozygous genotypes as homozygous alternate to better represent SNP frequencies in an allopolyploid (eg so that a fixed polymorphism on one genome would be at 100% frequency instead of 50%) and coded all homozygous alternate genotypes as missing data since in a selfer homozygous alternate sites should be rare and likely represent error. We also removed sites with over 90% missing data and invariant sites, leading to a final SNP set of 3801 SNPs with an average of 1.1% missing genotypes per site. Finally, we estimated the frequency of potential homeolog dropout in the GBS SNPs by calculating the rate that heterozygous genotypes from the whole genome dataset were called as homozygous in the GBS dataset and found dropout only occurred at an average of 2.2% of genotypes.

2.2.2 Phasing

We phased the whole genomes using GATK’s read-backed phasing, estimated phasing accuracy, and then used custom scripts to group phased sequences into the *C. grandiflora* and *C. orientalis* subgenomes. Read-backed phasing successfully phased over 96% of sites in each sample and produced haplotypes with a mean length of 17.35 SNPs. We estimated phasing accuracy by comparing the phased haplotypes to the most similar subgenome sequence at 11 Sanger sequenced loci (T. Slotte unpublished data) and by estimating the rate at which the assigned subgenome changed within a haplotype. Sanger validation showed we correctly phased 85.5% of SNPs from sequence validation, and haplotypes were on average 80.4% consistent. After assuring phasing quality, we then used a custom Perl script to sort phased haplotypes into the subgenomes. We identified continuously phased haplotypes for each sample and then used the number of sites that corresponded to each parental allele at fixed differences in 13 *C. grandiflora*, and 10 *C. orientalis* whole genome samples (genotyped by R. Williamson and G. Douglas, see Williamson and Josephs et al. 2014, and Douglas et al. 2014 for details) to assign each haplotype a subgenome. We assigned each haplotype to the subgenome with which it shared the most parental alleles, and ended haplotypes at any point where the phased alleles corresponded to a different subgenome than the preceding sequence. We coded sites on haplotypes that could not be assigned to a subgenome as missing data. We also removed phased sites with fewer than 10 reads per sample, and sites where there were less than 4 reads supporting
the assigned genotype. We then computed pairwise $r^2$ within and between the phased genomes along chromosome 8 using a custom Perl script and found significantly higher LD among SNPs within the same subgenome than with the other subgenome for both subgenomes in both European and Asian samples (Table S1). The resulting phased genomes have variable proportions of missing data per sample, but there are an average of 6 phased SNPs/kB without any missing data on both subgenomes.

We then used these phased whole genome sequences to phase the GBS data with a custom Perl script. Much like the approach used with the whole genomes, we assigned GBS reads to a subgenome based on their overlap with fixed differences in the phased subgenomes (Figure S2). We first found any fixed differences between the phased subgenomes that occurred within the length of the reads carrying a heterozygous SNP in GBS data. If the reads for a sample carrying each allele overlapped at least one fixed difference, and for each GBS read type there were at least 5 more and 1.5 times as many reads carrying the allele at a fixed difference for one subgenome than the other subgenome, we assigned the allele to that subgenome at that site. We then tallied how often each allele at the site being phased occurred with the alleles at fixed differences from each subgenome, and assigned each allele to the subgenome that contained the alleles it segregated with at most fixed differences. We left homozygous sites intact in both phased GBS subgenomes, and coded heterozygous sites that we could not phase as missing data. To assess phasing accuracy, we compared heterozygosity in the phased GBS datasets. As expected with correct phasing, based on the phased data we found significantly higher $\theta_\pi$ in the *C. grandiflora* than the *C. orientalis* subgenome ($t=1.99, P=0.046, df=10715.99$).

### 2.2.3 Demographic analysis

We used a variety of methods to investigate *C. bursa-pastoris*’ demographic history in Eurasia using both the phased and unphased datasets. We first analyzed the unphased data. We carried out PCA in Eigenstrat (Alexander et al. 2009) using unphased GBS SNPs from all samples, but also repeated PCA separately for Europe and Asia. We then verified these results by running STRUCTURE (Pritchard et al. 2000) on the same dataset for $K=1$ to $K=5$ with 20 replicates for each $K$, and determined the most probably $K$ using the Evanno method (Evanno et al. 2005), as implemented in StructureHarvester (Earl and vonHoldt 2011). We also ran STRUCTURE in the same way only in European and Asian samples for $K=1$ to $K=10$. We also
used the tree based algorithm Treemix (Pickrell and Pritchard 2012) on the same GBS dataset with the Greek population as the outgroup to assess population relationships and infer migration events. We chose Greece as the outgroup as it was the population with the highest mean pairwise Fst in the dataset. We tested models with m=1 to m=5 migrations to try and minimize residual variance in the model. Finally, we also repeated STRUCTURE and PCA analysis within China to examine regional population structure.

We then repeated PCA, STRUCTURE, Treemix, and IBD analysis using the phased GBS SNPs along with parental samples to try to elucidate the subgenome origins. We carried out PCA on the subgenomes with parental samples using only 536 and 98 SNPs without missing data in the C. orientalis and C. grandiflora sets, respectively, in R. We conducted STRUCTURE and Treemix analysis in the same way as we did for the unphased dataset, but did not run regional analyses. Additionally, we also included the genomic reference C. rubella sample (Slotte et al. 2013) in the C. grandiflora subgenome Treemix analysis to look to evidence of previously suggested introgression (Slotte 2007). Finally, we tested for regional isolation by distance (IBD) by conducting IBD analysis separately in Asia, Western Europe, and the Mediterranean. We used SpaGeDi (Hardy and Vekemans 2002) to compute pairwise Fst among populations in each region, and tested the relationship between linearized Fst (Fst/1-Fst) and Euclidean, earth adjusted, geographic distance with a Mantel test in vegan.

To try to identify a possible area of origin for C. bursa-pastoris in Eurasia, we used a similar approach as Ramachandran et al (2005) used to identify Africa as the most probable human origin and examined points of origin whose distance to our samples explained the most variation in heterozygosity. We computed the distance from 1000 random points throughout Europe and Asia to the locations our samples and used a linear model to test the relationship between distance from each origin, and expected silent site heterozygosity. We conducted the analysis separately on each subgenome, and only included European samples because of weak population structure and isolation by distance in Asia. As, unlike in the human studies, our regressions only explained a small fraction of heterozygosity, we regarded areas with the highest r² values as hypotheses for possible origin areas, rather than strong evidence for a given geographic origin.
We next investigated regional patterns of linkage disequilibrium in the phased whole genome dataset. We estimated pairwise $r^2$ in Europe and Asia separately using all SNPs on chromosome 1 in PLINK. To improve LD accuracy, we excluded the centromere region and eliminated all singletons. We then used GERMLINE to identify shared haplotypes among subgenome samples. We looked for shared haplotypes in the phased subgenomes within the complete dataset at sites without missing data. We constructed the necessary genetic map by interpolating points from the *C. grandiflora/ C. rubella* F2 linkage map made by K. Hazzouri (Slotte et al. 2012) using a sixth degree polynomial. We used the default parameters except we elevated the maximum number of mismatching homozygous markers in a haplotype to 9 to account for possible phasing error. To analyze patterns of haplotype sharing among our samples, we tested for non-independence in the number of shared haplotypes found within Europe and Asia and within each subgenome with a chi-squared contingency test. We then used a chi-square test to compare the number of shared haplotypes in the subgenomes and Wilcoxon rank sum tests to compare the length of haplotypes in Europe and Asia within each subgenome. Finally, we compared the length of shared haplotypes among samples in Europe, Asia, and the Mediterranean for each subgenome with non-parametric Kruskal Wallis tests.

We also compared the subgenome minor allele frequency spectra between regions and subgenomes. We calculated the minor allele frequencies of whole genome SNPs with less than 40% missing data for each subgenome in Europe and Asia. We tested for differences between the minor allele frequency distributions between subgenomes in a given region, and between regions within each subgenomes using Kolmogorov-Smirnov tests, and also compared the distribution of replacement and synonymous polymorphisms using 1000 bootstrap samples for 100kb windows from the genome. We also tallied the number of private and shared SNPs within the whole phased genomes in each region and with the parental samples, and computed diversity statistics using a modified version of Polymorphurama.

2.2.4 Environmental Associations

We used a trimmed SNP dataset and publicly available climate data to test for environmental associations. We tested environmental associations on the unphased GBS SNPs, but removed all singletons. We also removed SNPs that were in $>90\%$ LD within 10 SNP windows identified using PLINK to reduce multiple testing and to avoid detecting associations
from the same locus through multiple tightly linked SNPs. After trimming and removing singletons, 1567 and 1513 SNPs remained in Asia, and Europe respectively. We compared our SNP data to publicly available climate data at 30 arc seconds (~1km²) from the WORLDCLIM database (69 monthly precipitation, temperature, and BioClim variables) as well as an aridity index from the FAO Geonetwork. As many environmental variables were highly correlated and PCA did not yield easily interpretable axes, we reduced collinearity among the environmental variables by eliminating variables correlated at r>0.8 to obtain a reduced set of 18 environmental variables including aridity, altitude and most BioClim variables. We interpolated climate data for the GPS coordinates of each population using ArcGIS.

We first tested whether populations from more similar environments were more genetically similar by comparing genetic distance to climate distance. We summarized distance in climate by computing Gower’s distance among all populations using the 18 climate variables. We then tested for correlations between climate distance and Fst using a Mantel test with 1000 permutations in vegan.

We then used both an individual and population based approach to detect SNP-environment associations, and conducted both analyses separately in Europe and Asia. We used Bayenv2 to detect population level environmental associations, and highly differentiated loci. We calculated the population covariance matrix using all SNPs instead of silent sites as we had too few silent sites to produce consistent covariance matrices among iterations. We ran Bayenv separately in each region with 100000 iterations and also computed the XᵀX population differentiation parameter. We included all populations and all environmental variables, but removed any replicate samples within populations. Bayenv only produces Bayes Factors that reflect the strength of associations rather than p-values, so to identify meaningful SNP-variable associations we ranked SNPs by BayesFactor and used an outlier approach. As recommended in Hancock et al. (2010), we separated SNPs into allele frequency bins and took SNPs with the 5% highest BayesFactors in the two highest allele frequency bins for each variable. We also identified XᵀX outliers as SNPs in the two highest allele frequency bins with the 5% highest XᵀX value.

Due to our limited within-population sample size in certain populations, we also tested for SNP-environment associations with the individual-based Latent Factor Mixed Model
approach (Frichot et al. 2013). The LFMM approach is similar to the mixed model regression approach often used in GWAS and in landscape genomics as it also uses a regression to test how well environmental variables predict variance in genotypes while controlling for population structure at each SNP. It implements a model with the genotype matrix as a response variable, a fixed environmental variable term, but unlike mixed models it controls for background population structure using a K-dimensional latent factor matrix instead of a random effect kinship matrix. The resulting model shows higher power than traditional mixed model approaches, and a higher rate of true positives to false positives than Bayenv (Frichot et al. 2013). We ran LFMM with the trimmed SNP dataset and the reduced set of environmental variables separately in both Europe and Asia. As recommended in Frichot et al. 2013, we used the number of significant eigenvalues computed through Tracy-Widom statistics in Eigenstrat as the dimensions in the latent factor matrix. We then used q-values to correct for multiple testing through a 5% FDR cutoff (Frichot et al. 2013).

We then compared how well the candidate SNPs that each method identified explained variance in their associated variables. For each of the 18 reduced variables, we computed the variance explained by the major PCA axes summarizing variance in the associated SNPs. We conducted PCA on a matrix of genotypes for SNPs associated with the variable of interest, and used all PCA axes with an eigenvalue above the mean eigenvalue. These axes acted as explanatory variables in a linear model with the environmental variable as the response. We then qualitatively compared the range of adjusted $r^2$ among models.

After noting that LFMM identified notably more effective SNPs than Bayenv, we further explored patterns in LFMM SNP associations. We coded SNP associations in a binary SNP by variable matrix and analyzed patterns in variable associations through CCA in vegan. We then tested if Europe and Asia shared more putatively adaptive SNPs than expected by chance by comparing the number of SNPs with significant associations from our dataset in both regions to the distribution of shared SNPs with significant associations from running LFMM with 500 permutations of the environment data matrix in each region. We then examined if genes that contained SNPs with significant associations in both Europe and Asia generally showed associations to similar sets of variables, by computing the proportion of common variables associated with SNPs in shared genes. As variables are correlated and the same association could occur through different, correlated variables, we also classified variables as temperature or
precipitation related and computed the proportion of SNPs within shared genes that were associated with variables in each group in each region. To test if SNPs in shared genes had associations to the same type of variable in each region, we computed Spearman’s correlations for the proportions of associations in each variable type between each region. Finally, we tested if candidate SNPs had an excess of multiple or single associations by comparing the number of significant associations per SNP to the number of significant associations per SNP from running LFMM with 500 permutations of the environmental matrix in each region.

2.2.5 Enrichment analyses

Through the annotated *C. rubella* reference genome, we tested for site type enrichment sites as well as for conserved non-coding site enrichment. We obtained annotations for SNPs with putative associations form the *C. rubella* genome annotation using SnpEff (Cingolani et al. 2007). We tested for enrichment in annotations by comparing the number of putatively adaptive SNPs with a given annotation to the number of SNPs with that annotation in 1000 randomly drawn SNP sets the same size as the test SNP set. We considered an annotation significantly enriched or reduced if the number of test SNPs with that annotation was in the 2.5% tails of the random distribution. We test for enrichment in all major SnpEff annotations and also tested for coding and noncoding enrichment. We also used the same approach to test whether non-coding SNPs were closer to the nearest transcription start site than expected by chance. To control for the possible influence of minor allele frequency on site types without losing too much power, we repeated the enrichment analysis only comparing SNPs with MAF<0.15. We tested for enrichment in conserved non-coding sites in a similar way. If a SNP occurred within 500bp of a CNC we identified it as linked to that CNC. We then compared the number of SNPs linked to CNC’s in the significant SNP sets to the number of CNC associated SNPs in 1000 randomly drawn SNP sets of the same size as the test set. SNP sets were enriched for CNCs if they had more CNC linked sites than 95% of random draws.

In addition to site enrichment analysis, we also tested for functional enrichment in genes containing SNPs with putative associations. We only used sites that showed significant environmental associations through both Bayenv and LFMM for functional enrichment analysis as they likely represent the most reliable associations. We then identified any *C. rubella* genes that contained putatively adaptive SNPS within their coding region or introns and found their
BLAST best match in the *A. thaliana* genome. We repeated this protocol with the trimmed GBS SNP sets for each region to provide a background set of genes. Through the DAVID online tool (Huang *et al.* 2007), we then tested for GO functional enrichment in genes containing putatively adaptive SNPs compared to the respective background for each region. Additionally, we searched for flowering time candidate genes listed in Huang *et al.* 2012 among the putatively adaptive *A. thaliana* orthologs.

2.2.6 SNP origins

We examined how putatively adaptive SNPs segregated within the phased and parental genomes as well as geographically to better understand their possible origins. We noted if each putatively adaptive SNP in the LFMM, Bayenv, and common SNP sets segregated in either or both of the phased subgenomes, and tested for differences in the number of SNPs segregating in each subgenome with a chi-squared test. We then repeated this procedure with parental genomes, tallying SNP segregation in the 13 *C. grandiflora* and 10 *C. orientalis* samples. We also tested whether SNPs segregated in larger or smaller geographic areas than expected by chance within Europe and Asia. We computed the average pairwise geographic distance among samples containing the non-reference allele at a given SNP, and compared it to the 1000 samples of the mean geographic distances among the same number of random populations within that region.

2.2.7 Sweep scans and proximity analysis

We obtained a variety of different selection metrics to detect sweeps with different ages and characteristics. We conducted all sweep scans on phased whole subgenome SNPs without any missing data, and ran them separately in European and Asian samples, using the same inferred genetic map that we constructed for GERMLINE analysis where required. We coded the non-reference allele as derived in all analyses. We used CLR and XPCLR methods to detect older hard sweep signatures. We performed CLR with SweepFinder (Nielsen 2005) at every SNP and every 100 SNPs and identified significant sweep signatures as CLR values above the 95th percentile of CLR values from simulated neutral sequence. We simulated 500 neutrally evolving 250k chromosome datasets the same size as our dataset based on a demographic model inferred through *fastsimcoal2* from our phased whole genomes with Macs (Chen *et al.* 2009). We also used XP-CLR to detect differentiated SNPs indicative of hard sweep signatures in northern and southern Asia. We divided populations in Asia into two approximately even groups by latitude
and ran XP-CLR with 5 cM sliding windows with up to 10 SNPs per window. To identify recent and soft sweep candidates, we also used the same methods to detect iHS and nSL outliers. As the non-reference allele may not always be derived, we considered the absolute values of the statistics for outlier analysis. Additionally, for iHS we also separated the genomes into fixed 100 kB windows, and identified windows with the 95% highest proportion of iHS values >2 as recommended in (Voight et al 2006). Voight et al. suggest iHS>2 are likely sweep candidates, and this cutoff also coincided closely with the 95th percentile of iHS values. We tested for differences in the number of sweep candidates segregating in each subgenome for each metric using a chi-square test.

To test whether sweep outliers on the *Capsella bursa-pastoris* genomes overlapped with sweep outliers on the parental genomes we conducted the same sweep scans on the parental sample genomes. We first phased *C. grandiflora* samples (Williamson et al. 2014) with BEAGLE and assumed that *C. orientalis* samples were fully homozygous and did not require phasing. We then used CLR and nSL scans to identify sweep outliers on our samples. We conducted 100 simulations of the corresponding number of 250k chromosome for both *C. grandiflora* and *C. orientalis* using Macs based on the demographic parameters inferred in Douglas et al 2014. We identified significant CLR SNPs the same way as above, but as only one nSL value was significant when compared to simulations, we used a 1% cutoff to identify outliers. We tested whether 100kB windows between these and *C. bursa-pastoris*, overlapped more than expected by chance by randomly drawing the respective number of significant windows from each genome and counting how many out of 1000 draws showed as much or greater overlap. We also used this approach to test whether the number of windows containing CLR outliers in both Europe and Asia overlapped more than expected by chance.

We then tested whether selection sweep outliers occurred closer to putatively adaptive SNPs identified through environmental association analyses than expected by chance. We accomplished this through two complimentary approaches. We first compared the mean distance from potential sweep signatures to the nearest putatively adaptive SNP to the distribution of same statistic measured 10000 times in random SNP sets of the same size and from the same background (eg GBS or whole genome) as the datasets being tested. We also assessed the distances among the various selective sweep outliers in the same way. Additionally, we also computed the distance from each putatively adaptive SNP to selective sweep outliers on the same
chromosome and examined the distribution of distances from putatively adaptive SNPs to sweep outliers. We obtained 95% bootstrap confidence intervals by repeating this procedure with 500 bootstrap samples of each putatively adaptive SNP set. We carried out these proximity tests with significant SNPs identified through Bayenv, LFMM, X^T X, and both Bayenv and LFMM in Europe and Asia, and all sweep scan outliers. We also used these proximity tests to compare the distance among SNPs within putatively adaptive SNP sets to test for clustering.

We complemented sweep scan analyses with tests to examine whether, overall, putatively adaptive SNPs showed characteristics consistent with selection. We calculated $\theta_\Pi$, $\theta_w$, and Tajima’s D in 1000 SNP windows around each putatively adaptive SNP identified through both Bayenv and LFMM, and compared them to diversity values in windows of the same size around 10000 randomly drawn SNPs from the trimmed GBS data. We tested whether diversity around SNPs was significantly reduced or elevated by comparing it to the 95th percentile value for a given metric from randomly drawn SNPs. We compared replacement to synonymous $\theta_\Pi$ the same way. We also compared the distribution of LFMM putatively adaptive SNP’s minor allele frequency (MAF) to SNPs with significant associations in LFMM from 500 permutations of the environmental variable matrix. We compared the proportion of putatively adaptive and random SNPs with MAF in each of 10 equal size MAF bins. Putatively adaptive SNPs were enriched for MAF range if they had more SNPs within that bin than 95% of the random samples.

2.2.8 Linkage disequilibrium and effect size

We next examined how patterns of linkage and effect size in putatively adaptive SNPs differed among regions. We compared mean pairwise $r^2$ among putatively adaptive SNPs identified through LFMM, and both LFMM and Bayenv in Europe and Asia to the distribution of mean pairwise $r^2$ from 1000 random SNP draws of the same size as the test set from the corresponding trimmed GBS dataset. We computed LD both only among SNPs on the same scaffold, and among all SNPs. We then used a Wilcoxon rank sum test to compare the mean Z-score among significant SNPs identified through LFMM and both LFMM and Bayenv. Z-scores are standardized regression coefficients of the fixed environmental effects from the mixed model and reflect the strength of the association between the environmental predictor and the genetic response variables (Frichot et al. 2013).
2.2.9 Common garden GWAS

As a complementary approach to identifying putatively adaptive SNPs, we grew a sample of our genotypes in a common garden, measured phenotypic traits, and tested for genotype associations with ecologically important traits. To maximize genetic diversity in the common garden, we estimated a genetic similarity matrix based on identity by state in all GBS SNPs and eliminated any lines that were >97% identical. We then germinated greenhouse selfed seed from the 160 remaining lines on agar media in May 2014 and refrigerated them for 1 week. We then transferred the seedlings to soil cells and placed them in a greenhouse at 21°C for two weeks.

Once the seedlings had several true leaves but had not yet flowered, we transferred them to five common garden beds on an exposed University of Toronto rooftop. We replicated each genotype in five blocks, with one block per planting bed. Each planting bed was 13’ x 5’ and contained 5 cm of Promix soil. To encourage establishment, we thoroughly wet the soil before planting and covered the beds with shade cloth for two days after planting. However, aside from this initial support, we did not provide further water, shade, or nutrients throughout the course of the experiment. While plants were growing, we recorded bolting time, flowering time, leaf number, and rosette diameter. We checked for bolting and flowering approximately once every two days, and recorded leaf number and rosette diameter at bolting. We recorded rosette diameter at the largest point for every rosette. We continued monitoring the experiment until all plants had senesced, and harvested all samples in September of 2014. We subsequently counted fruits on each plant to estimate fitness, and also measures plant height, branch number, and dry biomass.

We analyzed the resulting phenotypic data using linear mixed models. We first tested for block, population, and line effects using mixed models through the lme package in R. We constructed models for each trait with block as a fixed effect, and line nested within population (defined by site of collection) as a random effect. We tested the significance of terms by comparing the -2 times difference in the log-likelihood of models with and without the term of interest to a chi-square distribution with 1 degree of freedom. We subsequently carried out GWAS analysis to look for variants associated with our phenotypes. We carried out GWAS using the GAPIT Package in R, and conducted the analysis with all samples together, as well as separately in Europe and Asia. In both region analyses we found a kinship matrix computed
through the EMMA algorithm best first the data, while for the joint analysis we computed the kinship matrix through the VanRaden method. In all analyses we allowed GAPIT to optimize the kinship grouping and summarizing parameters to increase power. For the European and full datasets we included 3 PCA axes in the Q matrix, while for Asia we included 4. We assessed significance using a 5% FDR cutoff.

In addition to GWAS, we also tested for phenotypic correlations with environmental variables. To further reduce multicollinearity, we first standardized our variables and refined the set of predictors to 7 and 10 environmental variables in Asia and Europe, respectively, with VIF<7. We included these variables as fixed effects in the mixed linear model described above and used backwards stepwise selection based on model significance through the lmerTest package to identify variables with the strongest associations. We then tested the significance of the remaining variables using a likelihood ratio test as above.

2.3 Results

2.3.1 Population genetics

Throughout all analyses, we observed populations separated into well-defined European and Asian clusters, along with a subtler Middle-Eastern grouping. Principle components analyses of both the full dataset and the subgenomes clearly showed these groupings, and also separated the Greek and Italian populations from the Western Europe cluster and closer to the Middle East Cluster (Figure 2). Treemix and STRUCTURE analyses for the unphased dataset also reflected this large-scale population structure, and populations clearly separated into regional branches and clusters, respectively (Figures 3,4). Furthermore, STRUCTURE Harvester analyses for both the unphased and phased data show two clusters, representing Europe and Asia, are most likely. Similarly, STRUCTURE analyses for European samples alone also identifies the division between Western European and Mediterranean samples, reflecting that regional groupings dominate population structure in our dataset (Figure 5). Treemix analysis on the unphased dataset supports the integrity of these groupings with little evidence of major migrations among them. We only found evidence of two main migration events between Italy and Greece, and
Greece and Bosnia (Figure 4). These groupings remain largely unaltered for all analyses of the phased genomes with the parental samples.

Through these analyses we also observed population groupings within regional clusters that may provide insight into the species’ colonization history. Consistently, the Russian populations grouped with the Western European cluster, while samples from the United States grouped with the Middle-Eastern cluster. Similarly, the Greek and Italian samples were often more similar to the Middle-Eastern than the Western European samples, and the Asian samples show little shared ancestry with samples from other regions. Although PCA, Treemix, and STRUCTURE show parental samples do not clearly group with any regional clusters (Figures 6-11), STRUCTURE plots suggest they are more closely related to the European than the Asian samples and the plots for K=4 suggest they may share some common ancestry with the Greek and Italian samples (Figure 7). We find no evidence that similarities among the parental samples and the Greek and Italian samples are explained by introgression as Treemix results do not identify any migration events among these populations (Figures 8,11). In contrast, we identify numerous migration events from *C. grandiﬂora* into the Western European and Middle Eastern branches, as well as from *C. rubella* into the Middle Eastern branch and one possible migration event from *C. orientalis* into the Western European branch. However, despite possible migration events among clusters, and disparities in geographical and population groupings, the three large regional clusters generally remain well defined.

In contrast, patterns of subgenome expected heterozygosity based on the GBS SNPs among clusters are largely similar and do not indicate a clear origin or colonization history. Silent site expected heterozygosity differs slightly but significantly for both subgenomes among Asia, Europe, and the Middle East (*C. gr*: P= 0.03, Kruskal Wallis $\chi^2$=7.03, df=2; *C.or*: P= 0.047, Kruskal Wallis $\chi^2$=6.1, df=2), with the highest median heterozygosity in Western Europe (Figure 12). Differences in diversity are more pronounced within the whole genome data, as both subgenomes show significantly higher $\theta_\pi$ and $\theta_w$ in Europe than Asia, and more negative Tajima’s D in Asia than Europe (Figure S3). Consistent with inheritance of more standing genetic variation from the ancestrally outcrossing *C. grandiﬂora* ancestor, this subgenome shows higher $\theta_\pi$ and $\theta_w$, as well as more positive Tajima’s D than the *C. orientalis* subgenome in Europe. However, the *C. orientalis* subgenome has slightly but significantly higher $\theta_\pi$ and $\theta_w$ in Asia (Figure S3). Similarly, only the western Europe *C. orientalis* genome shows significant
isolation by distance (Table 1), although the *C. grandiflora* Western Europe and both Middle-East clusters show marginal P-values. No region or subgenome showed significant isolation by environment when we compared genetic to environmental distance among populations. Random origin heterozygosity regressions generally show low adjusted $r^2$ values for both subgenomes (C. gr: -0.0014-0.15, C.or -0.014-0.18) but values increase when only considering European populations (C. g: -0.0028-0.21, C.or: -0.028-0.28). The most variation in heterozygosity is explained with distance from points in Eastern Europe between longitudes or 125 and 135 for both genomes (C. gr 0.196-0.22; C.or: 0.26-0.28), and maximum $r^2$ occurred when measuring distance from a point in the Khabarovsk Krai in South Eastern Russia for both genomes (Figure 13). Patterns of heterozygosity are then generally subtle in our sample and do not strongly differentiate clusters or clearly indicate species origin.

In contrast, differences in linkage disequilibrium and shared haplotype length among clusters are more pronounced. In both subgenomes, LD decays more rapidly in Asia than in Europe. LD decays at approximately 3kB in European populations and decays at approximately 1Kb in Asian samples for both subgenomes. Following LD decay, all groups then maintain $r^2$ near 0.25, substantially higher than the random expectation of 0.1, but show a notable portion of SNPs in complete LD (Figure 14). Contrary to expectations, LD decays more slowly and remains higher in the *C. grandiflora* than the *C. orientalis* subgenome. We also found more and longer shared haplotypes among the *C. grandiflora*, than the *C. orientalis* subgenomes regardless of region (P<0.01, $\chi^2$=853.24, df=1), (P<0.01,$W=39053$), respectively (Figure 15). Consistent with LD decay patterns, we observed longer shared haplotypes among European than Asian samples on both subgenomes (C. gr: P<0.01,$W=160399.5$; C. or: P=0.003, $W=7833.5$). We also observe longer shared haplotypes between Western Europe and the Middle East and in Asia and Western Europe than among samples in Asia and the Middle East in both subgenomes (Figure 16). Stronger LD and longer shared haplotypes in Europe than Asia may reflect a recent origin and expansion in Asia and longer establishment in Europe, combined with longer-term population subdivision.

Whole genome minor allele frequency (MAF) distributions are also consistent with more recent colonization in Asia than Europe. Both the *C. grandiflora* and *C. orientalis* subgenome showed significantly different MAF distributions in Europe and Asia (C. gr: D=0.33, P<0.0001, C.or: D=0.21, P<0.0001) with more low frequency mutations and fewer intermediate frequency
mutations segregating in Asia than Europe (Figure 17). The MAF distributions for the two subgenomes also differed significantly in a given region (Asia: D=0.14, P<0.0001, Europe: D=0.0325, P<0.0001), but the patterns are less consistent. In Europe, we observed more intermediate and fewer low frequency mutations in the *C. grandiflora* than the *C. orientalis* subgenome and observed the opposite pattern in Asia (Figure 17), consistent with each parental subgenome conserving more high frequency SNPs near the diploid progenitor’s range. The joint derived allele frequency distributions showed similar patterns (Figure 18). The *C. grandiflora* subgenome shows more intermediate frequency alleles in Europe that are fixed in Asia than the *C. orientalis* subgenome. However, the *C. orientalis* subgenome shows more intermediate and low frequency alleles segregating in both Europe and Asia than the *C. grandiflora* subgenome. It also contains more private alleles that are segregating in Asia than the *C. grandiflora* subgenome. Both these allele frequency distributions may then suggest different demographic processes influencing each subgenome in Europe and Asia.

Different evolutionary histories and demographic processes operating on each subgenome also led to differences in selection efficacy between the subgenomes. The *C. grandiflora* genome shows a significantly lower ratio of replacement to synonymous $\theta\pi$ in both Europe and Asia, consistent with more effective purifying selection (Figure 19). Similarly, the *C. grandiflora* subgenome shows a significant excess of low frequency SNPs and a significant lack of high frequency SNPs in Europe, although not Asia, further suggesting more efficient selection in the outcrossing than the selfing subgenome, at least in the ancestral species range (Figure S4). Low effective population size and recent colonization could contribute to overall weaker purifying selection in Asia, and subtler differences in selection efficacy between the two subgenomes.

After observing signatures of disparate demographic histories in the two subgenomes, we examined how SNPs segregated in the parental and tetraploid genomes to better characterize levels and origins of diversity. Shared and unique polymorphisms are also consistent with reduced diversity in Asia but also highlight differences between the two subgenomes. The Asian *C. grandiflora* and *C. orientalis* subgenomes both show fewer unique polymorphic sites than subgenomes in the European samples, reflecting an Asian bottleneck. However, we observe more unique SNPs in the *C. grandiflora* subgenome than the *C. orientalis* subgenome in both regions. The *C. grandiflora* European subgenome shows especially high diversity as it both
contains the most unique polymorphisms and shares the most polymorphisms with its parental species than any other subgenome (Table S2). In contrast, in both regions the C. orientalis subgenome shows few unique polymorphisms and few SNPs shared with the C. orientalis parental samples. Surprisingly, the C. orientalis subgenomes share more SNPs with the C. grandiflora parental samples than the C. orientalis parental samples, likely due to severely reduced diversity in C. orientalis. In Asia, the C. orientalis subgenome shares more SNPs with the C. grandiflora parent than the C. grandiflora subgenome. However, the C. orientalis subgenome shares more polymorphisms between Europe and Asia than the C. grandiflora subgenome.

Results from demographic inference based on the whole genomes supports an earlier origin in Europe followed by a recent expansion to Asia. In both genomes, and all models tested, Asia shows a substantially smaller effective population size than Europe and a recent split (Table S3), although growth rates were moderate in both Europe and Asia. Population size estimates were also similar in the two subgenomes for a given region, although they were consistently lower in the C. orientalis than C. grandiflora subgenome. For both genomes we used estimates from models without migration. This model showed the lowest AIC value for the C.orientalis subgenome. While a model with migration using the C. grandiflora subgenome had the lowest AIC value, results produced from models without migration in both subgenomes were more consistent with previous demographic history estimates (Cornille et al. 2015, submitted, Douglas at al 2014, Pannell et al. 2014) and seed fossils found in Pleiostocene sediments from the Ipswichian or Hoxne interglacials up to 245 000 years ago (Turner 2000).

After detecting strong regional population structure that could obstruct environmental association tests, we examined regional population structure in China to determine whether population structure was less defined in a recently colonized area. We identified some population structure within China, but it was much weaker than the population structure defining the major geographic clusters. STRUCTURE analysis shows regional grouping between populations approximately north and south of the Yangtze River (Figure S5), as well as the a third cluster dispersed throughout the country. Differences in environment may contribute to north and south groupings, as PCA and LDA based on environmental variables in Chinese populations shows populations in the north and south clusters are environmentally distinct (Figure S5, F=31.2,P<0.0001 ). However, genetic differentiation between the north and south populations is
weak in comparison to the strength of the worldwide regional groupings. In contrast to the strong regional clustering seen in Figure 1, north and south populations do not clearly group in PCA of the Chinese samples (Figure S5). STRUCTURE Harvester also identified two clusters without the north and south groupings as being most likely. Furthermore, Fst reflects weak differentiation between the two clusters (Fst=0.062), compared to overall Fst among the Asian samples among populations (Fst=0.21). These indications of subtler regional clustering then suggest China is less defined by population structure which may reduce the confounding effect of geography on environmental associations.

2.3.2 Environmental Associations

Sets of SNPs with significant environmental associations differed depending on the region and the method used, although we used the same 1567 and 1513 linkage trimmed GBS SNPs in Asia and Europe, respectively, for all analyses. Ranking using Bayenv BayesFactors, we identified 427 SNPs in Asia and 444 SNPs in Europe with the top 5% strongest environmental associations, while with LFMM we identified 281 and 742 SNPs with significant associations in Asia and Europe, respectively. In Asia, both methods identified 103 of the same SNPs, while in Europe the methods found 354 common SNPs. We also identified 79 and 76 $X^T X$ outliers in the top 5% of $X^T X$ values in Asia and Europe, respectively, of which over 86% showed significant associations in the LFMM analysis. PCA axes summarizing significant SNPs identified through LFMM explained more variance in their associated environmental variables than those identified through Bayenv in both Asia (LFMM: $r^2$ 0.46-0.89; Bayenv $r^2$ 0.049-0.23) and Europe (LFMM: $r^2$ 0.67-0.94; Bayenv $r^2$ 0.09-0.74). We emphasize candidate SNPs identified through LFMM, or identified through both analyses throughout the remainder of our analysis.

We found similar patterns of environmental associations in Europe and Asia. Both regions showed more associations with temperature than precipitation variables, although region and variable type had non-independent effects with an excess of environmental associations related to temperature in Europe, and an excess of associations linked to precipitation variables in Asia (P<0.01, $\chi^2$=24.8, df=1, Figure 20). Additionally, both regions showed multiple variable associations among SNPs. We found enrichment for >2 variable associations per SNP in Europe (P<0.001) and enrichment for 4 and >6 associations in Asia (P<0.05). Additionally, CCA
suggests many variables show similar significant SNP association patterns in both regions. However, mean temperature of the warmest and coldest quarter contributed most strongly to the first two CCA axes in both Asia and Europe (Figure 21). September precipitation and altitude also weighed heavily on the first two CCA axes in both Asia and Europe. Our analysis then reflects that in both regions seasonal and altitudinal temperature and precipitation extremes likely contribute to adaptation and, despite removing strongly correlated variables, it remains difficult to distinguish more specific associations among the general trends.

Although similar variables drove SNP associations in both regions, some SNPs with environmental associations overlapped between Europe and Asia, although few associations were shared. Over 20% (59) of putatively adaptive LFMM SNPs had significant associations in both Asia and Europe, far exceeding the maximum percent of common significant SNPs from 500 permutations of the environmental matrices (maximum 7.3%, P<0.0005). Similarly, Europe and Asia shared 109 out of a potential 642 (17%) combined unique genes that contained SNPs with significant associations (P=0.106). These genes only shared associations with an average of 26.2 ±1.9 % environmental variables and the proportion of associations with temperature and precipitation variables for shared genes were not significantly correlated between regions (Precipitation: P=0.86, rho=-0.016; temperature: P=0.46, rho=0.072). Furthermore, when we compared the proportion of significant SNPs within a shared gene segregating on each subgenome to test whether SNPs occurred on the same or opposite subgenomes, we found the proportion of significant SNPs within a shared gene segregating on each subgenome were not correlated between Europe and Asia. Only the proportion of SNPs occurring in both genomes in Europe, and the proportion of significant SNPs segregating only on the *C. grandiflora* subgenome in Asia were significantly negatively correlated (P=0.02, rho=-0.22), although this result is not significant under a Bonferroni correction for multiple testing. There then seems to be some evidence of shared putatively adaptive SNPs between regions, although these SNPs often associate with different environmental factors and subgenomes in each region SNP-environment associations between regions.

Finally, as population structure seems to dominate patterns in the genome, we repeated the LFMM analysis within north and south China both to try and further eliminate confounding population structure that could lead to false positives and explore how environmental associations can differ depending on the scale examined. We identified 131 SNPs with
significant associations in both north and the south based on 43 and 91 individuals, respectively. Of these associations, only 35 and 13 putatively adaptive SNPs in the north and south, respectively, were not also identified when we tested for environmental associations in all of Asia, and only three were common to both north and south China. Similarly, few SNPs with significant associations occur in genes that were detected in the continental analysis. Of 95 and 103 genes containing putatively adaptive SNPs in north and south China, only 15 and 2, respectively, were unique to this local scan. Genes with significant associations included four flowering time candidate genes in both regions, with all but one having been identified in the continental scan, although it was identified in Europe (ELF8). The strong overlap between the results of this local analysis and the continental analysis suggests that in our dataset environmental association scans at larger scales identify the majority of SNPs with adaptive patterns at regional scales, and local scans are unlikely to yield many new associations.

2.3.3 Enrichment analyses

We then used gene ontology enrichment analysis to explore the types of genes with possible environmental associations. Using gene ontology analysis on genes containing SNPs with significant associations through both LFMM and Bayenv, we saw possible enrichment for genes involved in response to abiotic factors and also detected flowering time candidate genes. In Asia, genes with putative environmental associations were most enriched for biological function GO terms involved in response to salt stress, response to abiotic stimulus, and response to osmotic stress (Table 2), although enrichment P-values are not significant with an FDR correction for multiple testing. These terms also comprised the top cluster of GO terms identified by DAVID. Similarly, top enriched GO terms in Europe included response to abiotic stimulus, response to chemical stimulus, and anatomical structure morphogenesis (Table 2), but enrichment was also non-significant after a FDR correction. The top enrichment cluster included response to abiotic stimulus, response to radiation, and response to light stimulus. Functional annotation of genes from both regions supported GO enrichment analysis, as we found genes with circadian clock, drought response, and light response functions in both regions. Additionally, we detected five flowering time candidate genes in Asia and ten flowering time candidate genes in Europe, of which three of which were common to both regions (Table 3). Two of the common flowering time candidate genes (LHY, TIC) are involved in regulating in the ‘morning’ circadian clock regulation pathway with CCA1. In Europe, we also found the
circadian clock gene LWD2 as well as two genes involved in vernalization (ELF8, FRL1). We also found two genes (FD in Europe and FY in Asia) whose mutants both cause late flowering. The presence of flowering time candidate genes and genes with gene ontology terms related to abiotic factor response align well with our expectations for SNPs associated with temperature and altitude in both regions.

Despite some putatively adaptive SNPs occurring in genes with ecologically relevant functions, SNPs with environmental associations were not enriched for coding sites and rather showed evidence of non-coding enrichment (Table 4). When we examined SNPs with significant environmental associations from each analysis to further investigate what kinds of SNPs were involved in adaptation, in Asia we only found significant enrichment for intergenic regions from the SNPs identified through Bayenv, and observed marginal enrichment for intergenic regions in SNPs found through LFMM for both the full SNP set and for SNPs with MAF<0.15. Similarly, in Europe we detected significant noncoding enrichment in the SNPs identified through LFMM for both the full and frequency controlled SNPs sets and marginal enrichment in the common SNPs. We also observed a significant reduction in non-synonymous coding sites in both SNPs identified through Bayenv, and low frequency MAF SNPs as well as in SNPs common between the Bayenv and the LFMM sets. No SNP sets showed significant coding, conserved noncoding, or stop codon gained site enrichment. We then tested whether noncoding SNPs were likely in part of the proximal promoters, but noncoding SNPs identified through both LFMM and Bayenv were not significantly closer to the nearest transcription start site than expected by chance (Asia: P=0.23, Europe P=0.72), although on average they are located within LD of the transcription start site of the nearest gene (Mean distance in Asia: 2052.18, Mean distance in Europe: 2298.75) suggesting they could either be linked to a nearby gene or are distal enhancers. To determine whether linked SNPs within genes were driving non-coding SNP environmental associations we investigated how often SNPs within a gene that were near a significant noncoding SNP also had significant associations. However linkage to SNPs within genes are unlikely to fully explain the noncoding signals as in both Europe and Asia the majority of noncoding SNPs that neighbor a gene containing a GBS SNP, most SNPs within the gene did not show a significant environmental association (25 out of 26 SNPs in Asia, 70 out of 86 SNPs in Europe).
2.3.4 SNP segregation patterns

When we examined how putatively adaptive SNPs segregated in both the phased GBS subgenomes, and the parental species genome samples, we found more SNPs with significant associations segregated in both the *C. grandiflora* subgenome and parental samples than the *C. orientalis* subgenome or samples in all SNP sets (Tables 5 and 6). While we did not find significantly more SNPs segregate exclusively on the *C. grandiflora* subgenome compared to the *C. orientalis* subgenome in all SNP sets except for XTX outliers in Asia. Additionally, in all but two SNP sets, more SNPs segregated in both genomes than in either genome exclusively. As all SNP sets had a high proportion of phased sites, the lack of significant differences in the number of SNPs segregating on either subgenome are likely not due to lower power because of poor phasing resolution. In contrast, when we compared the number of putatively adaptive SNPs segregating in the parental species samples, we consistently found highly significant excess of SNPs segregating in the *C. grandiflora* samples compared to the *C. orientalis* samples and differences in the number of SNPs segregating on the parental species genomes were highly significant for all SNP sets with consistently more SNPs segregating in the *C. grandiflora* than the *C. orientalis* samples. Relatively few SNPs segregated in both parental species, while more *C. bursa-pastoris* SNPs occurred at sites that are fixed differences in the parental species. However, the majority of SNPs did not segregate among samples from either species.

SNPs also often occurred in significantly larger or smaller geographic areas than expected by chance. All distance analyses showed highly bimodal p-value distributions, with the majority of P-values, based on the how many random samples of populations had larger mean pairwise distances than the populations where the SNP occurred, were of 1 or 0 (Figure 22). However, the distribution of P-values was slightly skewed towards higher P-values in Asia, but smaller P-values in Europe suggesting that in Europe more SNPs occur in larger areas than expected by chance, while Asia more SNPs occur in smaller areas than expected by chance. SNPs occurred in a similar number of populations in both areas (Asia 14.99,17.05; Europe 13.64, 16.00), but occurred in a larger mean area in Asia (898.72, 893.46 in the LFMM and common SNP sets) than in Europe (4564.14, 4766.88). These results did not change when we used Neslia to determine the area in which the derived allele segregated.
2.3.5 Selection scans

As expected, using scans on the whole genome samples based on haplotype homozygosity, we detected similar but slightly different sets of sweep candidates with the different selective sweep scans used, but consistently detected a higher proportion of sites with the highest 5% iHS or nSL values in the *C. grandiflora* than the *C. orientalis* subgenome in Europe (Table 7). All scans established more windows in the *C. orientalis* subgenome than the *C. grandiflora* subgenome regardless of region, likely because we marked the non-reference allele as derived. The higher number of points on the *C. orientalis* genome did not contribute to a significantly higher proportion of sweep candidates in Asia in most scans, except for iHS. Conversely, despite a greater number of points on the *C. orientalis* than the *C. grandiflora* subgenomes, European *C. grandiflora* subgenomes show significantly more sweep candidates than the *C. orientalis* subgenomes for all scans used (Table 7). We also observed significantly more sweep rich regions in the linkage scans for both iHS and nSL in Europe, and nSL in Asia when we compared the number of 100kb windows from each subgenome with the top 1% proportion of outliers, representing sweep candidate rich regions that are likely true positives (Voight et al 2006). Additionally, we detected significantly more XP-CLR outliers on the *C. grandiflora* than the *C. orientalis* subgenome when we compared northern and southern Asian populations.

When we used subgenome specific simulations to establish more stringent significance cutoffs for CLR sweep scans, we detected significantly more sweep scan candidates on the *C. grandiflora* subgenome in Europe, and an excess of sweep scan outliers on the *C. orientalis* subgenome in Asia both at all points and within 100kB windows (Table 7). This pattern remained significant when we divided the genomes into 100kb windows and compared the number of windows with at least a single significant CLR outlier (Table 7). Of the windows with significant outliers, three out of a possible seven windows on the *C. grandiflora* subgenome overlapped in Europe and Asia, and 11 out of a possible 22 windows on the *C. orientalis* subgenome overlapped between regions (P=0.0015 and P<0.0001 respectively), indicating significant overlap and possible convergent evolution. Gene ontology analysis showed significant enrichment among the *C. grandiflora* European windows containing CLR outliers for several terms including innate immune response (P<0.001), Toll-Interleukin receptors (P<0.001), immune response (P=0.0001), defense response (P<0.0001), apoptosis (P<0.001). *C. orientalis*
CLR outlier windows in Asia showed significant enrichment for a receptor-like protein kinase while CLR outliers in Europe were enriched for terms related to membrane proteins and receptor kinases (Table S4). Among genes found in CLR outlier windows in European samples we also detected two flowering time candidate genes in the \textit{C. orientalis} subgenome, and five in the \textit{C. grandiflora} subgenome, but none in the Asian CLR outlier windows.

When we examined overlap among CLR and nS\textsubscript{L} windows containing putative sweep outliers in \textit{Capsella bursa-pastoris} to sweep scan candidates in parental sample genomes we detected significant overlap in CLR but not nS\textsubscript{L} outliers. We detected significant overlap in \textit{C. bursa-pastoris grandiflora} and parental sweep outlier windows in both Europe and Asia, with 15.9\% of \textit{C. bursa-pastoris grandiflora} CLR outlier windows overlapping with outlier windows on the parental genome (P<0.0001) and 28.6\% overlapping in Asia (P=0.040). Similarly 20.9\% and 28.6\% of Asian \textit{C. bursa-pastoris orientalis} overlapped with parental CLR outliers in Europe and Asia respectively, (P=0.074, P=0.0209). In contrast, only \textit{C. bursa-pastoris orientalis} nS\textsubscript{L} outlier windows in Asia significantly overlapped with the \textit{C. orientalis} parental genome. As nS\textsubscript{L} outliers are often recent or incomplete sweeps, we expect more overlap between parental and \textit{C. bursa-pastoris} in the older CLR than nS\textsubscript{L} outliers.

Overlap among the top 5\% outliers for each selection scan in Asia generally conformed to our expectations based on the properties of the tests. iHS and XPCLR outliers were closest to outliers from the same scan on the other subgenome (P<0.0001), and nS\textsubscript{L} outliers also showed a notable but non-significant excess of nS\textsubscript{L} outliers from the sister genome within 5kB. Additionally, as expected due to their similar methodology, outliers from iHS scans were closer to outliers from nSL scans than expected by chance (P<0.001). No other pair of sweep scan outliers showed significant proximity to one another. No European scan outliers were significantly near each other. We found few outliers from any scan were shared in both Europe and Asia. No CLR outliers on either subgenome were common between Europe and Asia, and the two regions only shared 369 iHS outliers (4.3\%), 11 iHS windows (9.2\%), and 125 nS\textsubscript{L} outliers (1.4\%). Furthermore, of the shared outliers, 102 27.6\% of the iHS outliers, 63.3\% of the iHS outlier windows, and 2520\% of the nSL outliers did not occur on the same occurred on different subgenomes in Europe and Asia.
We did not observe strong evidence of putatively environmentally adaptive sites occurring near sweep scan outliers but found significant clustering in most SNP sets. In Asia, we found Bayenv SNPs and SNPs common to Bayenv and LFMM were significantly closer to both \textit{C. orientalis} and \textit{C. grandiflora} nSL outliers than expected by chance (P<0.001), and X^T\textbf{X} outliers were significantly near outliers from all sweep signals except CLR outliers on the \textit{C. grandiflora} genome. SNPs identified through both Bayenv and LFMM were also significantly near \textit{C. grandiflora} iHS and XP-CLR sites (P<0.001). Randomization did not show LFMM SNPs were significantly near any sweep scan outliers, but the distribution of distance to outliers showed an excess of nSL \textit{C. orientalis} outliers within 5kB of LFMM SNPs (Figure 23). Additionally, we observed an excess of LFMM and Bayenv SNPs within 5Kb of one another in both Europe and Asia (Figure S6). No SNP sets were significantly close to one another in Europe, but we did see an excess of nSL outliers within 5kB of both SNPs with associations identified through LFMM and Bayenv. SNPs shared from these two sets in Europe were also marginally significantly close to nSL outliers (P=0.13), showing slight but non-significant excess of nSL outliers within 5-10kB (Figure S7). We also detected significant clustering in \textit{C. orientalis} CLR, and both genome iHS, nSL, and LFMM sites in Asia and \textit{C. grandiflora} CLR, iHS, and nSL sites in Europe (P<0.001).

After finding little evidence of environmentally associated SNPs occurring near hard sweep outliers, we looked for subtler reductions in local diversity around outliers that would suggest selection. We found mixed evidence of selection around SNPs with significant environmental associations using more general population genetic patterns. Only a small fraction of 1Kb windows around SNPs jointly found in LFMM and Bayenv showed reduced diversity in either subgenome compared to windows around 10000 random SNPs. Under 5% and 10% of SNPs showed significantly reduced $\theta_\pi$ in Europe and Asia, respectively, and even fewer SNPs showed significantly reduced $\theta_w$ in both regions. Approximately 4.8% of SNPs showed more extreme Tajima’s D values compared to random SNPs in Europe, while only 5.8% (\textit{C. grandiflora}) and 2.9% (\textit{C. orientalis}) of Asian SNPs showed significantly extreme Tajima’s D values. However, the majority of significant Tajima’s D values were positive in all datasets (Table S5), suggesting an excess of intermediate frequency alleles. Similarly, we also found an excess of alleles with MAF from 0.4-0.5 and 0.9-1.0 we compared the LFMM SNP MAF distribution to the MAF distribution in sites with significant associations from 500 permutations.
of the environmental data matrix in both Europe and Asia (P<0.0001). We also observed as an excess of SNPs with MAF below 0.1 in Asia (P<0.0001) (Figure 24).

### 2.3.6 Linkage and effect sizes in putatively adaptive SNPs

Linkage and effect size patterns in putatively adapted SNPs generally reflect many linked putatively adaptive SNPs of small effect in Europe, and unlinked larger effect SNPs in Asia. Putatively adapted SNPs detected through both LFMM and Bayenv in Europe showed significantly higher mean pairwise $r^2$ both within (P<0.001) and across chromosomes (P<0.001) compared to 1000 random draws of SNPs from the corresponding dataset. European SNPs that were significant only through LFMM also showed significantly elevated linkage within (P=0.028) and across chromosomes (P<0.001). In contrast, LD among putatively adaptive SNPs in Asia was not significantly high within or among chromosomes compared to 1000 random samples. Effect sizes for SNPs with significant associations through LFMM (W=2024348, P<0.0001) and both LFMM and Bayenv (W=610101.5, P<0.0001) were also significantly higher in Asia than Europe (Figure 25).

### 2.3.7 Phenotypic experiment

Despite high survivorship to flowering, we detected extremely few significant SNP-phenotype associations. Over 87% of plants survived to transplantation, and of these 96.8% flowered. Although we detected a significant nested line within population effect (but varying heritability), as well as a significant population effect for all traits (Table 8 ), we only detected three significant SNP associations with biomass. These SNPs occurred within enzymes that suggested no obvious ecological importance and did not show significant environmental associations (Carubv10007831m.g, Carubv10003357m.g, Carubv10004089m.g ). No other traits, including fruit number, yielded significant associations through GWAS, although we found three marginal associations with leaf number (q-value<0.08). However, these SNPs are also not linked to any genes with functions immediately related to climate adaptation. The only marginally significant GWAS genes with potentially ecologically important functions contributed to disease and fungus resistance (CDR1,LOV1).

We then compared fitness to climate of origin to assess whether accessions from more similar climates to those experienced in Toronto had a fitness advantage indicative of local
adaptation. We used Gower’s distance to summarize climate distance from Toronto using the 18 WorldClim climate variables for each accession. While climate distance showed a significant effect on fruit number, this effect was no longer significant when we corrected for population structure both in the total and regional datasets. These results suggest that although populations differed in fitness, accessions from similar climates did not experience a strong home site advantage in the Toronto common garden. Fruit number in Toronto was instead significantly associated with population bolting time, biomass, branch number, and maximum inflorescence height, as well as through the block effect.

As overall local adaptation was too weak to detect through associations with overall habitat similarity, we looked for subtler signs of adaptation through associations with individual climate variables. We found several significant environment-phenotype correlations for all traits in both Europe and Asia suggesting subtler adaptive effects. In Asia bolting time, leaf number, branch number, maximum height and rosette diameter were all associated with mean temperature of the wettest quarter, while bolting time and rosette diameter were significantly associated with the annual temperature range and the precipitation of the driest quarter, respectively (Table 9). Maximum height was also associated with October precipitation. Fruit number was also associated with annual temperature range and was marginally associated with mean temperature of the wettest quarter when including other traits as covariates. None of the associations seen in Asian phenotypes occurred in European samples. In Europe both day to bolting and leaf number were only associated with the mean temperature of the warmest quarter, while rosette diameter was significantly associated with precipitation seasonality. The only post harvest trait associated with environmental variables in Europe was maximum height which was associated with September and October precipitation, as well as temperature annual range. Trait correlations may contribute to shared associations among traits as all traits were significantly correlated, but as no traits are fully correlated they are unlikely to completely explain them (Table S6).
2.4 Discussion

2.4.1 Local adaptation in Capsella bursa-pastoris

Overall, our results suggest *Capsella bursa-pastoris* locally adapted while spreading throughout Eurasia, although adaptation was not critical to its success. While environmental correlations can detect false positives due to confounding demography, SNPs identified through both individual and population based methods more likely represent true positives (de Villemereuil *et al.* 2014). Furthermore, SNPs with significant correlations, especially those identified through LFMM, explain a substantial portion of the variation in environmental variables, which likely reflect adaptive phenotypes. An excess of SNPs associated with precipitation in Asia are also ecologically sensible as tropical China was not glaciated and monsoons dominate climate fluctuations in this region (Zhuo *et al.* 1998). Some of the environmental variables with significant SNP correlations also align with associations in previous studies. Winter and early spring temperatures and precipitation measures were also shown to drive adaptive gradients in *Arabidopsis thaliana*, *Medicago truncatula*, and *Arabidopsis alpina* (Hancock *et al.* 2011; Manel *et al.* 2012; Yoder *et al.* 2014). These variables also explained the most variation along the principle axes of variation in SNP correlations in Asia and Europe, suggesting they may have had correlations with a large number of SNPs. Furthermore, some of these variables, or closely associated ones, also showed correlations with ecologically important phenotypic traits such as flowering time and vegetative size at lower which further asserts their adaptive importance (Nakazato *et al.* 2008).

Some putatively adaptive SNPs also occurred within key flowering time candidate genes that previously showed indications of adaptation both in *Capsella bursa-pastoris* and in other species (Slotte *et al.* 2007; Ni *et al.* 2009; Ma *et al.* 2010). Identifying flowering time and circadian clock candidate genes among the genes containing putatively adaptive SNPs affirm the adaptive value of SNPs identified through environmental associations. We identified several SNPs within genes involved in the circadian clock and flowering time regulation. Notably, the central circadian clock oscillator LHY showed significant environmental associations in both Europe and Asia, although through different SNPs in each region. LHY is a particularly strong adaptive candidate as it is both a transcription factor with a central role in the three interacting circadian rhythm loops (Hsu and Harmer 2014), and its differential expression contributes to
flowering time differences in *Capsella bursa-pastoris* (Slotte et al. 2007). In the current model of the Arabidopsis circadian clock, LHY and its paralog CCA1 are at the centre of three interlocking loops, where they integrate information from external light and temperature cues, and regulate evening circadian oscillators which then interact with flowering time regulators. Given its role as a key circadian oscillator, it is then not surprising that LHY is likely involved in flowering time adaptation as it shows differential expression in early and late flowering *Capsella bursa-pastoris* samples and latitudinal clines in Arabidopsis (Slotte et al. 2007; Ma et al. 2010). We also detected candidate genes involved in the vernalization flowering time pathway which is also known to shape flowering time clines in both *Capsella bursa-pastoris* and Arabidopsis (Slotte et al. 2007; Samis et al. 2012). Although we failed to detect significant environmental associations in the key vernalization genes FRI and FLC which often underlie flowering time clines in *Arabidopsis* and show latitudinal clines in both expression and polymorphism in *Capsella bursa-pastoris*, this is likely due to GBS limitations as no SNPs occurred within these genes in either Europe or Asia. Nevertheless these results further suggest the SNPs we detected with environmental associations are likely involved in local adaptation.

Despite finding promising environmental correlations, the absence of a significant home site advantage in the Toronto common garden suggests that environmental adaptation is subtle in *C. bursa-pastoris* and was likely not an essential factor in its spread and colonization. However, differential population and accession success suggests a genetic basis for fitness, and implies all plants did not perform equally well through plasticity like a general purpose genotype (Parker et al. 2003; Richards et al. 2006). Populations may have then differed in their ability to maintain fitness through plasticity. Adaptation, possibly to non-climatic factors, could have also subtly enhanced their performance. Indirect evidence of adaptation through trait associations with environmental variables, and SNP-environment associations suggests some adaptive differentiation in traits among *C. bursa-pastoris* populations throughout its range, and our experiment could have lacked sufficient power to detect their impact on fitness. Mixed signals of adaptation are consistent with previous studies that show *C. bursa-pastoris* employs plasticity, adaptation, and sorting of preadapted genotypes to colonize new environments, with the extent of adaptive differentiation depending on the trait (Neuffer and Hurka 1986a; b, 1999; Neuffer and Hoffrogge 2000). *C. bursa-pastoris* may then initially maintain fitness in new environments through plasticity, and later enhance fitness through adaptation in key traits (Lee 2002;
Ghalambor et al. 2007). Higher power and comparisons among multiple environments are then necessary to resolve the roles of plasticity and adaptation in determining fitness in new environments for *C. bursa-pastoris*.

### 2.4.2 Demographic and adaptive history

Our analysis of *Capsella bursa-pastoris*’s demographic history gives context to its recent adaptation, and is important in helping interpret signals of adaptation. The most prominent demographic pattern was the clear division between Asian and European populations. Asia and European clusters were clearly separated in all PCA, STRUCTURE, and Treemix analyses. Treemix analyses further emphasized the integrity of the division as there was little evidence of migration between regions and most trees show the division occurring deep in the tree before the separation of the Western European and Middle-Eastern populations. These results then support strong separation of the European and Asian *Capsella bursa-pastoris* populations, after its origin in Europe followed by expansion throughout Europe and recent growth in Asia from north to south, congruent with and A. Cornille’s Bayesian demographic models of our data (submitted) and Slotte’s analysis of six nuclear genes (2008). Limited linkage disequilibrium and isolation by distance suggests a recent, rapid expansion, especially in Asia, congruent with demographic modelling results from chloroplast and whole genome DNA in *Capsella bursa-pastoris* (Ceplitis *et al.* 2005; Douglas *et al.* 2014). Additionally, low isolation by distance and evidence of long distance gene flow through Treemix plots suggest *Capsella bursa-pastoris* is capable of long distance dispersal (possibly human-aided). Contemporary gene flow barriers may then not be as important to shaping population structure throughout Eurasia as demographic history.

Rapid expansion from glacial refugia likely helped create the patterns of population structure and diversity we observed. Recent *Capsella bursa-pastoris* expansion into Asia from a European refugial population during an interglacial could explain the observed phylogenetic relationships, reduction in diversity and the high number of shared alleles with *C. grandiflora* in Asia in both subgenomes if the refugial reserve was also ancestral to European populations. Slotte *et al.* (2008) supports this model as recent introgression between European but not Asian *Capsella bursa-pastoris* with *C. rubella* suggests Asian *Capsella bursa-pastoris* may have diverged before introgression occurred. This scenario is especially plausible if *C. rubella* is as recent as 25,000-50,000 years old and *Capsella bursa-pastoris* originated in the early Pleistocene.
over 300,000 years ago (Guo et al. 2009; Douglas et al. 2014; Roux and Pannell 2014), although estimates for *C. bursa-pastoris*’ age vary widely. High diversity in East Siberia that declines towards the West also suggests that *Capsella bursa-pastoris* may have had a refugium in Southeastern Siberia, which remained as taiga and steppe terrain during past glaciations (Tarasov et al. 2000; Bezrukova et al. 2008). It may have then expanded westward during interglacials like other terrestrial species such as the woodland tit, moose, and flying squirrel (Kvist et al. 2003; Oshida et al. 2005; Sipko and Kholodova 2009). Alpine glaciation in the Ursals to the West and the Stanovoy range to the south could have prevent gene flow from Russian to European and Asian populations (Sheinkman 2011), respectively, although we do observe some evidence of gene flow from Russia to Asia on the *C. orientalis* subgenome.

Expansion from traditional southern refugia in Europe was likely important to shaping patterns of diversity and population structure in Europe as well. Fossil evidence and genetic data suggest *Capsella bursa-pastoris* was present in Britain during the Ipswichian or even the Hoxne interglacial suggesting it endured one or two major glaciations (Coope et al. 1961, West et al. 1963). It is then likely that during these periods it retreated to some of the major southern refugia in Italy, the Balkans, and the Iberian Peninsula and subsequently expanded during interglacial periods. Expansion from the Iberian and Italian refugia, respectively, with the Alps acting as a barrier to gene flow, could help explain the division in the Western European and Mediterranean clades, as was hypothesized to have occurred in Arabidopsis, the European ash, and 22 tree species (Planck et al. 2000; Petit et al. 2003; Heuertz et al. 2004). Repeated expansion and contraction from these divergent refugia with long distance dispersal, as suggested by the lack of isolation by distance, could create the observed population structure and the overall low diversity we observed throughout our sample (Ibrahim et al. 1996; Hewitt 2000, 2004; Sexton et al. 2014). It would also account for the lack of a clear origin signal from regressions on heterozygosity.

This scenario is also consistent with the bottlenecks inferred through recent nuclear and chloroplast DNA Bayesian demographic inference (Ceplitis et al. 2005; Slotte et al. 2008). Although refugial expansion would predict higher diversity at southern than northern latitudes, higher diversity at northern latitudes could still occur if populations from distinct refugia mixed in these areas or if *Capsella bursa-pastoris* also remained in cryptic northern refugia which could have occurred as far north as Britain (Stewart and Lister 2001; Petit et al. 2003). Weedy, self-fertilizing species such as *Capsella bursa-pastoris* may have also had larger ranges during
glaciation, and may have been especially likely to occur further north on the disturbed moraine soil at the edge of the ice sheets, allowing them to quickly colonize northern latitudes during glacial retreat (Bhagwat and Willis 2008). Polyploids may have also been especially successful or even originated on these marginal glacial soils (Stebbins 1971). The movement of Neolithic farming from its origins in the Mediterranean throughout surrounding regions and Western Europe at the end of the Pleistocene (7-8 kya) (Colledge et al. 2004) could have also aided *Capsella bursa-pastoris*’s spread in Europe and influenced some contemporary patterns of population structure and help reinforce the observed East-West diversity gradient (A. Conille, unpublished manuscript), as it is likely did for *A. thaliana* (Planck et al. 2000).

Our analysis suggests demographic history shaped by Pleistocene events seems to be most important to shaping *Capsella bursa-pastoris*’s population structure and diversity, although uniquely polyploid features such as parental introgression and multiple origins may have also played a role. Through Treemix analysis we found evidence of limited parental introgression into nearby *Capsella bursa-pastoris* populations which could enhance adaptive success by transmitting adaptive alleles or enhancing local diversity. This parental introgression may have occurred prior to refugial expansion and spread as most introgression occurred at higher nodes rather than at branch tips. Parental introgression into Russia, and subsequent gene flow into Asia could also enhance diversity in the *C. bursa-pastoris* *C. orientalis* subgenome, also consistent with evidence of introgression between *C. orientalis* and Asian *C. bursa-pastoris* populations (Han et al. 2015). We also observed moderate admixture from *C. rubella* into Turkish populations, consistent with Slotte (2009)’s inferred introgression and *C. rubella*’s distribution in the Mediterranean. All signals of possible adaptation were also consistent with Stebbin’s (1971) prediction that gene flow among ploidy classes should be unidirectional from the diploid to the tetraploid species as inter-ploidy hybrids are likely to be triploid or tetraploid and can backcross more easily with tetraploid progenitors. Introgression from local parents is then likely to increase *Capsella bursa-pastoris*’s local diversity and transmit locally beneficial alleles, thus promoting the species’ success. In contrast, we found little evidence that multiple origins contribute to *Capsella bursa-pastoris*’s diversity despite previous assertions of high initial population size and the inference of multiple origins through high range-wide diversity. Only the highly divergent Greek and Italian samples show signals of a possible second origin. Given their central placement in the Mediterranean and its ample inferred admixture with other populations, it seems
unlikely that either divergent selection or drift explain their extreme divergence as seen in the PCA plots and maximum likelihood trees. However, the Greek sample did not show definite signals of a separate origin such as clustering with a distinct group of *C. grandiflora* samples or admixture from *C. grandiflora* or *C. orientalis* parental samples. Due to these shortcomings and our limited Greek sample size, a separate origin then remains a hypothesis. Together, these findings highlight important polyploid dynamics that can both contribute to population structure and enhance adaptive potential across the range.

Longer colonization history in Europe and recent, rapid adaptation in Asia also help explain the contrasting patterns of adaptive genetic architecture we observed in these regions. Fewer, more clustered adaptive loci, of greater effect in Asia are consistent with theoretical expectations for early stage adaptation with high migration (Feder *et al.* 2012b; Flaxman *et al.* 2014). Theory predicts that early in adaptation with gene flow adaptive loci will occur physically near strongly selected variants as they will then experience less gene flow in that region of the genome, resulting in few clusters of high divergence throughout the genome (Yeaman and Whitlock 2011; Nosil and Feder 2012). These predictions are consistent with our observation of significant clustering and higher effect sizes among adaptive SNPs in Asia. Clusters of adaptive variants are also more likely to form with limited long-range linkage disequilibrium, as we also observed in Asia but not in Europe. In contrast, the genetic architecture that we observed in Europe is more consistent with diffuse genome wide selection and polygenic adaptation predicted to occur later in adaptive divergence. Selection against migrants in between more differentiated populations where many loci have diverged can produce genome wide differentiation and diffuse genetic architecture for divergent loci involving many loci of small effect spread throughout the genome, as we observed in Europe (Nosil and Feder 2012; Feder *et al.* 2012b). Longer-range linkage disequilibrium in Europe, even among physically unlinked adaptive loci, as well as little clustering among adaptive loci and selection scan outliers are consistent with this theoretical framework. Strong European population structure created through expansion with bottlenecks from glacial refugia, selfing, and little long distance migration would further promote population divergence through drift and genome wide differentiation. Although we did not detect signals of isolation by adaptation, which theory predicts will occur with genome wide adaptive divergence, ecological differentiation can still occur if differentiation is recent and genetic drift also contributes to divergence, or if demographic history dominates.
spatial structure (Sexton et al. 2014). Strong gene flow can also overwhelm signals of isolation by adaptation (Nosil et al. 2009). As all these factors likely contribute strongly to genetic structure in our data, they may overwhelm genome wide signals of adaptive divergence even if they help shape the genetic architecture of environmental adaptation. Our data then emphasize how contrasting demographic histories and the resulting population structure can produce in disparate genetic architectures of adaptive traits in recently and more anciently diverged populations.

As the genome influences the effects of population level factors as well as selective pressures, *Capsella bursa-pastoris*’s polyploid features also likely contribute to its demographic history and adaptive success. *Capsella bursa-pastoris*’s polyploid genome and selfing ability likely helped it expand post-glaciation. Through selfing *Capsella bursa-pastoris* could readily colonize isolated areas through reproductive assurance, while maintaining diversity through its fixed heterozygous genome. High diversity and strong colonizing ability may explain why many polyploid species such as *Capsella bursa-pastoris* often occur in larger areas than their diploid ancestors in previously glaciated areas (Stebbins 1971), and could also enhance *Capsella bursa-pastoris*’s ability for long distance dispersal. Furthermore, higher mutation rates as well as buffering of maladaptive recessive mutations could allow new beneficial alleles to arise sooner in polyploids than in diploids, and allow polyploid populations to persist longer before extinction (Otto and Whitton 2000). These advantages may be especially important for selfing species, as populations may often go extinct after an environmental change if a rescuing mutation does not arise in time (Glémin and Ronfort 2013). Although we did not detect them in our dataset, post-polyploidization genome structural modifications such as deletions, inversions, and non-homeologous recombination could further enhance *Capsella bursa-pastoris*’s diversity and contribute to its post glacial expansion (Soltis and Soltis 2000). Additionally, these structural modifications, can also promote clustering of adaptive loci by locally reducing recombination (Kirkpatrick and Barton 2006). A polyploid genome then likely helped *Capsella bursa-pastoris* overcome the disadvantages of selfing and spread rapidly from refugia as glaciers retreated throughout the Pleistocene.
2.4.3 Polyploidy and adaptation

Diversity acquired from admixture of refugial populations as well as variation inherited from parents was likely important to *Capsella bursa-pastoris*’s adaptive success as many adaptive variants appear to have existed as standing variation throughout our sample. Several of our results suggest standing variation likely contributed importantly to adaptation in our sample. Our most direct evidence for the importance of standing variation in *Capsella bursa-pastoris*’s adaptation is that many adaptive variants from all putatively adaptive SNP sets segregate in the parental samples and a significant excess of CLR outliers overlap parental CLR outliers, particularly in the diverse *C. grandiflora* samples, suggesting they likely contributed to standing variation in *Capsella bursa-pastoris* populations prior to adaptation. Inherited standing variation from parents could also facilitate parallel adaptation in different regions and help explain the significant shared putatively adaptive SNPs we observed between Asia and Europe. An excess of intermediate frequency variants in Asia and especially an excess of intermediate and low frequency putatively adaptive variants in Europe are also consistent with expected allele spectra following adaptation from standing variation (Barrett and Schluter 2008). Positive significant Tajima’s D values around putatively adaptive SNPs also reflect that these SNPs are often at intermediate frequencies. Finally, GBS adaptive variants were most often significantly close to nSL and iHS outliers, both statistics which have more power to detect soft sweeps from standing variation, on both genomes (Voight *et al.* 2006; Ferrer-admetlla *et al.* 2014). Consistent with previous genomic scans in Arabidopsis, it then appears that much adaptation from *Capsella bursa-pastoris*’s range could originate from standing variation rather than new mutation (Hancock *et al.* 2011; Fournier-Level *et al.* 2011).

The importance of standing variation for adaptation in *Capsella bursa-pastoris* is not unexpected given its ecology and demographic history. Standing variation is likely to be especially important for rapid adaptation during population expansion, or after bottlenecks as small populations may go extinct in a new environment before an adaptive mutation occurs and rises in frequency (Hermisson and Pennings 2005; Prentis *et al.* 2008; Orr and Unckless 2008). Additionally, new mutations may easily have been lost to drift in small colonizing populations. Rapid adaptation may have been especially important for survival in the rapidly fluctuating climate of the Pleistocene. Several other Pleistocene species also show evidence of adaptation from standing variation (DE Carvalho *et al.* 2010; Fournier-Level *et al.* 2011). Aspen is
particularly interesting as it shows admixture among its divergent refugial populations during expansion further augmented its standing variation and facilitated adaptation (DE Carvalho et al. 2010), which could also have contributed to diversity in *Capsella bursa-pastoris*. Population divergence through both local adaptation and drift with moderate gene flow will substantially increase species wide diversity and increase the probability of adaptation from standing variation (Mitchell-Olds et al. 2007b; Lee and Mitchell-Olds 2011; Colautti et al. 2012; Hough et al. 2013). Admixture among refugia in *Capsella bursa-pastoris* could contribute to high silent site diversity observed in central Europe where Iberian and Italian refugial populations may have mixed as well as signatures of admixture between the Western European and Mediterranean clades seen throughout Europe in STRUCTURE and ADMIXTURE plots (A. Cornille, submitted), increasing *Capsella bursa-pastoris*’s adaptive potential. Given that alleles are more likely to contribute to adaptation when they have higher initial frequencies, adaptive variants segregating in the population due to parental diversity or demographic processes may have then been more likely to remain in the population. Despite its selfing habit and its history of strong population bottlenecks, *Capsella bursa-pastoris* may then have gained enough standing variation both from its parental genomes as well as from divergence and subsequent mixing of refugial that mutations segregating in the population were most often involved in adaptation.

However, demographic factors and methodological issues could also lead us to overestimate the importance of soft sweeps relative to new mutations. Demography often confounds sweep detection methods by producing extended linkage disequilibrium that nSL and iHS can interpret as sweeps (Teshima et al. 2006; Ferrer-admetlla et al. 2014; Huber et al. 2014). Areas of low recombination can also produce false positives in these tests (O’Reilly et al. 2008). Although simulations show nSL is particularly robust to false positives from recent bottlenecks, expansions, and recombination rate variation, population structure still reduces its accuracy (Ferrer-admetlla et al. 2014). Similarly, null distributions that do not sufficiently account for demographic history can increase both false positives and false negatives through MAF based tests such as CLR and XP-CLR (Huber et al. 2014). Inaccurately simulated null distributions may be an especially important concern for assessing differences between the subgenomes because if our estimates do not fully capture the disparate demographic history of each subgenome, higher linkage disequilibrium and standing variation in the *Capsella grandiflora* subgenome could then cause an excess of sweep signatures compared to the *Capsella orientalis*
subgenome, especially in Europe where we observe stronger population structure. Although using stringent significance cutoffs, a variety of tests, and ranking approaches should help mitigate these issues, reassessing sweep scan outliers against the best possible subgenome specific demographic distributions remains important for verifying our findings.

Even if sweep scan outliers reflect adaptation, they can produce an excess of soft sweep signals if adaptation occurs from hard sweeps at local scales. As selected alleles may occur on different backgrounds in each area, hard sweeps at small scales may then resemble soft sweeps in a regional sample (Long et al. 2013; Messer and Petrov 2013; Lee et al. 2014; Huber et al. 2014). Although environmental association tests are more robust to demographic effects, they may also overestimate the importance of standing variation for adaptation as they most easily detect SNPs segregating in large regional areas that incorporate a pronounced climatic gradient (Jones et al. 2013). We thus cannot discount the role of new mutations in Capsella bursa-pastoris’s spread. Despite Capsella bursa-pastoris’s recent origin, its rapid population expansion could have facilitated the spread of new mutations, as likely occurs in humans (Hawks et al. 2007; Pritchard et al. 2010; Fu and Akey 2013). Although genome scans are then subject to error and outliers need independent confirmation of their adaptive value, carefully accounting for demographic history and using varied genome scans that are robust to different demographic scenarios should help reduce the probability of false positives in our analysis and reveal general adaptive patterns.

Despite methodological uncertainty, the differing number of sweep signals on each subgenome in each region is consistent with evolutionary processes operating unevenly on allopolyploid subgenomes. While no previous work shows adaptation differs by subgenome, uneven gene expression and gene loss among subgenomes is common in allopolyploids (Tate et al. 2006; Rapp et al. 2009; Cheng et al. 2012). Maize, rice, Brassica, cotton, and Tragopogon all show biased fractionation among their subgenomes, implying different evolutionary forces acting on each subgenome (Liu et al. 2001, 2014; Rapp et al. 2009; Buggs et al. 2009; Flagel and Wendel 2010; Hohenlohe et al. 2010; Schnable et al. 2011). Although there is currently no evidence that gene loss patterns on each subgenome vary geographically, Tate et al. only examined samples within a small area and selection pressures on each genome may vary with stronger environmental gradients. Schnable (2011) proposed the more highly expressed subgenome will experience stronger selection and lose fewer genes. As subgenome expression levels can vary with the environment (Bardil et al. 2011), the level of selection each subgenome
experiences could also vary with the environment. Similarly, the excess of sweep signatures in *C. grandiflora* and *C. orientalis* in Europe and Asia, respectively, is consistent with subgenomes contributing more strongly towards adaptation in regions that resemble their ancestral range. Although we do not observe globally biased differences in gene expression between subgenomes (Douglas *et al.* 2014), we hypothesize that a subgenome may contribute more to adaptation in its native region if its alleles are closer to the adaptive optimum than the alleles on the other subgenome.

Parental subgenome alleles could be more fit in areas near their native range if they carry preexisting adaptive variation that selective sweeps can then modify in *C. bursa-pastoris*. While such a process has not been previously described in allopolyploids, the fitness advantage hybrids often experience near parental habitats due to parental similarity in ecologically important traits implies preadapted parental alleles can increase hybrid fitness in these environments. For example, introgressed herbivore and drought resistance QTL’s from a parental species that increase *Helianthus annuas* fitness in its parent’s native range clearly show preadapted parental alleles can contribute to a species’ success in a new environment (Rieseberg *et al.* 2003, 2007; Whitney *et al.* 2006). Similarly, introgressed genes for locally adapted ecologically important traits such as wing coloration in butterflies or warfarin resistance in mice often experience strong selection and show selective sweep signatures (Song *et al.* 2011).

Alleles on the local subgenome in *Capsella bursa-pastoris* could provide a similar fitness advantage where they provide partially adapted alleles that then experience strong selection as they acquire further adaptive changes from low frequency or new mutations, leaving evidence of selective sweeps. Further adaptations on partially pre-adapted alleles may have been necessary for adaptation to environments beyond the parental range as well as to the polyploid genome. Although confirmation of adaptive biases between subgenomes through methods that are less sensitive to confounding demographic factors are necessary to validate our results, if strong local biases in the number selective sweeps per genome occur in *Capsella bursa-pastoris* it would imply that parental pre-adaptation can also contribute towards polyploid adaptive success.

However, other parental genome features could also influences the differences in the adaptive contribution of each subgenome. Due to its high selfing rate, more deleterious mutations could have accumulated on the *C. orientalis* genome than the outcrossing *C.
grandiflora genome through Hill-Robertson effects (Wright et al. 2013). High selfing rates reduce effective population size as selfing alone reduces $N_e$ by $\frac{1}{2}$ and further diminishes it by reducing diversity and the effective recombination rate through increased heterozygosity (Charlesworth and Wright 2001; Charlesworth 2003; Wright et al. 2008). This reduction in effective population size then decreases the efficacy of background selection in purging new mutations (Charlesworth 2012). The lower replacement to synonymous diversity ratio in C. grandiflora compared to C. orientalis is consistent with reduces selection efficacy in the previously selfing subgenome. As the C. grandiflora subgenome harbors fewer deleterious alleles than the C. orientalis subgenome in Europe, adaptive mutations are probably less likely to be linked to deleterious mutations in the C. grandiflora subgenome and may sweep to high frequencies more easily (Douglas et al. 2014). A similar pattern may exist in Arabidopsis suecica alloployploids as they show greater gene loss and lower gene expression on the selfing A. thaliana subgenome compared to the outcrossing A. lyrata genome (Chang et al. 2010). Additionally, if adaptation occurs primarily from standing variation, higher diversity in C. grandiflora could also enhance its adaptive potential. C. grandiflora’s higher diversity and selection efficacy relative to C. orientalis could contribute to its excess of selective sweep outliers in Europe, and its consistently higher number of SNPs with environmental associations, which likely reflect adaptation from standing variation. These results then give insights into how parental genome history can shape the genomic architecture of adaptation in alloployploid genomes.

Distinct adaptive roles for each subgenome could also help explain why we did not observe clear reductions in pleiotropic constraint. Gene duplication is thought to be an essential way for proteins to avoid pleiotropic constraints and diversify in function (Hoekstra and Coyne 2007). If whole genome duplication reduced pleiotropic constraint, we should have observed fewer indicators of pleiotropic constraints including little parallel evolution and increased coding sequence adaptation. The significant excess of shared adaptive variants in our two geographically distinct regions and the enrichment of intergenic and noncoding adaptive SNPs outside of proximal promoter regions then suggest pleiotropy still limits the types of loci involved in Capsella bursa-pastoris adaptation. Although shared adaptive SNPs between regions may have already been present as standing variation in both regions, and are less likely to indicate strong pleiotropic constraints (Stern 2013), shared adaptive variants that segregate on
different subgenomes in each region may be more likely to have arisen independently and reflect developmental or pleiotropic constraints on adaptive sites. Additionally, further adaptive changes could occur on genes that already contain adaptive variants. Pleiotropy may still limit the genetic basis of adaptation in a recent allopolyploid such as *Capsella bursa-pastoris* if duplicated genes are not fully equivalent due to pre-existing deleterious or adaptive variation, or because of differences in expression. Gene loss that often occurs after whole genome duplication would further reduce the potential for weakening pleiotropic constraints through subfunctionalization in homeologous genes, although gene loss is limited in the *Capsella bursa-pastoris* genome (Douglas *et al.* 2014). If a copy of a gene performs more poorly than its homeolog, pleiotropic constraints could then remain important as the inferior homeolog would not perform the ancestral gene function as effectively as its diverging counterpart.

Given distinct adaptive values for genes on each subgenome, regulatory changes may also allow more precise expression control for each homeolog and better avoid pleiotropic consequences than protein coding changes. While their importance to adaptation remains controversial, regulatory mutations may have fewer pleiotropic consequences than coding mutations (Wray 2007; Carroll 2008; Stern *et al.* 2009). The modularity of cis-regulatory elements may allow for precise expression control of a sequence, reducing pleiotropic effects in sequences with diverse expression patterns throughout development and different tissues (Stern and Orgogozo 2008; Stern *et al.* 2009). Cis regulatory element adaptation may then allow for precise adjustments in phenotypes and contribute disproportionately to quantitative traits. The lack of pleiotropic consequences associated with cis regulatory mutations may then explain the excess of intergenic and non-coding putatively adaptive SNPs as they could allow for precise homeolog expression adjustment and phenotypic adaptation in quantitative traits. As the benefits of cis regulatory mutations may make them especially important to long term evolution (Stern *et al.* 2009), they may contribute disproportionately to adaptation if many adaptive SNPs in *Capsella bursa-pastoris* come from standing variation and have been previously been filtered by selection. Putatively adaptive non-genic SNPs may then indicate an evolutionary preference for adaptation through regulatory or structural mutations than coding mutations as they allow for precise changes in expression in *Capsella bursa-pastoris*’s allopolyploid genome.

Although previous prominent adaptation studies assumed genic SNPs would exceed non-coding SNPs among adaptive signals and supported the accuracy of their SNP associations with
coding enrichment, the assumption that coding sites are more likely to show adaptive signals likely does not apply to all species. Non-coding SNPs are clearly also important in evolution and the excess of non-coding putatively adaptive SNPs we observed does not necessarily invalidate our findings. Studies that identified putatively adaptive SNPs detected an excess of coding SNPs relative to non-coding SNPs, often use model species such as humans, and Arabidopsis and expression variation is also important for shaping intra and inter-specific adaptive trait divergence in these species (Hancock et al. 2010b, 2011; Lasky et al. 2014). In contrast, local adaptation scans in maize, poplar, black medic, fruit flies and sticklebacks all failed to detect a significant excess of genic relative to non-genic SNPs or detected an excess of non-genic SNPs, implying local adaptation may involve regulatory or structural variation in these species (Jones et al. 2012; Keller et al. 2012; Pyhäjärvi et al. 2013; Reinhardt et al. 2014; Yoder et al. 2014). Non-coding variants also contribute to domestication in maize, rabbits, and rice (Hufford et al. 2012; Carneiro 2014; Nabholz et al. 2014). In maize, genes controlling key domestication traits differ in expression level but not sequence in the domesticated and wild species and an excess of regulatory and non-coding sequences contribute to quantitative trait variation (Hufford et al. 2012). These studies show that cis regulatory variation can contribute substantially to phenotypic variation in traits under selection, suggesting that the excess of intergenic and non-coding sites we detected among putatively adaptive SNPs is likely to reflect a true adaptive pattern.

2.4.4 Methodological considerations and future directions

Comparing the diverse approaches we used to identify putatively adaptive loci gives empirical insights into the strengths and limitations of genome scan methods in a highly structured population. All genome scan methods compromise power or accuracy depending on the dataset, sampling scheme, and demographic scenarios. Our demographic analysis showed our dataset was highly historically structured, and did not clearly conform to a classic island or stepping stone model. These strong demographic patterns were also associated with environmental variation and likely raised the number of false positives we detected through environmental associations. Bayenv was particularly vulnerable to confounding demography as the excessive number and low explanatory power of associations it detected (over half of the SNPs), and the subsequent improvement in the quality of associations we found when we separated the Asian and European clades suggest it likely suffered a high rate of false positives.
In accordance to de Villemereuil’s (2014) findings, LFMM performed better given confounded demographic and environmental patterns but it also yielded a high number of significant associations in Europe, some of which are likely false positives. Using individual data rather than the more conservative approach of using population data could have inflated LFMM’s false-positive rate but it was likely the most appropriate approach for our data given our limited within population sampling. Our results then underline de Villemereuil’s recommendation of looking for shared signals through different environmental associations analyses is likely to yield the strongest set of true positives in a highly structured population, and agree with de Villemereuil’s general endorsement of LFMM as an adequate compromise of power and false discovery rate in complex demographic scenarios.

Consistent with findings from simulations, the differentiation ($X^TX$) scan was most conservative, yielding a small number of outliers which are likely truly adaptive as they almost completely overlapped with LFMM candidates and less so with Bayenv outliers (De Mita et al. 2013; de Villemereuil et al. 2014). However, given our limited GBS SNP set, there were too few $X^TX$ outliers to reveal larger adaptive patterns in our subsequent analysis. Although, unlike environmental associations, differentiation scans can reflect adaptation to factors not reflected in climate, $X^TX$ outlier annotation did not yield any strong candidates for biotic adaptation. Differentiation scans may then be most useful for highlighting the strongest adaptive signals, and verifying adaptive signals identified through more powerful approaches in small SNP sets.

Finally, the disappointing results from our common garden GWAS emphasize the usefulness of genome scans over traditional common gardens given logistical constraints. Given time and space constraints, we were only able to use a subset of our samples in the common garden GWAS which likely rendered it under-powered to detect SNP-trait associations. While we maximized genetic variation in the remaining sample, we had to eliminate 38% of the GBS-sequenced lines. Low sample size as well as a small SNP set, and large quantities of non-heritable genetic variation then likely impaired the experiment’s ability to detect SNP-trait associations. Although it remains problematic in Capsella bursa-pastoris, imputing SNPs using GBS data and a high quality phased reference panel would likely improve power. Furthermore, we only measured complex life history traits that involve a multitude of genes and are thus less likely to show strong associations with individual SNPs. Traits controlled by fewer genes, such as leaf shape, may then be more likely to show SNP associations, but are their ecological
relevance is often less clear. In contrast, using environmental factors as phenotypes produced numerous associations with SNPs as they benefited from larger sample size, greater variation, and clear ecological relevance. Although empirical validation is necessary to verify putatively adaptive SNPs identified through genome scans, our null common garden results suggest the increased power and flexibility of environmental association scans can be a powerful way to identify strong adaptive candidates and gain initial insights into adaptive processes given logistic constraints.

3 Conclusions

Although demographic history seems to most strongly shape the *C. bursa-pastoris* genome, subtle adaptive signatures suggest *C. bursa-pastoris* experienced some local adaptation throughout its range. Strong patterns of demographic structure, minor allele frequencies, and diversity suggest *C. bursa-pastoris* originated and spread in Europe and more recently colonized Asia. This demographic history led to signatures of early adaptive divergence in Asia, occurring through fewer SNPs of larger effect, and adaptation through more, highly linked SNPs of smaller effect in Europe, consistent with longer established genome wide, adaptive divergence among populations. Adaptation to temperature is more common in Europe, while adaptation to precipitation is more important in Asia. Candidate gene scans suggest circadian rhythm flowering time genes may contribute to adaptation in both ranges. However, phenotypic data suggests local adaptation is not a strong determinant of *C. bursa-pastoris*’ success in a new environment, and likely did not play a critical role in its spread. *C. bursa-pastrois*’ weedy lifestyle may have then allowed it to persist and spread in the shifting habitats of the Pleistocene while adaptation later adjusted traits towards local optima.

Patterns in putatively adaptive SNPs suggest *C. bursa-pastoris*’ duplicated genome did not facilitate adaptation by reducing pleiotropy. Consistent patterns of convergent adaptation in Europe and Asia both through environmental associations and selective sweep scans, as well as a
bias towards adaptation through noncoding SNPs suggests pleiotropy likely still constrains what sites can contribute to adaptation. *C. bursa-pastoris*’ subgenomes may be sufficiently different that genes duplicated on the two genomes may not be fully redundant, restricting at what sites sequence can change without pleiotropic consequences. Adaptive patterns consistent with pleiotropic constraint can also be remnants of adaptation in the diploid progenitors, as most environmental adaptation occurs through standing variation segregating in the parental genomes. The differing evolutionary history and features of *C. bursa-pastoris*’ recent allopolyploid subgenomes may then limit how much genome duplication releases pleiotropic constraint, especially when adaptation occurs from standing variation.

The contrasting features of the *C. grandiflora* than *C. orientalis* subgenomes contribute differently to adaptation from standing variation and new mutations. Lower selection efficacy and more high frequency alleles in the *C. grandiflora* than the *C. orientalis* subgenome may further enhance its adaptive potential in *C. bursa-pastoris*. The higher diversity in the *C. grandiflora* than *C. orientalis* subgenome likely lead it to contribute more to environmental adaptation, which most often occurs from standing variation. However, scans for selective sweeps from new mutations suggest *C. orientalis* may contribute more to adaptation near its native range in Asia as its alleles may already be closer to the local optimum. The importance and contribution of *C. bursa-pastoris*’ two subgenomes to its adaptation may then be regionally specific.
4 Figures and Tables

Figure 1. Sampled populations in A) the 261 accession GBS dataset B) the 24 accession whole genome dataset
Figure 2

PCA of the unphased dataset from Eigenstrat. The first two eigenvalues explain 0.38 and 0.19 of the variance, respectively.
Figure 3. Structure plots based on unphased GBS SNPS for K=2 to K=4.
Figure 4. Treemix results for the unphased dataset. **A.** Residual matrix for the ML without any migration edges, extreme colours indicate excess similarity not explained by the tree and likely migration events. **B.** Tree with four imposed, significant migration edges.
Figure 5. STRUCTURE plots from K=2 for European samples alone.
Figure 6. PCA plot of the phased grandiflora GBS SNPs with the grandiflora parental samples. The first two PCA axes explain 0.29, and 0.12 of the variance, respectively.
Figure 7. STRUCTURE plots for the phased grandiflora GBS samples with the parental samples. Structure harvester identifies K=2 as the most likely grouping.
Figure 8. Treemix trees of the phased grandiflora subgenome with grandiflora samples (A) and both grandiflora and rubella samples (B).
Figure 9. PCA plot of the phased orientalis GBS SNPs with the orientalis parental samples. The first two PCA axes explain 0.26, and 0.18 of the variance, respectively.
Figure 10. STRUCTURE plots for the phased orientalis GBS samples with the parental samples. Structure harvester identifies K=2 as the most likely grouping.
Figure 11. Treemix tree of the phased orientalis subgenome and the parental samples with 4 migration edges.
Figure 12. Boxplots of silent site expected heterozygosity by region in phased *C. grandiflora* subgenome (A.) and *C. orientalis* subgenome (B.) SNPs. Bars indicate the median.
Table 1. Mantel P and r values testing for a) isolation by distance and b) isolation by adaptation in each region and each subgenome.

A.

<table>
<thead>
<tr>
<th></th>
<th>Mantel P</th>
<th>Mantel r</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. gr Asia</td>
<td>0.14</td>
<td>0.11</td>
</tr>
<tr>
<td>C. gr Mediterranean</td>
<td>0.065</td>
<td>0.35</td>
</tr>
<tr>
<td>C. gr W. Europe</td>
<td>0.083</td>
<td>0.15</td>
</tr>
<tr>
<td>C. or Asia</td>
<td>0.19</td>
<td>0.10</td>
</tr>
<tr>
<td>C. or Mediterranean</td>
<td>0.052</td>
<td>0.36</td>
</tr>
<tr>
<td>C. or W. Europe</td>
<td>0.041</td>
<td>0.19</td>
</tr>
</tbody>
</table>

B.

<table>
<thead>
<tr>
<th></th>
<th>Mantel P</th>
<th>Mantel r</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. gr Asia</td>
<td>0.35</td>
<td>0.033</td>
</tr>
<tr>
<td>C. gr Mediterranean</td>
<td>0.14</td>
<td>0.23</td>
</tr>
<tr>
<td>C. gr W. Europe</td>
<td>0.67</td>
<td>-0.061</td>
</tr>
<tr>
<td>C. or Asia</td>
<td>0.10</td>
<td>0.17</td>
</tr>
<tr>
<td>C. or Mediterranean</td>
<td>0.24</td>
<td>0.17</td>
</tr>
<tr>
<td>C. or W. Europe</td>
<td>0.77</td>
<td>-0.089</td>
</tr>
</tbody>
</table>
Figure 13. Maps of random points used as potential origins, and the resulting $r^2$ when we regressed the distance from that point with expected silent site heterozygosity. Points with maximal are squares are marked black, and lines mark the $125^\circ$ and $135^\circ$ longitude lines.

A.

B.
Figure 14. Mean pairwise $r^2$ and physical distance among SNPs in the *C. grandiflora* (A,B) and *C. orientalis* (C,D) subgenomes, in Europe (A,C) and Asia (B,D).
C.

D.
Figure 15. Mean shared haplotype length by genome and region with standard error.
Figure 16. Mean haplotype length shared among samples in the three main regional clusters in the *C. grandiflora* (A) and *C. orientalis* (B) subgenomes

A.

B.
Figure 17.

Minor allele frequency distribution from the phased subgenomes. A and B compare subgenomes within Asia (A) and Europe (B), while C and D compare the distributions between regions within the *C. grandiflora* (C) and *C. orientalis* (D) subgenomes.
Figure 18. The joint derived frequency spectra of the A C. grandiflora and B. C. orientalis subgenomes

A.

B.
Figure 19. Ratio of replacement to synonymous $\theta_r$ in the *C. grandiflora* and *C. orientalis* subgenome
Figure 20. Number of significant associations for each variable by region through LFMM (A) and observed and expected (italics) number of associations for precipitation and temperature variables by region (B). Expected numbers are based on expectations for the $\chi^2$ contingency test.

<table>
<thead>
<tr>
<th></th>
<th>Precipitation</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>1071</td>
<td>680</td>
</tr>
<tr>
<td></td>
<td>987.79</td>
<td>763.21</td>
</tr>
<tr>
<td>Europe</td>
<td>1678</td>
<td>1444</td>
</tr>
<tr>
<td></td>
<td>1761.21</td>
<td>1360.78</td>
</tr>
</tbody>
</table>
Figure 21. CCA plots based on a SNP by variable matrix of significant associations showing only variables for Asia (A) and Europe (B) as well as a key to the variable names (C).
C.

<table>
<thead>
<tr>
<th>Number</th>
<th>Variable</th>
<th>Precipitation or Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aridity</td>
<td>P</td>
</tr>
<tr>
<td>2</td>
<td>Altitude</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Prec. March</td>
<td>P</td>
</tr>
<tr>
<td>4</td>
<td>Prec. Sept</td>
<td>P</td>
</tr>
<tr>
<td>5</td>
<td>Prec. Oct</td>
<td>P</td>
</tr>
<tr>
<td>6</td>
<td>Annual Mean Temperature</td>
<td>T</td>
</tr>
<tr>
<td>7</td>
<td>Mean Diurnal Range (Mean of monthly (max temp - min temp))</td>
<td>T</td>
</tr>
<tr>
<td>8</td>
<td>Isothermality</td>
<td>T</td>
</tr>
<tr>
<td>9</td>
<td>Temperature Annual Range</td>
<td>T</td>
</tr>
<tr>
<td>10</td>
<td>Mean Temperature of Wettest Quarter</td>
<td>T</td>
</tr>
<tr>
<td>11</td>
<td>Mean Temperature of Warmest Quarter</td>
<td>T</td>
</tr>
<tr>
<td>12</td>
<td>Mean Temperature of Coldest Quarter</td>
<td>T</td>
</tr>
<tr>
<td>13</td>
<td>Annual Precipitation</td>
<td>P</td>
</tr>
<tr>
<td>14</td>
<td>Precipitation Seasonality (Coefficient of Variation)</td>
<td>P</td>
</tr>
<tr>
<td>15</td>
<td>Precipitation of Wettest Quarter</td>
<td>P</td>
</tr>
<tr>
<td>16</td>
<td>Precipitation of Driest Quarter</td>
<td>P</td>
</tr>
<tr>
<td>17</td>
<td>Precipitation of Warmest Quarter</td>
<td>P</td>
</tr>
<tr>
<td>18</td>
<td>Precipitation of the Coldest Quarter</td>
<td>P</td>
</tr>
</tbody>
</table>
Table 2. GO terms with significant enrichment in Asia and Europe

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>%</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Asia</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009651~response to salt stress</td>
<td>5</td>
<td>6.097561</td>
<td>0.010891</td>
</tr>
<tr>
<td>GO:0010035~response to inorganic substance</td>
<td>5</td>
<td>6.097561</td>
<td>0.016691</td>
</tr>
<tr>
<td>GO:0010038~response to metal ion</td>
<td>5</td>
<td>6.097561</td>
<td>0.016691</td>
</tr>
<tr>
<td>GO:0008270~zinc ion binding</td>
<td>12</td>
<td>14.63415</td>
<td>0.020211</td>
</tr>
<tr>
<td>GO:0003677~DNA binding</td>
<td>15</td>
<td>18.29268</td>
<td>0.023259</td>
</tr>
<tr>
<td>GO:0009628~response to abiotic stimulus</td>
<td>10</td>
<td>12.19512</td>
<td>0.029158</td>
</tr>
<tr>
<td>GO:0006970~response to osmotic stress</td>
<td>5</td>
<td>6.097561</td>
<td>0.033285</td>
</tr>
<tr>
<td><strong>Europe</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009628~response to abiotic stimulus</td>
<td>25</td>
<td>10.72961</td>
<td>0.023134</td>
</tr>
<tr>
<td>GO:0009628~response to abiotic stimulus</td>
<td>25</td>
<td>10.72961</td>
<td>0.023863</td>
</tr>
<tr>
<td>GO:0042221~response to chemical stimulus</td>
<td>25</td>
<td>10.72961</td>
<td>0.040317</td>
</tr>
<tr>
<td>GO:0009653~anatomical structure morphogenesis</td>
<td>12</td>
<td>5.150215</td>
<td>0.049205</td>
</tr>
</tbody>
</table>
Table 3. Flowering time candidate genes detected through SNP environment associations, with abbreviated TAIR descriptions

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>FES1</td>
<td>Represses flowering along with FRI</td>
<td>Asia</td>
</tr>
<tr>
<td>FY</td>
<td>Flowering time regulator</td>
<td>Asia</td>
</tr>
<tr>
<td>LHY</td>
<td>Transcription factor involved in circadian rhythm</td>
<td>Asia, Europe</td>
</tr>
<tr>
<td>MYB91</td>
<td>Involved in regulating development through auxin response</td>
<td>Asia, Europe</td>
</tr>
<tr>
<td>TIC</td>
<td>Circadian rhythm regulator, regulates CCA1 through LHY</td>
<td>Asia, Europe</td>
</tr>
<tr>
<td>ARF6</td>
<td>Auxin response factor</td>
<td>Europe</td>
</tr>
<tr>
<td>DWARF-2</td>
<td>Kinase involved in flowering time regulation through BR1 signal transduction</td>
<td>Europe</td>
</tr>
<tr>
<td>ELF8</td>
<td>Flowering time regulator that</td>
<td>Europe</td>
</tr>
</tbody>
</table>
works with FLC to produce winter annual habits

<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession</th>
<th>Description</th>
<th>Continent</th>
</tr>
</thead>
<tbody>
<tr>
<td>FD</td>
<td>AT4G35900</td>
<td>Flowering time regulator</td>
<td>Europe</td>
</tr>
<tr>
<td>FRL-1</td>
<td>AT5G16320</td>
<td>Frigida related gene that contributes to a winter annual habit</td>
<td>Europe</td>
</tr>
<tr>
<td>HASTY 1</td>
<td>AT3G05040</td>
<td>Importin/exportin gene involved in shoot maturation</td>
<td>Europe</td>
</tr>
<tr>
<td>LWD2</td>
<td>AT3G26640</td>
<td>Circadian clock protein</td>
<td>Europe</td>
</tr>
</tbody>
</table>
Table 4. P-values testing for annotation enrichment in SNPs with significant associations,* indicates significant enrichment, † indicates significant reduction. For each dataset, enrichment P-values for the full dataset are listed above enrichment P-values for SNPs with MAF<0.15.

<table>
<thead>
<tr>
<th></th>
<th>Coding</th>
<th>Non-coding</th>
<th>Coding/noncoding</th>
<th>Non-synonymous coding</th>
<th>Synonymous coding</th>
<th>Stop Gained</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bayenv Asia</strong></td>
<td>0.852</td>
<td>0.112</td>
<td>0.871</td>
<td>0.842</td>
<td>0.207</td>
<td>0.026*</td>
</tr>
<tr>
<td></td>
<td>0.386</td>
<td>517</td>
<td>0.434</td>
<td>0.544</td>
<td>0.125</td>
<td>0.072</td>
</tr>
<tr>
<td><strong>LFMM Asia</strong></td>
<td>0.792</td>
<td>0.303</td>
<td>0.747</td>
<td>0.534</td>
<td>0.707</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>0.467</td>
<td>0.331</td>
<td>0.564</td>
<td>0.324</td>
<td>0.584</td>
<td>0.243</td>
</tr>
<tr>
<td><strong>Common Asia</strong></td>
<td>0.557</td>
<td>0.321</td>
<td>0.628</td>
<td>0.17</td>
<td>0.425</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>0.467</td>
<td>0.331</td>
<td>0.564</td>
<td>0.324</td>
<td>0.584</td>
<td>0.243</td>
</tr>
<tr>
<td><strong>Bayenv Europe</strong></td>
<td>0.805</td>
<td>0.165</td>
<td>0.822</td>
<td>0.999†</td>
<td>0.011*</td>
<td>0.518</td>
</tr>
<tr>
<td></td>
<td>0.703</td>
<td>0.235</td>
<td>0.732</td>
<td>0.984†</td>
<td>0.044</td>
<td>0.409</td>
</tr>
<tr>
<td><strong>LFMM Europe</strong></td>
<td>0.962</td>
<td>0.026*</td>
<td>0.969</td>
<td>0.831</td>
<td>0.774</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>0.994†</td>
<td>0.336</td>
<td>0.958</td>
<td>0.886</td>
<td>0.951</td>
<td>0.012*</td>
</tr>
<tr>
<td><strong>Common Europe</strong></td>
<td>0.881</td>
<td>0.087</td>
<td>0.901</td>
<td>0.992†</td>
<td>0.125</td>
<td>0.461</td>
</tr>
<tr>
<td></td>
<td>0.716</td>
<td>0.185</td>
<td>0.776</td>
<td>0.952</td>
<td>0.104</td>
<td>0.465</td>
</tr>
</tbody>
</table>
Table 5. Putatively adaptive SNP segregation in the parental subgenomes. Common refers to SNPs identified through both Bayenv and LFMM.

<table>
<thead>
<tr>
<th></th>
<th>C. grandiflora</th>
<th>C. orientalis</th>
<th>Both</th>
<th>$\chi^2$</th>
<th>P value</th>
<th>% phased</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X\textsuperscript{TX}</td>
<td>18</td>
<td>27</td>
<td>23</td>
<td>1.8</td>
<td>0.17</td>
<td>100</td>
<td>76</td>
</tr>
<tr>
<td>Bayenv</td>
<td>124</td>
<td>107</td>
<td>149</td>
<td>1.25</td>
<td>0.26</td>
<td>95.27027</td>
<td>444</td>
</tr>
<tr>
<td>LFMM</td>
<td>196</td>
<td>188</td>
<td>231</td>
<td>0.167</td>
<td>0.68</td>
<td>92.18329</td>
<td>742</td>
</tr>
<tr>
<td>Common</td>
<td>103</td>
<td>85</td>
<td>118</td>
<td>1.72</td>
<td>0.18</td>
<td>94.63277</td>
<td>354</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X\textsuperscript{TX}</td>
<td>21</td>
<td>20</td>
<td>26</td>
<td>0.0244</td>
<td>0.87</td>
<td>100</td>
<td>79</td>
</tr>
<tr>
<td>Bayenv</td>
<td>100</td>
<td>106</td>
<td>125</td>
<td>0.174</td>
<td>0.67</td>
<td>91.33489</td>
<td>427</td>
</tr>
<tr>
<td>LFMM</td>
<td>68</td>
<td>55</td>
<td>96</td>
<td>1.34</td>
<td>0.24</td>
<td>88.6121</td>
<td>281</td>
</tr>
<tr>
<td>Common</td>
<td>32</td>
<td>23</td>
<td>31</td>
<td>1.47</td>
<td>0.22</td>
<td>88.34951</td>
<td>103</td>
</tr>
</tbody>
</table>
Table 6. Putatively adaptive SNP segregation in the parental genomes. Common refers to SNPs identified through both Bayenv and LFMM.

<table>
<thead>
<tr>
<th></th>
<th>C. gr</th>
<th>C. or</th>
<th>both</th>
<th>Fixed difference</th>
<th>total</th>
<th>$\chi^2$</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X^T X</td>
<td>19</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>76</td>
<td>13.76</td>
<td>0.0002</td>
</tr>
<tr>
<td>Bayenv</td>
<td>132</td>
<td>7</td>
<td>14</td>
<td>54</td>
<td>444</td>
<td>112.41</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LFMM</td>
<td>186</td>
<td>19</td>
<td>17</td>
<td>76</td>
<td>742</td>
<td>136.04</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Common</td>
<td>105</td>
<td>7</td>
<td>10</td>
<td>44</td>
<td>354</td>
<td>85.75</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X^T X</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>79</td>
<td>2.78</td>
<td>0.09</td>
</tr>
<tr>
<td>bayenv</td>
<td>75</td>
<td>18</td>
<td>6</td>
<td>35</td>
<td>427</td>
<td>34.94</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LFMM</td>
<td>36</td>
<td>6</td>
<td>6</td>
<td>21</td>
<td>281</td>
<td>21.42</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Common</td>
<td>16</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>103</td>
<td>7.2</td>
<td>0.007</td>
</tr>
</tbody>
</table>
Histograms of P-values reflecting if putatively adaptive SNPs occur in smaller areas than expected by chance based on the average 1000 random samples (i.e., P-value=0 means the mean pairwise distance among populations where the alternate SNP occurs was smaller than all 1000 random population samples). The histograms are from analyses based on SNPs that had significant environmental association in both LFMM and Bayenv in A. Asia and B. Europe.
Table 7. The number of selection scan outliers found on each subgenome. A) All outliers and iHS and nSI 100kb windows with the highest proportion of outliers. B) The number of 100kB windows with at least one outlier

A)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C. grandiflora</th>
<th>C. orientalis</th>
<th>X²</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>iHS outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>3098</td>
<td>5375</td>
<td>9.10</td>
<td>0.00255</td>
</tr>
<tr>
<td>Europe</td>
<td>13909</td>
<td>11068</td>
<td>72.41</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>iHS windows</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>16</td>
<td>10</td>
<td>1.38</td>
<td>0.2393</td>
</tr>
<tr>
<td>Europe</td>
<td>18</td>
<td>7</td>
<td>4.84</td>
<td>0.0278</td>
</tr>
<tr>
<td>nSL outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>2889</td>
<td>5593</td>
<td>3.29</td>
<td>0.07</td>
</tr>
<tr>
<td>Europe</td>
<td>13320</td>
<td>11671</td>
<td>8.95</td>
<td>0.0028</td>
</tr>
<tr>
<td>nSL windows</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>18</td>
<td>6</td>
<td>6</td>
<td>0.01</td>
</tr>
<tr>
<td>Europe</td>
<td>19</td>
<td>7</td>
<td>5.54</td>
<td>0.019</td>
</tr>
<tr>
<td>CLR outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>7</td>
<td>22</td>
<td>4.21</td>
<td>0.0040</td>
</tr>
<tr>
<td>Europe</td>
<td>73</td>
<td>48</td>
<td>5.16</td>
<td>0.023</td>
</tr>
<tr>
<td>XP-CLR outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>3604</td>
<td>2333</td>
<td>383.89</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
B)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C. grandiflora</th>
<th>C. orientalis</th>
<th>X²</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>iHS</td>
<td>Asia</td>
<td>733</td>
<td>859</td>
<td>8.06</td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>657</td>
<td>897</td>
<td>13.91</td>
</tr>
<tr>
<td>nSL</td>
<td>Asia</td>
<td>793</td>
<td>1038</td>
<td>28.98</td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>884</td>
<td>1077</td>
<td>17.28</td>
</tr>
<tr>
<td>CLR outliers</td>
<td>Asia</td>
<td>7</td>
<td>21</td>
<td>8.81</td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>69</td>
<td>43</td>
<td>6.06</td>
</tr>
</tbody>
</table>
Figure 23. Count of pairwise distances from Bayenv (A,C) and LFMM (B,D) SNPs to selection scan outliers in Asia (A,B) and Europe (C,D). Error bars show 95% confidence intervals based on 500 bootstrap samples.
C.
D.
Figure 24. Minor allele frequency spectra of SNPs with significant environmental associations in
A. Asia and B. Europe with 95% confidence intervals based on 500 permutations.

A.

B.
Figure 25. Boxplot of Z-scores of SNPs with significant associations in Asia and Europe from LFMM (A) and both LFMM and Bayenv (B). Z-scores are standardized β terms from the LFMM mixed model regression (Frichot et al 2013).
Table 8. Significance of block, line, and population from mixed models on phenotypic traits in the common garden

<table>
<thead>
<tr>
<th>Trait</th>
<th>Term</th>
<th>Test statistic</th>
<th>P-value</th>
<th>Percentage of variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolting</td>
<td>Block</td>
<td>4.94</td>
<td>0.026</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Line</td>
<td>173.45</td>
<td>&lt;0.0001</td>
<td>30.8</td>
</tr>
<tr>
<td></td>
<td>Population</td>
<td>34.11</td>
<td>&lt;0.0001</td>
<td>55.6</td>
</tr>
<tr>
<td>Leaf number</td>
<td>Block</td>
<td>0.067</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Line</td>
<td>19.18</td>
<td>&lt;0.0001</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>Population</td>
<td>93.76</td>
<td>&lt;0.0001</td>
<td>58.5</td>
</tr>
<tr>
<td>Diameter</td>
<td>Block</td>
<td>5.06</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Line</td>
<td>44.51</td>
<td>&lt;0.0001</td>
<td>16.82</td>
</tr>
<tr>
<td></td>
<td>Population</td>
<td>36.67</td>
<td>&lt;0.0001</td>
<td>38.33</td>
</tr>
<tr>
<td>Maximum height</td>
<td>Block</td>
<td>0.0034</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Line</td>
<td>27.592</td>
<td>&lt;0.0001</td>
<td>14.25</td>
</tr>
<tr>
<td></td>
<td>Population</td>
<td>36.24</td>
<td>&lt;0.0001</td>
<td>29.5</td>
</tr>
<tr>
<td>Dry mass</td>
<td>Block</td>
<td>3.75</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Line</td>
<td>8.52</td>
<td>0.0035</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Population</td>
<td>25.73</td>
<td>&lt;0.0001</td>
<td>21.5</td>
</tr>
<tr>
<td>Fruit number</td>
<td>Block</td>
<td>7.18</td>
<td>0.0075</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Line</td>
<td>12.32</td>
<td>0.00045</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>Population</td>
<td>15.9</td>
<td>&lt;0.0001</td>
<td>14.7</td>
</tr>
<tr>
<td>Branch number</td>
<td>Block</td>
<td>0.15</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Line</td>
<td>66.57</td>
<td>&lt;0.0001</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>Population</td>
<td>7.37</td>
<td>0.0066</td>
<td>11.9</td>
</tr>
</tbody>
</table>
Table 9. P-values for selected environmental variables from mixed models on phenotypic traits in Europe and Asia

<table>
<thead>
<tr>
<th></th>
<th>Asia</th>
<th>Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bolting</strong></td>
<td>0.00038</td>
<td>0.0063</td>
</tr>
<tr>
<td><strong>Leaf Number</strong></td>
<td>0.0037</td>
<td></td>
</tr>
<tr>
<td><strong>Diameter</strong></td>
<td>0.00013</td>
<td>0.00044</td>
</tr>
<tr>
<td><strong>Fruit Number</strong></td>
<td>0.021</td>
<td>0.027</td>
</tr>
<tr>
<td><strong>Maximum height</strong></td>
<td>0.097</td>
<td>0.0094</td>
</tr>
<tr>
<td><strong>Branch Number</strong></td>
<td>0.044</td>
<td></td>
</tr>
<tr>
<td><strong>Bolting</strong></td>
<td>0.011</td>
<td>0.0099</td>
</tr>
<tr>
<td><strong>Leaf Number</strong></td>
<td></td>
<td>0.038</td>
</tr>
<tr>
<td><strong>Diameter</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Maximum height</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Temperature Annual Range

Mean Temperature of Wettest Quarter

Mean Temperature of Warmest Quarter

Precipitation Seasonality (Coefficient of Variation)

Precipitation of Driest Quarter

September Precipitation

October Precipitation
References


Anderson J. T., Willis J. H., Mitchell-Olds T., 2011 Evolutionary genetics of plant adaptation. Trends Genet. 27: 258–266.


Carneiro M., 2014 Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. 1074.


Hahn M. a, Kleunen M. van, Müller-Schärer H., 2012 Increased phenotypic plasticity to climate may have boosted the invasion success of polyploid Centaurea stoebe. PLoS One 7: e50284.


Hewitt G., 2000 The genetic legacy of the Quaternary ice ages.


Kirkpatrick M., Barton N., 2006 Chromosome inversions, local adaptation and speciation. Genetics **173**: 419–34.


Neuffer B., Hoffrogge R., 2000 ECOTYPIC AND ALLOZYME VARIATION OF CAPSELLA BURSA-PASTORIS AND C. RUBELLA (BRASSICACEAE) ALONG LATITUDE AND ALTITUDE GRADIENTS ON THE D3ERIAN PENÍNSULA by


Pannell J. R., 2015 Inferring the mode of origin of polyploid species from next-generation sequence data.


Petit C., Thompson J. D., 1999 Species diversity and ecological range in relation to ploidy level in the flora of the Pyrenees. : 45–66.


Sheinkman V. S., 2011 *Glaciation in the High Mountains of Siberia*. Elsevier Inc.


Slotte T., 2007 *Evolution of Flowering Time in the (Brassicaceae)*.


Turelli M., Barton N. H., 2006 Will population bottlenecks and multilocus epistasis increase additive genetic variance? Evolution 60: 1763–76.

Turner C., 2000 The Eemian interglacial in the North European plain and adjacent areas. Geol. en Mijnbouw/Netherlands J. Geosci. 79: 217–231.


Whitlock M. C., Barton N. H., 1997 The Effective Size of a Subdivided Population. 441.


## Appendix

Table 1. Comparison of within and between phased subgenome linkage disequilibrium

<table>
<thead>
<tr>
<th>Subgenome</th>
<th>Mean within subgenome $r^2$</th>
<th>Mean between subgenome $r^2$</th>
<th>t statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. grandiflora:</td>
<td>0.45</td>
<td>0.36</td>
<td>341.22</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>C. orientalis:</td>
<td>0.37</td>
<td>0.36</td>
<td>25.38</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. grandiflora:</td>
<td>0.48</td>
<td>0.38</td>
<td>717.11</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>C. orientalis:</td>
<td>0.39</td>
<td>0.38</td>
<td>42.35</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>


Figure S1. Cumulative heterozygosity across each scaffold. Different coloured lines in each plot represent an individual genome.
Figure S2. Illustration of GBS phasing technique. FD sites represent fixed differences in the phased whole genomes.
Figure S3. Pi (A), Watterson’s theta (B), and Tajima’s D (C) between subgenomes in each region.

A.

B.
C.
Figure S4. Minor allele frequency spectra comparing synonymous and replacement polymorphisms in the *C. grandiflora* (A,C) and the *C. orientalis* (B,D) subgenomes.
Table S2. The number of private (A) and shared (B) subgenome and parental SNPs between regions. The first six columns show private SNPS. Names with p are the parental samples.

A.

<table>
<thead>
<tr>
<th>site type</th>
<th>or as</th>
<th>or eu</th>
<th>gr as</th>
<th>gr eu</th>
<th>gr p</th>
<th>or p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0fold</td>
<td>26</td>
<td>42</td>
<td>47</td>
<td>261</td>
<td>48027</td>
<td>157</td>
</tr>
<tr>
<td>intergene</td>
<td>159</td>
<td>189</td>
<td>367</td>
<td>823</td>
<td>213620</td>
<td>1059</td>
</tr>
<tr>
<td>3utr</td>
<td>16</td>
<td>29</td>
<td>18</td>
<td>121</td>
<td>33299</td>
<td>26</td>
</tr>
<tr>
<td>5utr</td>
<td>3</td>
<td>9</td>
<td>6</td>
<td>27</td>
<td>11051</td>
<td>19</td>
</tr>
<tr>
<td>intron</td>
<td>63</td>
<td>117</td>
<td>72</td>
<td>468</td>
<td>132408</td>
<td>126</td>
</tr>
<tr>
<td>stop</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>185</td>
<td>0</td>
</tr>
<tr>
<td>exon</td>
<td>25</td>
<td>42</td>
<td>34</td>
<td>202</td>
<td>43650</td>
<td>86</td>
</tr>
<tr>
<td>istop</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>4fold</td>
<td>15</td>
<td>24</td>
<td>26</td>
<td>139</td>
<td>34064</td>
<td>45</td>
</tr>
<tr>
<td>total</td>
<td>308</td>
<td>453</td>
<td>570</td>
<td>2043</td>
<td>516331</td>
<td>1518</td>
</tr>
</tbody>
</table>

B.

<table>
<thead>
<tr>
<th>site type</th>
<th>gr p</th>
<th>gr as</th>
<th>gr eu</th>
<th>or as</th>
<th>or eu</th>
<th>gr as</th>
<th>gr eu</th>
<th>gr p</th>
<th>or p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0fold</td>
<td>1799</td>
<td>239</td>
<td>304</td>
<td>680</td>
<td>1446</td>
<td>3030</td>
<td>621</td>
<td>2740</td>
<td>921</td>
</tr>
<tr>
<td>intergene</td>
<td>6640</td>
<td>867</td>
<td>1095</td>
<td>2602</td>
<td>7889</td>
<td>16781</td>
<td>3033</td>
<td>9201</td>
<td>3772</td>
</tr>
<tr>
<td>3utr</td>
<td>459</td>
<td>50</td>
<td>73</td>
<td>413</td>
<td>1333</td>
<td>2678</td>
<td>430</td>
<td>1870</td>
<td>675</td>
</tr>
<tr>
<td>5utr</td>
<td>18</td>
<td>34</td>
<td>137</td>
<td>447</td>
<td>970</td>
<td>131</td>
<td>618</td>
<td>238</td>
<td>395</td>
</tr>
<tr>
<td>intron</td>
<td>2476</td>
<td>364</td>
<td>544</td>
<td>2171</td>
<td>6682</td>
<td>13228</td>
<td>2048</td>
<td>9348</td>
<td>3239</td>
</tr>
<tr>
<td>stop</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>17</td>
<td>4</td>
<td>14</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>exon</td>
<td>1556</td>
<td>223</td>
<td>247</td>
<td>720</td>
<td>1607</td>
<td>3481</td>
<td>662</td>
<td>3505</td>
<td>989</td>
</tr>
<tr>
<td>istop</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4fold</td>
<td>1086</td>
<td>124</td>
<td>169</td>
<td>523</td>
<td>1312</td>
<td>2789</td>
<td>527</td>
<td>2839</td>
<td>891</td>
</tr>
<tr>
<td>total</td>
<td>14183</td>
<td>1887</td>
<td>2467</td>
<td>7251</td>
<td>20725</td>
<td>42977</td>
<td>7456</td>
<td>30136</td>
<td>10730</td>
</tr>
</tbody>
</table>
Table S3. Point estimates from demographic inference using *fastsimcoal*

### *C. grandiflora*

<table>
<thead>
<tr>
<th>migration</th>
<th>NAS</th>
<th>NEU</th>
<th>TDIV</th>
<th>migr12</th>
<th>migr21</th>
<th>RAS</th>
<th>REU</th>
<th>MaxEstLhood</th>
<th>MaxObsLhood</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>28622</td>
<td>5870411</td>
<td>389419</td>
<td>2.81E-05</td>
<td>2.38E-05</td>
<td>-7.97E-08</td>
<td>-7.15948E-07</td>
<td>-344301</td>
<td>-282275.09</td>
<td>688617.4</td>
</tr>
<tr>
<td>no migration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAS</td>
<td>NEU</td>
<td>TDIV</td>
<td>RAS</td>
<td>REU</td>
<td>MaxEstLhood</td>
<td>MaxObsLhood</td>
<td>AIC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>188919</td>
<td>778098</td>
<td>61499</td>
<td>-3.78E-08</td>
<td>-1.54E-08</td>
<td>-357172</td>
<td>-282275.088</td>
<td>714356.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### *C. orientalis*

<table>
<thead>
<tr>
<th>migration</th>
<th>NAS</th>
<th>NEU</th>
<th>TDIV</th>
<th>migr12</th>
<th>migr21</th>
<th>RAS</th>
<th>REU</th>
<th>MaxEstLhood</th>
<th>MaxObsLhood</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>39520</td>
<td>7823324</td>
<td>406753</td>
<td>3.23E-05</td>
<td>3.48E-05</td>
<td>-2.62E-08</td>
<td>-2.82451E-07</td>
<td>-400130</td>
<td>-331876.67</td>
<td>800275.8</td>
</tr>
<tr>
<td>no migration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAS</td>
<td>NEU</td>
<td>TDIV</td>
<td>RAS</td>
<td>REU</td>
<td>MaxEstLhood</td>
<td>MaxObsLhood</td>
<td>AIC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>178033</td>
<td>764789</td>
<td>51603</td>
<td>-3.49E-08</td>
<td>-2.88E-08</td>
<td>-357037</td>
<td>-282275.088</td>
<td>714085.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure S5. STRUCTURE cluster membership for individuals plotted on a map of China (A), a PCA based on environmental variables in Chinese populations coloured by their membership to the northern or southern STRUCTURE groups (B), and a PCA based on SNPs coloured by latitude (C).
Table S4. Significant gene ontology term enrichment within CLR outlier windows

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>%</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. orientalis</strong> subgenome Asia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIRSF036788:receptor-like protein kinase</td>
<td>9</td>
<td>8.571429</td>
<td>2.08E-09</td>
</tr>
<tr>
<td>IPR002902:Protein of unknown function DUF26</td>
<td>9</td>
<td>8.571429</td>
<td>2.65E-06</td>
</tr>
<tr>
<td>domain:Gnk2-homologous 2</td>
<td>9</td>
<td>8.571429</td>
<td>6.00E-06</td>
</tr>
<tr>
<td>domain:Gnk2-homologous 1</td>
<td>9</td>
<td>8.571429</td>
<td>6.00E-06</td>
</tr>
<tr>
<td><strong>C. orientalis</strong> subgenome Europe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0016021~integral to membrane</td>
<td>55</td>
<td>18.15182</td>
<td>0.021861</td>
</tr>
<tr>
<td>PIRSF036788:receptor-like protein kinase</td>
<td>13</td>
<td>4.290429</td>
<td>1.10E-11</td>
</tr>
<tr>
<td>IPR002902:Protein of unknown function DUF26</td>
<td>13</td>
<td>4.290429</td>
<td>5.35E-07</td>
</tr>
<tr>
<td>domain:Gnk2-homologous 1</td>
<td>13</td>
<td>4.290429</td>
<td>4.13E-06</td>
</tr>
<tr>
<td>domain:Gnk2-homologous 2</td>
<td>13</td>
<td>4.290429</td>
<td>4.13E-06</td>
</tr>
<tr>
<td>protein kinase</td>
<td>8</td>
<td>2.640264</td>
<td>0.008348</td>
</tr>
<tr>
<td><strong>C. grandiflora</strong> subgenome Europe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPR002182:NB-ARC</td>
<td>15</td>
<td>3</td>
<td>3.69E-06</td>
</tr>
<tr>
<td>GO:0045087~innate immune response</td>
<td>21</td>
<td>4.2</td>
<td>4.10E-05</td>
</tr>
<tr>
<td>IPR000157:Toll-Interleukin receptor</td>
<td>12</td>
<td>2.4</td>
<td>1.05E-04</td>
</tr>
<tr>
<td>GO:0006955~immune response</td>
<td>21</td>
<td>4.2</td>
<td>1.91E-04</td>
</tr>
<tr>
<td>SM00255:TIR</td>
<td>12</td>
<td>2.4</td>
<td>1.49E-04</td>
</tr>
<tr>
<td>GO:0006952~defense response</td>
<td>45</td>
<td>9</td>
<td>2.17E-07</td>
</tr>
<tr>
<td>IPR002182:NB-ARC</td>
<td>15</td>
<td>3</td>
<td>3.69E-06</td>
</tr>
<tr>
<td>GO:0006915~apoptosis</td>
<td>16</td>
<td>3.2</td>
<td>4.13E-06</td>
</tr>
<tr>
<td>Plant defense</td>
<td>26</td>
<td>5.2</td>
<td>9.10E-06</td>
</tr>
<tr>
<td>GO:0012501~programmed cell death</td>
<td>18</td>
<td>3.6</td>
<td>5.19E-05</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----</td>
<td>----</td>
<td>----------</td>
</tr>
<tr>
<td>GO:0008219~cell death</td>
<td>18</td>
<td>3.6</td>
<td>9.47E-04</td>
</tr>
<tr>
<td>GO:0016265~death</td>
<td>18</td>
<td>3.6</td>
<td>9.47E-04</td>
</tr>
<tr>
<td>PIRSF036788:receptor-like protein kinase</td>
<td>10</td>
<td>2</td>
<td>2.38E-05</td>
</tr>
<tr>
<td>atp-binding</td>
<td>70</td>
<td>14</td>
<td>2.36E-04</td>
</tr>
<tr>
<td>nucleotide-binding</td>
<td>76</td>
<td>15.2</td>
<td>2.85E-04</td>
</tr>
<tr>
<td>nucleotide phosphate-binding region:ATP</td>
<td>38</td>
<td>7.6</td>
<td>0.020922</td>
</tr>
<tr>
<td>GO:0001883~purine nucleoside binding</td>
<td>80</td>
<td>16</td>
<td>0.020286</td>
</tr>
<tr>
<td>GO:0030554~adenyl nucleotide binding</td>
<td>80</td>
<td>16</td>
<td>0.020286</td>
</tr>
<tr>
<td>GO:0032559~adenyl ribonucleotide binding</td>
<td>76</td>
<td>15.2</td>
<td>0.021193</td>
</tr>
<tr>
<td>GO:0001882~nucleoside binding</td>
<td>80</td>
<td>16</td>
<td>0.022441</td>
</tr>
<tr>
<td>GO:0005524~ATP binding</td>
<td>75</td>
<td>15</td>
<td>0.025391</td>
</tr>
<tr>
<td>protein kinase</td>
<td>9</td>
<td>1.8</td>
<td>0.028675</td>
</tr>
<tr>
<td>GO:0017076~purine nucleotide binding</td>
<td>85</td>
<td>17</td>
<td>0.032584</td>
</tr>
</tbody>
</table>
Figure S6. Distribution of distances among significant LFMM SNPs and Xt’X and Bayenv outliers in A. Asia and B. Europe
Figure S7. Distribution among significant SNPs in both Bayenv and LFMM and selection scan outliers in A. Asia and B. Europe

A.
Table S5. The number of 1000kb windows on each subgenome around putatively adaptive SNPs with significant population genetic selection signatures.

<table>
<thead>
<tr>
<th></th>
<th>( \pi )</th>
<th>( \Theta_w )</th>
<th>Tajima’s D</th>
<th>Significant windows with a positive Tajima’s D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. grandiflora</em></td>
<td>16</td>
<td>11</td>
<td>17</td>
<td>14</td>
<td>354</td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>17</td>
<td>12</td>
<td>16</td>
<td>14</td>
<td>354</td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>103</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>103</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table S6. Pearson R among phenotypic traits in the common garden in (A) Asia and (B) Europe. The bottom half of the matrix shows Pearson correlation coefficients while the top shows the corresponding p-value.

<table>
<thead>
<tr>
<th></th>
<th>Branch number</th>
<th>Height</th>
<th>Fruits</th>
<th>Mass</th>
<th>Bolting time</th>
<th>Leaf number</th>
<th>Rosette diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branch number</td>
<td>1</td>
<td>0.0024</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.71</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>0.36</td>
<td>1</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.26</td>
<td>0.023</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Fruits</td>
<td>0.66</td>
<td>0.56</td>
<td>1</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.62</td>
<td>0.37</td>
</tr>
<tr>
<td>Mass</td>
<td>0.47</td>
<td>0.73</td>
<td>0.75</td>
<td>1</td>
<td>0.0038</td>
<td>0.46</td>
<td>0.0056</td>
</tr>
<tr>
<td>Bolting time</td>
<td>-0.49</td>
<td>-0.14</td>
<td>-0.58</td>
<td>-0.35</td>
<td>1</td>
<td>0.57</td>
<td>0.46</td>
</tr>
<tr>
<td>Leaf Number</td>
<td>0.05</td>
<td>0.28</td>
<td>-0.06</td>
<td>0.09</td>
<td>0.57</td>
<td>1</td>
<td>0.63</td>
</tr>
<tr>
<td>Rosette diameter</td>
<td>0.07</td>
<td>0.49</td>
<td>0.11</td>
<td>0.33</td>
<td>0.46</td>
<td>0.63</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Branch number</th>
<th>Height</th>
<th>Fruits</th>
<th>Mass</th>
<th>Bolting time</th>
<th>Leaf number</th>
<th>Rosette diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branch number</td>
<td>1</td>
<td>0.024</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.31</td>
<td>0.0072</td>
<td>0.015</td>
</tr>
<tr>
<td>Height</td>
<td>0.26</td>
<td>1</td>
<td>0.0033</td>
<td>&lt;0.0001</td>
<td>0.0036</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Fruits</td>
<td>0.55</td>
<td>0.33</td>
<td>1</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.045</td>
<td>0.13</td>
</tr>
<tr>
<td>Mass</td>
<td>0.53</td>
<td>0.70</td>
<td>0.73</td>
<td>1</td>
<td>0.62</td>
<td>0.011</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Bolting time</td>
<td>-0.11</td>
<td>0.32</td>
<td>-0.49</td>
<td>-0.058</td>
<td>1</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Leaf Number</td>
<td>0.31</td>
<td>0.45</td>
<td>-0.091</td>
<td>0.29</td>
<td>0.75</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Rosette diameter</td>
<td>0.27</td>
<td>0.76</td>
<td>0.17</td>
<td>0.68</td>
<td>0.51</td>
<td>0.62</td>
<td>1</td>
</tr>
</tbody>
</table>