Adaptive genetic variation and gene flow potential in the alpine plant *Arabis alpina*

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by

**DOMINIQUE BUEHLER**

MSc in Env. Sciences (University of Zurich)
born May 31, 1983
from Sigriswil, Bern

Accepted on the recommendation of

Prof. Dr. Alex Widmer, ETH Zurich, examiner
Dr. Felix Gugerli, WSL, Birmensdorf, co-examiner
Prof. Dr. Rolf Holderegger, ETH Zurich and WSL, Birmensdorf, co-examiner
Prof. Dr. Barbara Neuffer, University of Osnabrück, co-examiner

2011
Table of Contents

Summary .................................................................................................................................................. 1

Zusammenfassung .................................................................................................................................. 3

Introduction ............................................................................................................................................ 6

Chapter 1  An outlier locus relevant in habitat-mediated selection .................................................. 17
           in an alpine plant across independent regional replicates

Chapter 2  Concordance of outlier analysis in an independent dataset: ............................................. 41
           does a dominant regional effect expose a habitat-related “outlier”? 

Chapter 3  Using the 454 pyrosequencing-based technique in the ............................................... 62
           development of nuclear microsatellite loci in the alpine plant
           Arabis alpina (Brassicaceae)

Chapter 4  Contemporary gene flow and mating system of Arabis alpina .................................... 71
           in an alpine landscape

Chapter 5  Landscape genetics in plants ............................................................................................ 94

Acknowledgments .................................................................................................................................. 122

Curriculum Vitae .................................................................................................................................. 124
Summary

The chance of a species to persist under changing environmental conditions largely depends on genetic variation, gene flow and, on mating strategies. A sufficient amount of adaptive genetic variation, meaning genes beneficial for adaptation to environmental change, may be already available within populations (standing genetic variation). In turn, gene flow may spread beneficial alleles across landscapes and breeding systems. It may also promote outcrossing and, hence, the reception and sharing of these alleles. In this thesis, I investigated habitat-mediated selection and functional connectivity by gene flow of the alpine rock-cress, *Arabis alpina* (Brassicaceae) to assess the evolutionary potential of alpine plants. *Arabis alpina* is currently being established as a model organism for eco-genomic studies in alpine plants, and it is also a close relative of the model organism *Arabidopsis thaliana*, for which ample genomic resources are available.

In Chapter 1, I aimed to detect genetic markers (outliers) under potential selection for habitat type using *A. alpina* individuals sampled in three habitat types (rock/scree, nutrient rich and moist habitats) across the Swiss and French Alps. Using outlier analysis, one genetic marker was detected to be under putative habitat-mediated selection in both alpine regions separately and across the entire dataset. This genetic outlier locus was considered robust because it was replicated across spatial scales and it did not show genetic structuring possibly confounding the signal of selection. In addition, I sequenced characterized this marker and compared it to the genome of *A. thaliana*. This study shows that a putative outlier locus may be related or even underlying habitat-mediated selection in *A. alpina*, important to further understand the evolutionary potential of this plant species.

Outliers are rarely validated for their signal of selection. In Chapter 2, I aimed to test whether the signals of selection detected at the above outlier locus could be replicated in an independent *A. alpina* dataset. I sampled 30 populations occurring in 3 habitat types across 5 biogeographic regions of the Swiss Alps. Individual haplotypes were genotyped using single nuclear polymorphisms (SNPs) and haplotype frequency was subsequently tested using a general linear model for habitat type and geographic area. I found that habitat type was not detected as driving divergence at this locus. Instead, haplotype frequency was significantly associated with geographic area. This study is one, among a few studies, that attempted to validate a potential
locus under selection in independent datasets. The results show that outlier loci screened in different populations or regions can show inconsistent haplotype frequency distributions. Nevertheless, this approach has the potential to add confidence to genomic regions of putative adaptive significance.

Assessing contemporary gene flow patterns is important to understand the dispersal potential of beneficial alleles across a landscape. In order to test for contemporary gene flow, biparentally inherited nuclear genetic markers were established in Chapter 3. I established 19 polymorphic microsatellite markers which should be useful to estimate contemporary gene flow patterns and to further develop *A. alpina* into a model species for eco-genomic studies. In Chapter 4, I estimated (i) the extent to which current pollen transfer connects *A. alpina* individuals, (ii) the realized mating patterns in this potentially selfing plant species and (iii) the fine-scale genetic structuring in the landscape. I found that *A. alpina* showed substantial genetic structuring but that contemporary gene flow occurs over larger distances ($\leq 1$ km) than previously expected for alpine plant. Thus, *A. alpina* has the potential to spread genes across landscapes.

In Chapter 5, I reviewed the field of landscape genetics in plants. Landscape genetics is an emerging field combining the principles of landscape ecology, population genetics and spatial statistics to investigate how genetic variation is influenced by landscape elements and environmental factors. Applying landscape genetics in plants is promising and can be applied using existing landscape genetic methods on adequate spatial scales.

In conclusion, my results show that habitat-mediated divergence in *A. alpina* may be underlined by a genetic component, but it is currently still challenging to detect the genomic regions under selection. In addition, I have shown that gene flow has the potential to spread genes at distances up to 1 km within a single year. Therefore, gene flow could be more important in population divergence and in the evolutionary fate of alpine plants, than previously expected. The dispersal of genetic variation is especially vital in plants confronted with climate change.
Zusammenfassung


Ausreisser-Marker werden selten auf ihre Verlässlichkeit überprüft. In Kapitel 2 überprüfe ich die Hypothese, dass das AFLP Bandenmuster des genetischen Markers aus Kapitel 1 in einem unabhängigen Datensatz Ähnlichkeiten aufweist. Individuen von *A. alpina* wurden in


General Introduction
Alpine areas and the adaptive potential of alpine organisms

Alpine biomes are characterized by extreme topographic, abiotic and microclimatic heterogeneity on small spatial scales which cause natural isolation of populations (Körner, 2003). Steep environmental gradients exist in the alpine landscape such as increasing altitude which causes shortening of the vegetation period or varying exposition of the terrain which affects soil moisture content, solar radiation or temperature (Körner, 2003). The environmental gradients can drive fine-scale local adaptation in organisms. The alpine flora is found worldwide and its distribution and genetic structure is part of a complex historical interaction with climatic warming and cooling causing retreat and recolonization of populations. Signatures of these historical processes, i.e. refugia and post-glacial expansion routes, are still distinctly imprinted in the genome and reflected in the present-day genetic structure of alpine species (Thiel-Egenter et al., 2010).

Currently, the distribution and persistence of ecosystems is susceptible to global environmental change induced by climate change and land-use change, causing a loss of biological diversity (Vitousek et al., 1997). In alpine areas, certain discernible biological effects are already visible, such as the upward range movements of plants (i.e. migration; Grabherr et al., 1994; Hughes, 2000; McCarthy, 2001; Parmesan, 2006; IPCC, 2007). Besides migration, one possible way for alpine species to adapt to climate change, is to have enough suitable standing genetic variation or to successfully disperse such genetic variation across populations to allow genetic adaptation (Fig 1). Therefore, the likelihood for the survival of populations of alpine species under a changing environment depends to a large part on the available genetic variation, gene flow patterns and the mating strategies (of plants) influencing gene flow (Byars et al., 2009, Fig 1). Since alpine areas are heavily fragmented and have high temporal variation, plants are exposed to selection, expected to result in a large amount of genetic variation (Till-Bottraud and Gaudeul, 2002).

Adaptive genetic variation

Genetic variation is important in all organisms for the adaptation to changing environmental conditions which ultimately results in evolution. Adaptive genetic variation is defined as the variation in genes that affects the fitness of an organism (Holderegger et al., 2006). Adaptive traits have been of interest since Darwin revolutionized the scientific perception of the world
General Introduction

(Darwin, 1859). In Darwin’s theory of evolution, fitness changes that are passed down to an offspring are predicted to make organisms more apt to survive and reproduce in their environment. In present-day evolutionary research, identifying genes underlying such adaptive traits has become a major goal to help scientists understand the evolutionary potential in a species, for instance, under changing climate (Stinchcombe and Hoekstra, 2007; Hoffmann and Willi, 2008).

Alpine plants are especially suited for adaptive studies because of their distribution across varying habitat types in a heterogeneous landscape (Holderegger et al., 2008). Habitat types are recognized to be important drivers of adaptive and genetic population divergence (Turesson, 1922). For instance, differing patterns of snow melt were detected to create habitat-specific phenological patterns between fell field and snow bed plants in an alpine landscape (Kudo and Hirao, 2006). In another study, a glasshouse experiment showed that the size of plants from high- and low-altitude habitats differed significantly even when grown under identical conditions (Pluess and Stöcklin, 2005). However, neither neutral nor adaptive genetic variation is well understood in alpine plants. This is partly due to morphological characters (e.g. leaf shape, size, color, etc.) used to quantify genetic variation which are laborious to measure (Till-Bottraud and Gaudeul, 2002).

Genetic research on alpine plants consists of many phylogeographic studies (e.g. Schönswetter et al., 2005; Thiel-Egenter et al., 2011), but few alpine plants have been studied at the genomic level. Alpine plants are generally considered as non-model species because their genomic set-up is largely unknown. Here, I studied the Alpine rock-cress, Arabis alpina L., a perennial diploid (2n = 16) alpine herb which is widely distributed from the amphi-Atlantic region to the European mountain system (Koch et al., 2006). In the European Alps, A. alpina grows from the montane to alpine zones and prefers cool open habitats found in scree fields, on humus rich soils, on moist cliffs, and along springs and small streams (Hess et al., 1976; Schultze-Motel, 1986). Studies on A. alpina have shown significant inbreeding in the European Alps (Ansell et al., 2008; Buehler et al., 2011). As a wild relative of the model species Arabidopsis thaliana and A. lyrata, A. alpina is currently being developed into an eco-genomic model system for alpine plants (Wang et al., 2009; Poncet et al., 2010). As a result of the genomic resources available from related model species and its vast distribution across
heterogeneous landscapes, *A. alpina* provides a suitable study example to investigate adaptive genetic variation and gene flow.

To measure adaptive genetic variation, genomic regions underlying traits affected by environmental selection can be isolated. Several methods are available to analyze and map genetic markers under potential selection (reviewed in Mitchell-Olds *et al.*, 2007; Reusch and Wood, 2007; Stinchcombe and Hoekstra, 2007; Hoffmann and Willi, 2008). One approach applicable to non-model species is a genome scan followed by outlier analysis. Outlier analysis uses the principle that allele frequencies at genes under divergent selection show higher differentiation among populations under divergent selection than allele frequencies at neutral loci (Storz, 2005; Nosil *et al.*, 2009). In non-model organisms, outlier analysis scans hundreds of anonymous loci (i.e. markers with unknown genomic information such as amplified fragment length polymorphisms; AFLPs) to detect loci putatively under selection. The identified outlier loci are either directly under divergent selection (the rare case) or are linked to a genomic region under divergent selection (the common case; Nosil *et al.*, 2009).

In most studies using outlier analysis, the investigation of putative outlier loci has been conducted without sequence-characterizing the identified outlier loci or the genomic regions in which they are positioned. In addition, the putative loci are rarely verified for the signals of selection as identified by the outlier analysis, and functional validation of outliers is rarely done. This means that outlier loci can represent false positives (i.e. loci which may be associated with neutral processes such as genetic drift, inbreeding, bottlenecks, or gene conversions) instead of reflecting the signals of (past) selection (Teshima *et al.*, 2006; Stinchcombe and Hoekstra, 2007; Excoffier *et al.*, 2009; Buerkle *et al.*, 2011). However, it would be appropriate to isolate genetic markers under potential selection in alpine plants especially in *A. alpina* and validate the signals of selection.

**Gene flow and mating system**

Gene flow is essential for the dispersal of adaptive genetic variation among and within populations (Fig. 1). In the past, gene flow was seen as spatially limited, but new evidence has shown that gene flow occurs at evolutionary significant rates and over large distances (Ellstrand, 1992, 2003). Maintaining gene flow is vital for sessile organisms because isolated populations are at higher risk of losing genetic diversity as a consequence of drift (Reusch and Wood, 2007).
In alpine plants, gene flow has been measured indirectly using inferences from genetic structure based on, e.g. Wright’s (1984) population differentiation $F_{ST}$. This measure essentially gives historical rates of gene flow based on pollen and seed dispersal. Historical gene flow has been detected across large distances in alpine organisms (Paun et al., 2008). Inferences from pollinator flight distance also provide an estimate of pollen dispersal. However, this approach likely underestimates the amount and the distance of pollen and gene dispersal (Bingham and Orthner, 1998). Studies estimating contemporary in contrast to historical gene flow are rare in alpine plants and have been mostly restricted to pollinator observations across short distances at field sites (e.g. Hirao et al., 2006; Stöcklin et al., 2009). To study actual (contemporary) gene flow on a heterogeneous landscape at an ecological timescale, paternity analysis is an appropriate approach (Sork et al., 1999). In paternity analysis, the offspring and their known mother plants along with all potential fathers across an area are sampled and genotyped. Subsequently, the father plant that has sired the offspring of a mother plant is determined using a maximum likelihood approach such as implemented in CERVUS (Marshall et al., 1998). Analyzing contemporary gene flow across an alpine landscape will reveal the realized distances and, hence, the actual potential of gene dispersal of both neutral and adaptive genes.

In addition to gene flow, mating strategies strongly affect the standing genetic variation of populations. In alpine plants, clonal reproduction and/or selfing have been speculated to increase with altitude (Bliss, 1962). However, even few sexually produced and outcrossed offspring will maintain genetic variation in a population (Watkinson and Powell, 1993; Morjan and Rieseberg, 2004). Breeding systems that promote outcrossing also promote pollen movement and gene flow. As a result, the sharing of alleles among populations reduces genetic differentiation among the populations of a species (Morjan and Rieseberg, 2004). However, not much is known about contemporary gene flow in alpine landscape and the influence of the breeding system in spreading beneficial alleles.

**Research approach**

In the context of the subjects introduced above, the objective of this thesis was to study habitat-mediated selection and population connectivity in the alpine plant *A. alpina*. Isolating genetic markers under selection and understanding gene flow patterns and mating strategies is of high priority for alpine plants confronted with global environmental change.
The present project formed part of the collaborative research project “BioChange” of the Competence Center Environment and Sustainability (CCES) of the ETH-domain. BioChange dealt with the question of whether alpine biodiversity has the adaptive potential for meeting the challenges imposed by global environmental change. The goal of BioChange was to use molecular tools and classical transplant experiments to estimate ecological parameters, contemporary evolution, and the maintenance of genetic diversity in natural populations in a variety of species. Thus, the BioChange project extended across several specialized institutions of the ETH-domain.

The scope and ecological questions addressed in my PhD project as part of the BioChange project are elucidated below.

Chapter 1: In this chapter, I performed a series of outlier analyses to detect outlier loci under habitat mediated selection in A. alpina sampled in different habitat types, namely rock/scree, nutrient rich and moist habitats across the Swiss and French Alps. The outlier loci were also tested for population genetic structure indicative of interfering with historical processes such as recolonization after glaciations. Subsequently, I aimed to sequence-characterize and compare the detected outlier loci to the whole-genome sequence of Arabidopsis thaliana and A. lyrata. I then evaluated the potential of outlier loci under habitat-mediated evolutionary divergence for further studies on the adaptation of alpine plants.

Chapter 2: Disentangling selection from neutral processes such as genetic drift, inbreeding, bottlenecks, or gene conversions is difficult (Teshima et al., 2006; Stinchcombe and Hoekstra, 2007; Excoffier et al., 2009; Buerkle et al., 2011). Furthermore, the reliability of outlier analysis has rarely been evaluated in independent datasets (Wiener et al., 2011). In this chapter, I used an AFLP fragment detected in chapter 1 to be an outlier locus for habitat-mediated selection in A. alpina. I addressed the hypothesis that this locus would be verified by the haplotype frequencies screened per habitat type using SNaPshot analysis in an independent dataset of additional populations collected in the Swiss Alps. If so, the locus would be validated for signals of habitat-mediated selection.
Chapter 3: Microsatellite loci are important molecular markers to investigate ecological and evolutionary processes such as gene flow (Selkoe and Toonen, 2006). However, in species that are inbred, isolating microsatellite loci is difficult and may lead to a low success rate. Here, I used the next-generation 454 pyrosequencing technique to isolate and characterize 19 polymorphic microsatellites in *A. alpina*.

Chapter 4: Gene flow is a key factor in constraining or adding genetic variation to populations (Mayr, 1963; Wright, 1984). Maintaining functional connectivity (i.e. gene flow) and genetic diversity is important especially for sessile species in order to decrease the continuous loss of genetic diversity owing to genetic drift. Here, I asked the question whether local occurrences of *A. alpina* were connected by contemporary gene flow by pollen as detected in paternity analysis. This also allowed to study realized mating patterns in an alpine landscape of the Swiss Alps.

Chapter 5: Landscape genetics combines the fields of landscape ecology, population genetics and spatial statistics to understand landscape scale processes such as gene flow and adaptation (Manel et al., 2003). Many landscape genetic studies have been performed on animals. In plants, it is challenging to study landscape scale processes given the particularities of gene flow (e.g. wind-pollinated pollen; animal-dispersed seed; gene dispersal by pollen and seed). In this collaborative study, I reviewed the promises of landscape genetic in plants, but also the reasons why studies on plants are lagging behind those on animals.
Figure 1 Conceptual steps necessary for organisms to adapt to global environmental change and flowchart of the present thesis.
Literature Cited


IPCC. 2007. Climate change 2007: the physical science basis, contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge: Cambridge University Press.


Chapter I

An outlier locus relevant in habitat-mediated selection in an alpine plant across independent regional replicates

Dominique Buehler, Bénédicte N. Poncet, Rolf Holderegger, Stéphanie Manel, Pierre Taberlet, and Felix Gugerli

*Submitted to Evolutionary Ecology*
Abstract

Local environmental conditions and habitat types induce genetic responses in species. Thus, habitat types are strong drivers of adaptive differentiation and evolutionary divergence of populations. Genome scans followed by outlier analysis are powerful tools to improve our understanding of the genetic basis of adaptation. Especially in non-model organisms which lack genomic information, outlier analysis can detect potential genetic loci affected by environmental variation. Based on this approach, we aimed at detecting loci indicative of adaptation for different habitat types in the alpine plant *Arabis alpina*. In a dataset consisting of two independent regional replicates of the European Alps, we consistently detected a single outlier locus indicative of habitat-mediated selection. Subsequently, we sequence-characterized this outlier locus and compared it to the *Arabidopsis* genome. The outlier locus was revealed to be a putative homologue to the SIT4 phosphatase-associated protein family. We consider this locus to be a strong candidate worth further exploration in the habitat-mediated selection and genetic adaptation of natural populations in an alpine plant.

**Keywords:** Adaptation, *Arabis alpina*, Genome scan, Habitat type, Natural selection, Outlier analysis
Outlier locus under habitat-mediated selection

Introduction

The genetic response of a species to its local environment and to different habitat types is of great importance in evolution. Once adapted to different habitat types, populations of a species may experience reduced gene flow and become reproductively isolated, which ultimately results in evolutionary divergence of ecotypes (Rieseberg and Willis 2007; Sobel et al. 2009). Thus, habitat types have been recognized as important drivers of adaptive differentiation and genomic population divergence. For example, the marine gastropod *Littorina saxatilis* has been shown to harbour particular loci which are under divergent selection for contrasting habitat types along coastal gradients (Galindo et al. 2011; Wilding et al. 2001). Likewise, strong adaptation to soil type has been detected in *Arabidopsis lyrata* by mapping candidate mutations showing serpentine soil-mediated selection (Turner et al. 2010). However, it remains a major challenge, especially in natural populations of non-model organisms, to identify genes and genomic regions underlying adaptive traits (Alonso-Blanco et al. 2009) such as involved in adaptation to habitat types.

Genome scans followed by outlier analysis is seen as a major method to improve our understanding of the genetic basis of adaptation (Sgrò et al. 2011; Storz 2005). Outlier analysis is of particular use in non-model organisms since *a priori* information on the genomic background of the study species is not necessary (Storz 2005; Vasemägi and Primmer 2005). Instead, hundreds to thousands of usually anonymous genetic markers are genotyped for each sampled individual (Beaumont and Nichols 1996) and loci that show a higher differentiation than expected under neutrality are detected as outlier loci (Luikart et al. 2003; Storz 2005; Vasemägi and Primmer 2005). These outlier loci are either directly located in adaptive genes or they are only linked to genes and genomic regions under selection (Nosil et al. 2009). Thus far, a large number of outlier loci have been identified across a wide range of organisms and study questions (reviewed in Holderegger et al. 2008; Vasemägi and Primmer 2005). However, the genomic information underlying outlier loci has rarely been characterized (but see Minder and Widmer 2008; Wood et al. 2008), although this is a first step to trace genes or genomic regions under adaptive selection.

The adaptive signal of outlier loci (i.e. especially high population divergence) should be caused by environmental selection. However, historical and demographical processes may cause
substantial spatial genetic structure (e.g. bottlenecks, range expansion, admixture zones) and can thus be confounded with the signal of adaptation at outlier loci (Robertson 1975; Schlötterer 2003). Outlier analysis tends to underestimate these demographical effects, which increases the number of false positives (Excoffier et al. 2009; Hofer et al. 2009). Substantial population genetic structure is especially found in alpine ecosystems since populations diverged in isolation during glacial periods and subsequently came into secondary contact through re-expansion (Alvarez et al. 2009; Körner 2003; Schönswetter et al. 2005; Thiel-Egenter et al. 2010). Under such circumstances, limiting confounding effects on signals of divergent selection and detecting real outlier loci is challenging. This can be achieved by testing for hierarchical population structure (Excoffier et al. 2009) and/or by studying independent regional replicates and restricting confidence to outlier loci detected across regions, based on the reasoning that neutral historical or demographical processes would not create the same genetic pattern across independent replicates (e.g. Manel et al. 2010b; Poncet et al. 2010).

We studied the alpine plant Arabis alpina L. (Brassicaceae), which occurs in distinct habitat types, namely rock/scree, nutrient rich and moist habitats, across two alpine regions in the Swiss and French Alps. A previous study used the same genome scan dataset to search for loci involved in the general response of A. alpina to different environmental variables related to temperature, precipitation and topography using regression analysis (Poncet et al. 2010). Here, our objective was to analyze the dataset using outlier analysis to detect specific genomic regions involved in genetic adaptation to different habitat types.

The strategy of this study was to detect outlier loci by grouping individuals from all sampling locations according to the three habitat types in which they were collected. We performed outlier analyses using these three habitat type groupings within and across two alpine regions. We then selected those outlier loci which were detected in both alpine regions and which did not show patterns of hierarchical population structure. Finally, we characterized the molecular basis of the most consistent outlier locus and compared it to the whole-genome information of the model species Arabidopsis thaliana.
Materials and Methods

Study species, study area and AFLP genotyping

The alpine rock-cress *Arabis alpina* L. (Brassicaceae) is a perennial rosette herb. It has a broad distribution ranging from the amphi-Atlantic region to the European mountain system (Koch et al. 2006). In the European Alps, *A. alpina* grows between 400 m and 3200 m a.s.l. on calcareous substrates in various open habitats (Titz 1971). It reproduces sexually but presumably with a substantial rate of inbreeding (Ansell et al. 2008), or vegetatively via stolons (Schultze-Motel 1986). *Arabis alpina* is a wild relative of *Arabidopsis thaliana* for which whole-genome sequence information is available (The Arabidopsis Initiative 2000).

We used an existing amplified fragment length polymorphism (AFLP) dataset on *A. alpina* (Herrmann et al. 2010; Poncet et al. 2010). Samples from 192 sites in two regions (i.e. Swiss and French Alps) were collected at elevations ranging from 440 m to 3133 m a.s.l. in summer 2006 (Fig. 1). Samples originated from three different habitat types: rock/scree, nutrient rich and moist (Table 1). Rock/scree habitats are found in large scree fields that are dynamic, dry or with irregular water availability and low humus content. The nutrient rich habitat is characterized by high nutrient content and is naturally found in the alpine ecosystem under exposed rocks and along ridges frequently visited by wild animals. Nutrient rich habitats also occur at anthropogenically influenced sites such as alpine pastures near cattle farms. The moist habitat is defined by high water availability and high humus content, it is mostly found along small streams. At each sampling location, fresh leaf material was collected from three, at some locations from nine individuals and immediately dried in silica gel.

Initially, 2386 AFLP loci were genotyped with 19 primer/enzyme combinations (Herrmann et al. 2010; Poncet et al. 2010). AFLP loci were automatically selected according to the stringent procedure used by SCANAFLP (Herrmann et al. 2010). AFLP loci with low reproducibility or minor polymorphisms (i.e. < 3 individuals with a different presence/absence score than all other samples) were discarded. Linkage disequilibrium was detected for only 3.5% of loci pairs (Poncet et al. 2010). The final dataset consisted of 825 polymorphic loci in 634 individuals.
Outlier detection for habitat type

To detect outlier loci, we searched for loci that exhibited higher genetic differentiation (i.e. divergent selection) among habitat types (N = 3) than expected under neutrality with the program DFDIST (Beaumont and Balding 2004; Beaumont and Nichols 1996). DFDIST uses coalescent simulation under Wright’s (1951) symmetrical island model. A neutral distribution of$F_{ST}$values is generated conditional on expected heterozygosity ($H_e$) and based on thousands of simulated loci with a trimmed mean$F_{ST}$identical to the mean empirical$F_{ST}$. This null distribution is used to separate outlier loci from neutral loci based on confidence intervals. For each analysis, DFDIST was configured to generate 50 000 loci and $N_e$ was set to 1 000. We designated those loci as outlier with an observed$F_{ST}$value higher than the upper confidence limit ($P = 0.05$).

We first performed the above DFDIST analysis for habitat types separately in the Swiss and French Alps, and secondly in the cumulative dataset combining the samples from the two alpine regions. The number of loci used, as well as the number of individuals per habitat type in the Swiss and French Alps and in total IS given in Table 1. We considered loci to be consistent outlier loci for habitat type if they were identified in both regions and in the cumulative analysis.

Identification of hierarchical population genetic structure

In order to detect hierarchical population genetic structure, we first applied the Bayesian clustering in STRUCTURE v2.3.1 modified for dominant data (Falush et al. 2003, 2007; Pritchard et al. 2000). A previous analysis by Poncet et al. (2010) detected two clusters in the whole dataset ($F_{ST} = 0.1652$), namely the Swiss and the French Alps. Based on this observation, we considered the Swiss Alps and the French Alps to represent two independent regional replicates. We thus performed STRUCTURE analyses separately for each of the two alpine regions to detect hierarchical genetic structure within these regions. All individuals per region were first grouped into groupings detected in a previous study with an independent dataset (Alvarez et al. 2009). Individuals were clustered into K discrete clusters (K = 1-6) with the admixture model using the population groupings as location prior. We performed three independent runs for each K, with a burn-in period of 50 000 cycles and 50 000 Markov Chain Monte Carlo replications. Next, analysis of molecular variance (AMOVA) as implemented in ARLEQUIN v3.1 (Excoffier et al. 2005) was carried out to test for genetic differentiation among STRUCTURE groups using 1 000 permutations. AMOVA computes$\Phi_{ST}$, an analog of$F_{ST}$statistics (Excoffier et al. 1992).
Identification of hierarchical population structure in the outlier loci detected

To test whether the outlier loci we detected showed hierarchical population genetic structure, we performed locus-specific $\Phi_{ST}$ (Excoffier et al. 2005) for all 825 polymorphic loci. We determined three separate $\Phi_{ST}$-values per locus, namely (1) between the Swiss Alps and the French Alps to identify loci with large-scale population genetic structure, (2) between STRUCTURE groups in the Swiss Alps and (3) between STRUCTURE groups in the French Alps. The latter two should detect loci with small-scale population genetic structure. Significance was estimated with 1 000 permutations. Loci significant at $P = 0.05$ were designated as loci exhibiting significant hierarchical population genetic structure. Note that outlier loci showing hierarchical population signals are not necessarily non-adaptive, as variation in hierarchical population structure may correlate with a regional or large-scale environmental gradient (Holderegger et al. 2008).

Sequencing of outlier locus

Only one locus was identified as a consistent outlier locus across all outlier analyses and was also not affected by hierarchical population genetic structure (see Results). For this locus EM74.7, we generated AFLP genotypes using the procedures in Herrmann et al. (2010) for eight samples from different habitat types: four samples that lacked the corresponding AFLP band and four samples with the AFLP band present. Genomic DNA was digested using EcoRI/MseI and the selective PCR was done using primers EcoRI 5′-GACTGCGTACCAATTC with selective bases ATC and MseI 5′-GATGAGTCCTGAGTAA with selective bases CAC. The AFLP locus EM74.7 was then isolated using the procedure of Roden et al. (2009). Electrophoresis of AFLP bands was done on Spreadex® EL 500 mini gels (Elchrom Scientific, Cham, Switzerland) in 30 mM Tris-acetate EDTA (TAE) buffer using 10 μL of selective PCR product and M3 Marker (Elchrom Scientific, Cham, Switzerland) as size standard. After pre-electrophoresis for 10 min at 50 V, samples were loaded and run for 82 min at 120 V. The gel was stained with SYBR GOLD (Clare Chemical Research, Dolores, CO, USA) for 30 min, and the bands were viewed in an Epi Chemi II Darkroom (UVP Laboratory Products, Upland, CA, USA). We excised locus EM74.7 from samples with the amplified band using a cylinder of 1 mm diameter. As a control, two longer fragments were also excised from the gel. Each gel cut
was eluted in 25 µL deionized water for 24 h at 4 ºC. We repeated this procedure and excised locus EM74.7 from a second Spreadex® gel.

The excised DNA fragments were amplified using primers designed for the adaptor sequences (EcoRI + 0 and MseI + 0) as described in Roden et al. (2009). The PCR products were purified using the MinElute Kit (QIAGEN, Hilden, Germany), and sequencing reactions were performed in both directions with BigDye® Terminator Cycle Sequencing Kit v1.1 (Applied Biosystems, Foster City, CA, USA). Sequences were determined on an ABI 3130 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Alignment of sequences was done using CLC SEQUENCER VIEWER (CLC bio, Aarhus, Denmark), and the obtained consensus sequence was BLASTed against the nucleotide database of GenBank (Altschul et al. 1990).

Results

Outlier analysis

The DFDIST analyses led to the detection of 34 outlier loci (7.7% of all loci) in the Swiss Alps, 37 outlier loci (7.4%) in the French Alps and 32 outlier loci (6.1%) in the cumulative dataset of the Swiss and French Alps combined (Fig. 2). Thus, the percentage of outlier loci detected was consistent with previous genome scan studies (Bonin et al. 2006). Two loci, EM74.7 and PM179.6 were the only outlier loci detected in both alpine regions (Fig. 2, 3a). Locus EM74.7 was also detected as an outlier locus in the cumulative dataset. Furthermore, locus EM74.7 consistently had a higher frequency of fragment presence in moist habitats in all analyses (Fig. 4).

Hierarchical population structure and outlier loci with hierarchical population structure

STRUCTURE detected two clusters in the Swiss Alps, namely the southeastern Swiss Alps and the central/northeastern Swiss Alps (\(\Phi_{ST} = 0.1318, P < 0.001\)). These two clusters only partially matched the groupings created by Poncet et al. (2010). In the French Alps, three clusters were detected, namely the two mountain massifs Vercors and Chartreuse and the region Brionçonnais (\(\Phi_{ST} = 0.2812, P < 0.0001;\) Figs. 1, S1).
We found significant large-scale hierarchical population structure for a total of 113 loci (13.7% of all loci; threshold value $\Phi_{ST} = 0.25$). This reduced the number of confident outlier loci in each dFdist analysis (Fig. 3b), but neither affected locus EM74.7 nor locus PM179.6.

In the next step, we detected 104 loci (12.6% of all loci; threshold value $\Phi_{ST} = 0.15$) with small-scale population genetic structure among the STRUCTURE groups within the Swiss Alps. This reduced the number of confident outlier loci in the Swiss Alps, but also did not affect loci EM74.7 and PM179.6 (Fig. 3c). Finally, we detected 110 loci (13.3% of all loci; threshold value $\Phi_{ST} = 0.30$) with small-scale population genetic structure between the STRUCTURE groups in the French Alps, reducing the number of confident outlier loci in the French Alps, but again not affecting loci EM74.7 and PM179.6 (Fig. 3c).

**Outlier sequence and its localization in the Arabidopsis genome**

We sequenced 41 bp (78 bp including selective bases and both adaptors) of locus EM74.7 (GenBank accession no. HM594277.1), which was the only AFLP locus consistently detected as an outlier locus in all analyses. In a nucleotide BLAST search, locus EM74.7 gave four significant hits (E-value < 0.16; note that short sequences often have relatively high E-values).

The sequence of locus EM74.7 matched with the SIT4 phosphatase-associated family protein in *Arabidopsis lyrata* (GenBank accession no. XM_002890826.1, max. score 42.1, max. identity = 96%). In *Arabidopsis thaliana*, the outlier sequence also matched with the SIT4 phosphatase-associated family protein (At1g30470; GenBank accession no. NM_102783.4, max. score 42.1, max. identity = 96%) as well as with a full length cDNA sequence associated with gene At1g30470 (GenBank accession no. BX818019.1, max. score 42.1, max. identity = 96%), and a complete BAC sequence on chromosome 1 (GenBank accession no. AC009917.2, max. score 42.1, max. identity = 96%). By aligning these sequences, regions of high conservation were revealed among species. Figure 5 illustrates the alignment of the short sequence of locus EM74.7 in *Arabis alpina* and the SIT4 phosphatase-associated family proteins in *Arabidopsis lyrata* and *Arabidopsis thaliana*. 
Discussion

Species inhabiting different habitat types may be exposed to divergent selection pressures, which could lead to adaptive population differentiation. Genome scans followed by outlier analysis are a feasible method to detect genes or (more likely) genomic regions under selection for habitat types, especially in non-model organisms which lack genomic background information. However, sequence-characterizing outlier loci is a step scarcely done in outlier studies using non-model organisms. In this study, our goal was to detect habitat-mediated selection in the alpine plant *Arabis alpina* by comparing samples from different habitat types across two independent regions of the European Alps. Across a suite of stringent analyses, we found one locus as a consistent outlier locus for downstream analysis, whose sequence matched to a coding region in *Arabidopsis* genomes. In the following, we interpret the environmental link of the consistent outlier locus detected and discuss the stringency and limitations of our study.

Outlier loci

The outlier analyses performed in this study revealed that divergent selection affected 6 – 7.7% of loci among all analyses. Almost all of these outlier loci were limited to one alpine region (either the Swiss or French Alps) and were not replicated across the two alpine regions. This may be a result of (i) most outlier loci showing only weak selection, (ii) selection mainly acting on a local scale or (iii) the detection of false positive outlier loci (Minder and Widmer 2008). On the other hand, outlier loci exhibiting replicated divergence across independent regions or environmental gradients may be considered as convergent (Nosil et al. 2009; Schmidt et al. 2008). Therefore, we applied such a strict criterion to our dataset and only considered those outlier loci detected in both alpine regions under study, taking into account that some potentially relevant outlier loci are discarded due to our stringency. We remained with two consistent outlier loci, EM74.7 and PM179.6. Similar studies have also remained with few consistent outlier loci once replicates across regions or environmental gradients were considered (Miller et al. 2007; Nosil et al. 2009; Oetjen et al. 2010; Poncet et al. 2010). Nevertheless, these consistent outlier loci show the strongest signals of selection and are not likely to be affected by population genetic structure (see below). Therefore, they are best suited for downstream applications such as
sequence-characterization, reciprocal transplant experiments or expression studies (Holderegger et al. 2008).

In our study, we considered locus EM74.7 to be a particularly strong candidate because it was replicated across regions and detected in a cumulative analysis. Moreover, this AFLP marker showed a significantly higher frequency of fragment presence in moist habitat type as compared to the other two habitat types (Fig. 4a). This fit the assumption that an allele identified to be under selection for a particular environmental response such as habitat type is potentially targeted by selection across all sampling sites with different selection intensities or directions (Verhoeven et al. 2008). Therefore, a significant change in allele presence/absence should be detected in alleles of outlier loci among habitat types or along environmental gradients. In contrast, locus PM179.6 was considered of less interest because fragment frequencies were not correlated to a particular habitat type (Fig. 4b). Therefore, we consider locus EM74.7 to be best suited for further downstream analysis.

**Habitat-mediated divergence**

Previous studies using outlier analysis in habitat-mediated selection studies have identified particular loci responsible for local adaptation (Galindo et al. 2011; Keller et al. 2010; Shikano et al. 2010; Wilding et al. 2001). The classical example is the detection of outlier loci in *Littorina saxatilis* for ecotypes with different shell types occurring in distinct habitat types across coastal shores (Wilding et al. 2001). Recently, a genome scan using next-generation sequencing has revealed that functional annotations of contigs containing outlier loci for these *Littorina* ecotypes are coding for shell matrix and muscle proteins (Galindo et al. 2011).

In order to detect the underlying function of locus EM74.7, we, likewise, isolated and sequence-characterized this locus and compared it to the *Arabidopsis* genome. The sequence obtained matched the SIT4 phosphatase-associated family protein (SAPs) in *A. thaliana* and *A. lyrata*. This protein seems to be highly conserved in the Brassicaceae family; however the role of SAPs in plants is unknown. In *Saccharomyces cerevisiae*, SAPs interact with a catalytic subunit (SIT4) of a type 2A-related protein phosphatase (Luke et al. 1996). SIT4 executes the start in the cell cycle influenced by factors such as nutrient limitation or the presence of mating pheromones (Luke et al. 1996). Experiments revealed that SAPs are unable to function in the absence of SIT4 and vice versa (Luke et al. 1996). However, without additional information of the underlying
function or the selection factors involved in *A. alpina*, the role of SAPs in habitat-mediated selection remains speculative.

The selective force behind habitat-mediated selection in *A. alpina* is not likely as clear as in the case of shell shape in *Littorina*. Outlier analysis only tests for population divergence and does not directly determine the particular selection factor acting upon the outlier loci or the linked genomic region (Stinchcombe and Hoekstra 2007; Storz 2005). In another analysis of the same dataset, Poncet et al. (2010) used a regression analysis in an environmental association study, to detect outlier loci correlated to numerous specific ecological factors extracted from eco-climatic GIS layers. This particular approach has the advantage that a putative selection force can be determined such as seasonal rainfall or average minimum temperature. However, the selective agent behind adaptation to rock/scree, nutrient rich, and moist habitat types is most likely too specific and complex to detect in such an analysis. Habitat-mediated selection could involve the interaction of multiple selection factors ranging from water use efficiency or anoxia tolerance to nitrogen uptake or competition (reviewed in Reich et al. 2003). As Turner et al. (2010) speculated in their study of serpentine-soil mediated selection in *Arabidopsis lyrata*, the detection of candidate mutations may be instigated by deleterious mutations causing non-adapted individuals to perish. Similarly, we could argue that non-adapted individuals (individuals with a mutation at locus EM74.7 or the corresponding linked gene) had a lower fitness in moist habitats and were ultimately unable to grow and reproduce. However, follow-up investigations should be conducted to test whether such fitness advantages exist under natural conditions.

**Hierarchical population structure**

To rigorously test the reliability of our consistent outlier loci we attempted to distinguish signals of selection from neutral signals of historical or demographical events causing genetic structure. Historical or demographical signals can mimic selection and lead to the detection of a large number of false positive outlier loci in outlier analysis (Excoffier et al. 2009). Locus EM74.7 and locus PM179.6 in *A. alpina* were the only two AFLP markers detected as consistent outliers across two independent regional replicates. Therefore it is unlikely that the effects of neutral processes caused high population divergence among habitat types in both regions (Bonin et al. 2006; Campbell and Bernatchez 2004; Luikart et al. 2003; Nosil et al. 2009; Storz 2005; Vasemägi and Primmer 2005). As a second test, we *a posteriori* searched for substantial
hierarchical genetic structure at a small and a large-spatial scale in identified outlier loci. This differs from recently developed methodologies for outlier detection, which a priori test for population structure (Excoffier et al. 2009). As expected, a large number of outlier loci showed confounding hierarchical population structure at both scales. However, the strongest candidates, locus EM74.7 and locus PM179.6, were not affected by hierarchical genetic structure. This confirms that neutral processes, as far as we could test, were not interfering with the signals of selection at these outlier loci.

**Conclusion**

Identifying loci linked to genomic regions potentially under selection is still part of the exploratory stage in the work to find genes underlying relevant adaptive genetic variation (Manel et al. 2010a; Reusch and Wood 2007). Although we have gone one step further than most previous investigations in this field by sequence-characterizing the outlier locus of interest, we are still limited by the functional proof of the evolutionary relevance of the genomic polymorphisms detected. In theory, the underlying genomic position of an outlier locus is not assumed to be the direct target of selection, and at this point we lack the experimental evidence to interpret the gene(s) to which this outlier locus is linked. At the same time, we cannot exclude the possibility that the coding region to which our outlier locus matched indeed represents the "needle in the haystack" and is directly involved in habitat-mediated selection. Nevertheless, we believe that we found a strong candidate for further evaluation of adaptive responses in field trials and functional tests at the molecular level.

The detection of a putative outlier locus that is consistently related to habitat types offers the possibility to study adaptive genetic diversity across complex landscapes. Especially with the present concern of climate change, isolating adaptive genes will enable us to measure how beneficial gene variants are distributed and may spread among populations (Holderegger et al. 2010), revealing the adaptive potential of populations and their putative range changes. A next step is to detect the polymorphisms causing the genetic pattern at outlier loci and to confirm the linkage of identified outlier loci to known genes under selection. Final proof should come from testing the adaptive relevance of putative outlier loci in their natural environment. For instance, the allelic frequency of an outlier locus found in a particular environmental setting could be verified across a set of independent natural samples. In conclusion, this study represents a first
step towards understanding the molecular basis of habitat-mediated selection for the alpine plant *A. alpina*, and we consider locus EM74.7 indicative of a genomic region of high interest in this species’ adaptation to habitat type worth further genomic exploration.

**Acknowledgments**

We would like to thank René Graf, Conny Thiel-Egenter, Annina Bürgi, Nathalie Baumgartner, Fabio Rimensberger, Rolland Douzet, Serge Aubert, Ludovic Gielly, Delphine Rioux and Claire Redjadj for help in sampling and AFLP genotyping, Sabine Brodbeck for sequence characterization, György Sipos and Christoph Sperisen for advice on the functioning of SIT4 phosphatase-associated proteins, and Sarah Bryner and Debbie Zulliger for their valuable comments on the manuscript. Funding was provided by the CCES-BIOCHANGE project of the ETH domain. S.M. was supported by the Institut Universitaire de France.
Outlier locus under habitat-mediated selection

References


Outlier locus under habitat-mediated selection

Hofer T, Ray N, Wegmann D et al (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73:95–108


Robertson A (1975) Gene frequency distributions as a test of selective neutrality. Genetics 81:775–785


Stinchcombe JR, Hoekstra HE (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100:158–170


**Fig. 1** Map showing the 192 sampling locations of *Arabis alpina* in the European Alps. The two alpine regions, in which outlier analysis was carried out separately, are circled (Swiss Alps shown and French Alps). Different symbols represent STRUCTURE clusters (Falush *et al.* 2003, 2007; Pritchard *et al.* 2000) of the sampling locations within the Swiss and French Alps. Habitat types are given by different shading rock/scree: black, nutrient rich: grey and moist: white.
Fig. 2 Plots of DFDIST analyses (Beaumont and Balding 2004; Beaumont and Nichols 1996) to detect outlier loci that are more strongly differentiated among habitat types than expected under neutrality in *Arabis alpina*. The distribution of $F_{ST}$ values is shown as a function of heterozygosity ($H_e$) for (a) the Swiss Alps, (b) the French Alps, and (c) the cumulative dataset (i.e. the Swiss and French Alps combined). The solid line depicts the 95% confidence interval, and each dot represents a single AFLP locus. Filled dots above the solid line indicate outlier loci indicative of divergent selection, and open circles below the solid line designate neutral loci. Marked with an arrow are the outlier loci EM74.7 and PM179.6.
Outlier locus under habitat-mediated selection

Fig. 3 Venn diagrams showing the effect of significant hierarchical population genetic structure on the number of outlier loci among habitat types in *Arabis alpina*. Each Venn diagram shows the overlap of outlier loci detected in the Swiss Alps, the French Alps, and the cumulative dataset (i.e. the Swiss and French Alps combined). (a) Total number of outlier loci found in all three outlier analyses. (b) Number of outlier loci showing no hierarchical genetic structure on large spatial scale, between the Swiss and French Alps. (c) Number of outlier loci showing no hierarchical genetic structure at small spatial scale among *STRUCTURE* groups within the Swiss and French Alps. (d) Total number of outlier loci showing no hierarchical genetic structure.
Fig. 4 Frequency of presence of (a) locus EM74.7 and (b) locus PM179.6 in *Arabis alpina* across the three habitat types rock/scree, nutrient rich and moist. Crosses: Swiss Alps; Triangles: French Alps; Squares: Cumulative dataset.
Fig. 5 Alignment of sequences of locus EM74.7 in (a) *Arabis alpina* and the putative homologue sequences of the SIT4 phosphatase-associated family protein in (b) *Arabidopsis thaliana* (GenBank accession no. NM_102783.4) and (c) *Arabidopsis lyrata* (GenBank accession no. XM_002890826.1). Nucleotide differences in the sequences are marked with dots.
**Table 1** Number of AFLP loci (N) and individuals of *Arabis alpina* collected for each habitat type (rock/scree, nutrient rich and moist) used in three analyses of the Swiss Alps, the French Alps and the cumulative dataset (i.e. the Swiss and French Alps combined).

<table>
<thead>
<tr>
<th></th>
<th>Number of individuals</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Rock/scree</td>
<td>Nutrient Rich</td>
<td>Moist</td>
</tr>
<tr>
<td>Swiss Alps</td>
<td>443</td>
<td>177</td>
<td>96</td>
<td>71</td>
</tr>
<tr>
<td>French Alps</td>
<td>503</td>
<td>236</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>Cumulative dataset</td>
<td>523</td>
<td>413</td>
<td>126</td>
<td>107</td>
</tr>
</tbody>
</table>

**Fig. S1** Genetic structure as inferred by STRUCTURE (Falush *et al.* 2007; Pritchard *et al.* 2000) of the alpine plant *Arabis alpina* in the Swiss and French Alps. (a) Plotting of the mean maximum likelihood probability of the data as a function of the number of cluster (K) using locprior in the Swiss Alps (triangles) and the French Alps (circles). Individual assignment to the most probable cluster for K=2 and K=3 in the (b) Swiss Alps and (c) French Alps is also given.
Chapter II

Concordance of outlier analysis in an independent dataset: does a dominant regional effect expose a habitat-related “outlier”?

Buehler, D., Brodbeck, S., Holderegger, R., Gugerli, F.

In preparation
Abstract

Outlier analysis has become a common method to detect genomic regions showing genetic differentiation, that are interpreted as signals of divergent selection. However, a validation of such outlier loci in independent datasets has rarely been done. Here, our aim was to validate an outlier locus in *Arabis alpina* previously detected as indicative of habitat-mediated selection across two alpine regions. This outlier locus did not show strong association to population structure in the previous study and was, therefore, considered a strong candidate locus for adaptive divergence. To find concordance of the AFLP band pattern detected in the previous outlier analysis and the haplotype frequency found in an independent dataset, we sampled 30 populations occurring in the same habitat types sampled across five regions of the Swiss Alps. In the present study, habitat type was not detected as a driver of adaptive divergence; instead there was a significant association of haplotype frequency with geographic area. Therefore, we present alternative processes that are potentially underlying the observed pattern. Additionally, we show that inconsistent haplotype frequency distributions might appear when outlier loci are screened in different populations. Despite these restrictions, we can conclude that validating outlier loci in an independent dataset could potentially add confidence to particular loci found in genomic regions of adaptive significance.

**Keywords:** *Arabis alpina*, habitat types, natural selection, SNaPshot
Validation of habitat-mediated outlier locus

Introduction

Outlier analysis has become a common approach to detect genomic regions that show genetic differentiation and, that are potentially under divergent selection (Storz 2005; Vasemägi & Primmer 2005; Stinchcombe & Hoekstra 2007). Unraveling these genomic regions makes it feasible to isolate putative candidate genes important to explore the adaptive genetic diversity and evolutionary potential of populations (Hoffmann & Willi 2008). Due to new technologies, e.g. next-generation sequencing and widely available genomic resources, the detection of candidate genes is now possible also in non-model organisms (Nielsen et al. 2007; Manel et al. 2010a). To date, sets of putative adaptive molecular markers have been identified as outliers, which may be used to characterize adaptive genetic diversity across genomes (e.g. Keller et al. 2010; Shikano et al. 2010; Galindo et al. 2011; Midamegbe et al. 2011). However, the true adaptive significance of these identified loci is difficult and laborious to prove (Reusch & Wood 2007). At the same time, the power and consistency of outlier detection are scarcely evaluated. This means that outlier loci are not validated for the signals of selection and thus, can represent false positives.

An important task in outlier analysis is to discriminate selection from neutral evolutionary processes (Vitalis et al. 2001; Beaumont & Balding 2004; Storz 2005; Manel et al. 2010a). Different factors can interfere with the signals of selection and result in the detection of false outlier loci. These false outlier loci are associated with neutral processes such as genetic drift, inbreeding, bottlenecks or gene conversions (Teshima et al. 2006; Stinchcombe & Hoekstra 2007; Excoffier et al. 2009; Buerkle et al. 2011). Furthermore, local environmental conditions can cause signals of selection independent of a selective force tested in outlier analysis (Shikano et al. 2010). In previous studies, geographic area (Shikano et al. 2010) and hierarchical genetic population structure (Excoffier et al. 2009) were detected as conflicting the signals of adaptive population divergence. Thus, detected outlier loci may be products of an array of neutral and selective processes that are difficult to disentangle.

Independent datasets to validate outlier loci and the signals of selection across different regions and populations are rarely available (Wiener et al. 2011). The few studies using independent datasets to test candidate genes have delivered surprising results. For example, Nachman et al. (2003) found that coat color variation in pocket mice was associated with a
single-gene mutation in one population. However, the mutation did not show the same association in a replicated population. A validation approach, as we propose, could therefore, provide a better understanding of the genetic function of the candidate loci detected in an outlier analysis. It may also help to separate large-scale selective responses from local adaptation.

Previously, we had used a large-scale genome scan of amplified fragment length polymorphisms (AFLPs) to search for outlier loci under habitat-mediated selection in the alpine perennial plant, Arabis alpina, a close relative of the model organism Arabidopsis thaliana (Buehler et al. submitted). In two alpine regions, the Swiss and the French Alps, we had collected plants in rock/scree, nutrient rich and moist habitat types. Using rigorous selection criteria, we detected one AFLP locus, EM74.7, as a consistent outlier locus across the two alpine regions. This outlier locus showed consistently higher AFLP band presence in moist than in rock/scree or nutrient rich habitat types. A test for hierarchic genetic differentiation revealed that the locus was not affected by spatial structure on a small (within the Swiss and French Alps) as well as on a large scale (between the Swiss and French Alps).

In this study, our aim was to find concordance between the AFLP band presence detected in the previous study and the corresponding haplotype frequency detected in an independent dataset of additional populations. Therefore, we sampled 30 populations occurring in the same habitat types across 5 alpine regions in the Swiss Alps. We expected locus EM74.7 to show a genotype pattern associated with habitat types as was detected in the previous study using AFLP markers (Buehler et al. submitted). In this study, single-nucleotide polymorphisms (SNPs) underlying the AFLP pattern were genotyped. Thus, we addressed the hypothesis that the haplotype frequency at locus EM74.7 would be higher in moist habitat types than in nutrient rich and rock/scree habitat types across five regions within the Swiss Alps.

Materials and Methods

Characterization of single nucleotide polymorphisms

Locus EM74.7, previously detected as a consistent AFLP outlier locus (Buehler et al. submitted), was sequenced following the method of Roden et al. (2009). In a next step, we characterized the distinct mutation(s) underlying the AFLP pattern, i.e. we determined the sequences at the AFLP-restriction sites and the adjacent selective bases. In a BLAST search, sequences of A. thaliana
(GenBank accession no. NM_102783.4) and *A. lyrata* (GenBank accession no. XM_002890826.1) were detected as putative homologues to locus EM74.7 (Buehler et al. submitted). Primer pairs were designed in conserved upstream and downstream regions using PRIMER3 (Rozen & Skaletsky 2000). This yielded the forward primer 5’ TCA CAC TAC CTT CTC TGG TTC C 3’, the reverse primer 5’ GCT TGG GTT GAG TGG AGA GA 3’ and a fragment length of 486 bp. To detect single nucleotide polymorphisms underlying locus EM74.7 band presence/absence. We applied standard polymerase chain reaction (PCR) conditions for Sanger sequencing in a selection of 56 individuals of *A. alpina* collected from the former AFLP dataset (Buehler et al. submitted).

**Sampling area and design**

For the independent dataset of the present study, we collected ten sampling locations in summer 2010. Each sampling location consisted of three *A. alpina* populations (total of 30 populations) occurring in distinct habitat types, namely rock/scree, nutrient rich, and moist (Fig. 1; Table S1). In each population we had an equal number of individuals leading to a balanced design. The sampling locations were distributed in five biogeographic regions of the Swiss Alps according to the plant-geographical division of the Central European Alps given in Hess et al. (1976). The alpine regions included the Prealps, the northern Alps, the central eastern Alps, the central western Alps and the southern Alps. The rock/scree habitats were found in rock-scree fields along mountain slopes and are characterized by dynamic conditions such as unstable substrate, low levels of organic matter, and irregular water availability. The nutrient rich habitats are found along alpine pastures or underneath rocky cliffs where nutrients naturally accumulate and are characterized by high humus content and organic fertilization. The moist habitats are found along small alpine streams and are defined by high water availability or even occasional flooding. For more information about habitat type characteristics see Buehler et al. (submitted). We sampled 20 plants in each habitat type at each location at distances of ≥ 2m (600 plants in total). Leaf material was dried in silica gel. DNA extraction was done using the DNeasy 96 Plant Kit (Qiagen, Hilden, Germany) according to the manufacturer’s protocol.
DNA amplification around locus EM74.7

To amplify the genomic region around the selected SNPs, we ran PCRs with the forward and reverse primer described above. PCRs were carried out in a total volume of 10 µL using 1x MM (Multiplex PCR Kit; Qiagen, Hilden, Germany), 0.2 µM of forward and reverse primers, and approximately 1 ng of DNA template. Reactions were amplified as follows on a Veriti Thermocycler (Applied Biosystems, Foster City, California, USA): initial activation for HotStart Taq DNA polymerase at 95 °C of 15 min, followed by 35 cycles of denaturation at 94 °C for 30 s, primer annealing at 56 °C for 90 s and 72 °C for 60 s, followed by a final extension at 72 °C for 10 min. PCR products were purified to remove primers and un-incorporated dNTPs by incubating 1 µL of ExoSAP–IT® (USB, Cleveland, Ohio, USA) in 5 µL PCR product at 37 °C for 15 min followed by 80 °C for 15 min for enzyme inactivation.

Single-base extension SNaPshot® reaction

Primers for SNaPshot® reactions were designed to anneal in the flanking regions of the three SNPs detected (1b, 3f and 4h; Fig. 2). BATCHPRIMER3 (You et al. 2008) was used to design primers on the sense or anti-sense DNA strand with an annealing temperature of 50–60 °C. To test for possible hairpin structures and primer dimers, AUTODIMER (Vallone & Butler 2004) was used. All primers were purified by HPLC to remove incomplete primer synthesis products. The selected primers were combined into a multiplex in which the PCR products ranged from 24 to 42 bp (Table 1). Poly-(T) tails were added to primers to increase the length of the extension product. Single-base extension (SBE) was performed using 2.5 µL SNaPshot® ready reaction mix (Applied Biosystems, Foster City, California, USA), 0.2 µM of each primer, and 1.5 µL of purified PCR product in a 6 µL total volume. SBE reactions were carried out on a Veriti Thermocycler (Applied Biosystems, Foster City, California, USA) with 27 cycles comprising of 96 °C for 10 s, 52 °C for 5 s and 60 °C for 30 s. After the SBE reaction, a post-extension treatment to remove high background signals was performed using 5 µL volume PCR product treated with 0.5 µL of shrimp alkaline phosphatase (SAP; USB Corporation, Cleveland, Ohio, USA) at 60°C for 30 min followed by incubation at 80°C for 15 min for enzyme inactivation. The SBE products were run on an ABI 3130 capillary sequencer (Applied Biosystems, Foster City, California, USA) by mixing 0.5 µL SBE product with 9 µL Hi–Di Formamide and 0.5 µL GeneScan 120LIZ size standard (Applied Biosystems, Foster City, California, USA). Final
Validation of habitat-mediated outlier locus

results were analyzed on GENEMAPPER 3.7 (Applied Biosystems, Foster City, California, USA). We verified the confidence of the multiplex SNaPshot® assay on 20 previously sequenced and AFLP genotyped individuals from the former AFLP dataset. Finally, the multiplex SNaPshot® assay was used to genotype the three SNPs in the independent dataset consisting of 600 individuals.

Statistical analysis

Haplotype frequencies per population were estimated using GENEPOP 4.0 (Rousset 2008). General linear models were used to test the effects of habitat type and region on haplotype frequencies using SPSS 17.0 (SPSS, Chicago, IL, USA) after arcsin transformation, suitable for percentages. Habitat was treated as a fixed factor; region and location were treated as random factors, with location nested within region. The interaction of habitat and region was also tested.

Results

The sequences of the 56 A. alpina individuals from the former AFLP dataset (Buehler et al. submitted) showed several polymorphisms within the sequence of locus EM74.7 (Fig. 2). At the 5’ end, there was a polymorphism in one of the selective bases (T-G). At the 3’ end, there was a two-base pair polymorphism in the selective bases (A-G, C-T). These polymorphisms were linked, meaning that all individuals with band presence displayed the same mutations. A few individuals also had a polymorphism in the MseI restriction site (T-G). All of these polymorphisms are potential candidates accounting for AFLP band presence/absence. In addition, there was also a polymorphism within the fragment sequence (T-C; Fig. 2). Sequence variants are available from GenBank (accession nos. HM594277–HM594279).

In the newly sampled independent dataset of populations from the Swiss Alps, we amplified three of the sequence-characterized mutations using SNaPshot®. Two mutations were linked (1b and 3f) and one mutation was not linked (4h; Fig. 2). The electrophoretic mobility of the SBE products as determined by the automated sequencer was slightly different to the actual size of the products. However, the spacing between SBE primers was large enough to obtain clearly separated peaks.
In the independent dataset, two loci, 1b and 3f, were biallelic and showed similar haplotype frequencies in all samples due to linkage (T: $H_E = 0.267$; G: $H_E = 0.267$ respectively; Table 1). Locus 4h however, was monomorphic (T: $H_E = 1.000$, Table 1) and not linked with SNP loci 1b and 3f. The general linear model analysis did not detect a significant effect of habitat type on haplotype frequencies at locus EM74.7 ($P = 0.191$; Table 2). Instead, the effect of the biogeographic regions was significant ($P = 0.038$). Especially, the regions central western Alps and southern Alps showed deviations in allele frequencies from all the other regions (Fig. 3). The interaction of habitat and region, however, did not have a significant effect on haplotype frequencies at locus EM74.7 ($P = 0.166$; Table 2). See Table S1 for AFLP fragment frequency in each population.

**Discussion**

*Inconsistent validation of outlier locus*

The use of outlier analysis to detect candidate loci potentially under adaptive selection has become a widely applied method (Storz 2005; Vasemägi & Primmer 2005; Stinchcombe & Hoekstra 2007). Still the extent to which these outlier loci can be verified has remained largely unexplored. Here we attempted to validate an outlier locus, EM74.7, previously identified in an outlier analysis among habitat types in two regions of the European Alps (Buehler et al. submitted). By using an independent dataset sampled across the Swiss Alps, we expected to find the same pattern in haplotype frequencies as previously detected in the AFLP band presence of the outlier analysis.

In the former study of Buehler et al. (*submitted*) locus EM74.7 had shown a higher affinity for moist habitat types (Fig. 4) than for the nutrient rich or rock/scree habitat types across the Swiss and French Alps. The result of the present study, however, did not support this pattern. Instead, the haplotype frequencies displayed a significant association with the regions ($n = 5$). This suggested that geographic area may be a strong driver of population divergence at locus EM74.7, thus indicating a contention to the true significance of the outlier locus. Similarly, the previously mentioned study by Nachman et al. (2003) investigating a candidate gene for coat color in rock pocket mice failed to find concordance between populations from Arizona and New Mexico. The authors conclude that coat color might be determined by different genes in different
regions. Such examples are valuable since validating the reliability of candidate loci is rarely done (Wiener et al. 2011), especially for outlier analysis.

In this study, the pattern of haplotype frequencies detected in the independent populations could not be attributed to distinct signatures of natural selection. Different forces of neutral and adaptive selection affect the detection of putative outlier loci and may cause an excess of false positives (Excoffier et al. 2009; Nosil et al. 2009). Therefore, Buehler et al. (submitted) applied rigorous selection criteria to identify locus EM74.7 as a habitat-mediated outlier loci. For one, the locus was detected as an outlier across two alpine regions. In addition, the locus did not show genetic structuring possibly confounding the signal of selection. Nevertheless, this study could not confirm the expected genotype patterns within different habitat types. Several explanations are plausible in interpreting the differential distribution of alleles among habitat types and, at the same time, the affinity of locus EM74.7 to the biogeographic regions. Below, we will evaluate and discuss five potential explanations in more detail.

Potential causes for the observed haplotype frequency pattern

First, local adaptation could cause locus EM74.7 to be indicative of habitat mediated selection only in the regions formerly sampled in the AFLP dataset. Plant populations often adapt to local environmental conditions driving the evolution of local genotypes and ecotypes (Linhart & Grant 1996; Joshi et al. 2001). In this study, we detected that one of the 10 sampling locations, located in the region previously sampled in the AFLP dataset, showed the same genotype frequency as in the outlier analysis. However, in order to confirm local adaptation, more sampling locations would need to be sampled in the original area. Furthermore, classical transplant experiments should be performed to effectively prove that local adaption is causing habitat mediated selection (Holderegger et al. 2008).

Second, if locus EM74.7 was not the target of selection but linked to a gene under selection, it is possible that by crossing-over and recombination between locus EM74.7 and the genomic target of selection, different EM74.7 haplotypes were favored in the habitat types. This could explain the inconsistency we detected in genotype frequencies. The recombination between the outlier locus and the selective gene would result in the selected alleles changing the allele frequencies at the outlier locus in opposite directions in the different habitat types. Thus, respective haplotypes could occur in any of the three habitat types depending on where and when crossing-over occurred between the locus and the genes under selection. In contrast, the genes
under selection would still present a consistent pattern among habitat types. However, this explanation would only seem plausible if a significant effect of the interaction between habitat and region was detected, this was not the case in this study.

Third, locus EM74.7 may represent a false positive (neutral) outlier locus, since the haplotype frequencies detected showed high division in geography or demography. An association of allele frequencies with geographic area can arise from restricted gene flow or other neutral processes such as bottlenecks and range contraction/expansions during glaciation cycles, or by a combination of such factors (Templeton et al. 1995). Alpine ecosystems are known to show distinct population genetic structures since populations diverged in separate glacial refugia and came into secondary contact through re-expansion (Körner 2003; Schönswetter et al. 2005; Alvarez et al. 2009; Thiel-Egenter et al. 2010). The previous outlier analysis did not detect a spatial genetic structure at locus EM74.7 by testing for hierarchical population structure within and between alpine regions (Buehler et al. submitted). However, the phylogeographic history of species is complex and in most cases unknown (Excoffier et al. 2009). It is likely that the way in which, genetic structure was previously tested in A. alpina may not have reflected the actual population structuring across the entire Swiss Alps; the area which we considered in the sampling of the present validation study. Thus, it is plausible that locus EM74.7 is not under selection even though it was identified as a habitat-related outlier locus in the previous study.

Fourth, locus EM74.7 could be under adaptive divergence indicating that the regional differences in haplotype frequencies were attributed to ecological gradients which were not tested. The geographic area encompassing a species range can leave distinct signatures of selection in the genome of a population. Especially temperature and precipitation were shown to be determinants of allele distribution in A. alpina (Manel et al. 2010b; Poncet et al. 2010) and in other alpine plants (Manel et al. submitted). These climatic factors are likely to be intermingled with geographic area. Therefore, they could explain the discrepancies in haplotype frequencies found in the present study if these were triggered by unaccounted large-scale environmental gradients. In addition, historical contingencies could also have caused genetic differentiation at locus EM74.7 across geographical subdivision. For instance, ecological factors (e.g. bedrock type), which are historically important for determining refugial survival areas and postglacial migration routes (Alvarez et al. 2009), can profoundly impact the genome. On a similar premise, Shikano et al. (2010) explored the effect of habitat type, marine vs. freshwater and geographic
area on population divergence in nine-spined sticklebacks. The authors found that most loci under selection detected for salinity differences were instead associated with geographic area (Shikano et al. 2010).

In a last scenario, a new mutation may have arisen locally, increasing population subdivision in genomic regions affected by selection (Nielsen et al. 2007). An adaptive allele can arise and locally be fixed in one geographic location, followed by its spread to multiple geographic areas (Nosil et al. 2007). However, since A. alpina experiences high inbreeding in the European Alps (Ansell et al. 2008; Buehler et al. 2011) and, most likely, reduced gene flow, alleles may not spread effectively across geographic areas. Thus, an adaptive allele may have arisen in the southern Alps and may have reached the highest frequencies in the southern and central western Alps but was not spread to the remaining Swiss Alps.

Conclusion

A major drawback of outlier analyses is that outlier loci might show inconsistent selection patterns when screened under different population genetic structures. Here, we demonstrated that a validation approach may help in unraveling the nature of these inconsistencies. We developed several potential explanations accounting for the haplotype frequency pattern independent of habitat types detected at locus EM74.7, which was supposed to be a habitat-specific outlier locus. It seems clear that one or combinations of the five scenarios outlined could describe the pattern observed in this study. However, disentangling the effects of adaptive, demographic or historical factors remains difficult. Unfortunately, we cannot present conclusive evidence proving that one particular of the five scenarios discussed is the most likely.

In addition, detected outlier loci often corresponded to unusual allele frequencies observed in a single, not reproduced case (Buerkle et al. 2011). This leads to outlier loci not confidently determined as potential adaptive genomic targets or linked to genes under selection. Also, experimental proof is seldom provided to unravel the true nature of selection at outlier loci (Manel et al. 2010b). Although, we could not convincingly validate locus EM74.7 in this study, we were able to demonstrate that using independent datasets to test outlier loci is important. Validation can provide a better understanding of the signatures of selection and has the potential to add confidence to outlier loci detected.
Acknowledgments

We would like to thank Alex Bösch and Deborah Zulliger for help in sampling and Sarah Bryner for comments on the manuscript. This work was funded by the CCES-BIOCHANGE project of the ETH domain.
Validation of habitat-mediated outlier locus

References


Validation of habitat-mediated outlier locus


Figure 1. Map of sampling locations of *Arabis alpina* in each of the five regions of the Swiss Alps. Shown are pie charts of haplotype frequencies at locus EM74.7 for each habitat type per sampling site (moist: light grey; nutrient rich: dark grey; rock/scree: black). Different symbols represent different biogeographic regions (▲ Prealps; + northern Alps; ♦ central eastern Alps; ■ central western Alps; ● southern Alps) and abbreviations denote sampling locations (FL: Flendruz; EA: Ebenalp; GR: Grindelwald; KP: Klausenpass; SM: Samnaun; AL: Albula; BA: Bachalp; TA: Täsch; PR: Piora; SB: San Bernardino).
Figure 2. Alignment of sequences in *Arabis alpina* at AFLP locus EM74.7. Shown are sequences (a) in an *A. alpina* individual with the amplified locus EM74.7, and in two *A. alpina* individuals without the amplified locus EM74.7 because of (b) polymorphisms in the selective bases, as well as (c) a polymorphism in the *MseI* restriction site. The restriction sites (*EcoRI* and *MseI*) are underlined and marked at the cutting positions (ʹ), the original selective bases are in bold face, the nucleotide polymorphisms are also marked (∙), the nucleotide polymorphisms leading to amino acid changes are highlighted with grey boxes, the corresponding amino acids are written above the nucleotide sequence, and the three single nucleotide polymorphisms (SNPs) used in the analysis are labelled.
Validation of habitat-mediated outlier locus

**Figure 3.** Haplotype frequency of locus EM74.7 in different habitat types (rock/scree, nutrient rich and moist) in five Alpine regions (Prealps, northern Alps, central eastern Alps, central western Alps, southern Alps). Dashed line indicate rock/scree habitats, dotted line indicate nutrient rich habitats and solid line indicates moist habitats.

**Figure 4.** Frequency of presence of locus EM74.7 as described in Buehler et al. (Submitted) for *Arabis alpina*. Shown are the three habitat types rock/scree, nutrient rich and moist. Crosses signify the Swiss Alps; Triangles signify the French Alps; Squares signify both alpine regions.
Table 1. Single nucleotide polymorphisms (SNPs) at locus EM74.7 in *Arabis alpina* in the independent dataset of the Swiss Alps. Given are the SNP marker names, nucleotide base changes, haplotype frequencies, single-base extension (SBE) primers including the poly(T) tail and primer length (bp).

<table>
<thead>
<tr>
<th>Marker names</th>
<th>Change</th>
<th>Haplotype frequency</th>
<th>SBE Extension Primer</th>
<th>Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b</td>
<td>G</td>
<td>0.733</td>
<td>GATAAAAAGAGTGCAGAGAATTCA</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.267</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3f</td>
<td>A</td>
<td>0.733</td>
<td>(T)26CCAGTTGCAACAAAGTG</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0.267</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4h</td>
<td>A</td>
<td>1.000</td>
<td>(T)17AGGTCTCAGTGTTTTTA</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. General linear model of allele frequency of locus EM74.7 in different habitat types (rock/scree, nutrient rich and moist) of *Arabis alpina*. Tested were habitat, treated as fixed factor, region, treated as a random factor, sampling site nested within region, treated as a random factor, and the interaction of habitat and region. The asterisk indicates significance at $P = 0.05$.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habitat</td>
<td>2</td>
<td>0.657</td>
<td>0.329</td>
<td>2.051</td>
<td>0.191</td>
</tr>
<tr>
<td>Region</td>
<td>4</td>
<td>6.225</td>
<td>1.556</td>
<td>4.865</td>
<td>0.038 *</td>
</tr>
<tr>
<td>Sampling site [Region]</td>
<td>5</td>
<td>1.217</td>
<td>0.243</td>
<td>2.905</td>
<td>0.071</td>
</tr>
<tr>
<td>Habitat x Region</td>
<td>8</td>
<td>1.282</td>
<td>0.16</td>
<td>1.912</td>
<td>0.166</td>
</tr>
</tbody>
</table>
**Table S1.** Sampling locations of *Arabis alpina* in the Swiss Alps. Given are the biogeographic regions, sampling sites, habitat types (M: moist; NR: nutrient rich; RS: rock/scree), spatial coordinates, and AFLP fragment frequency (absence/presence).

<table>
<thead>
<tr>
<th>Biogeographic Regions</th>
<th>Population name</th>
<th>Habitat type</th>
<th>Latitude (°N)</th>
<th>Longitude (°E)</th>
<th>Absence</th>
<th>Presence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prealps</td>
<td>Ebenalp</td>
<td>M</td>
<td>47.269837</td>
<td>9.401633</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>47.277425</td>
<td>9.398223</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>47.279720</td>
<td>9.397660</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Fiendruz</td>
<td></td>
<td>M</td>
<td>46.511263</td>
<td>7.161996</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.518907</td>
<td>7.153864</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.511290</td>
<td>7.144223</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Northern Alps</td>
<td>Grindelwald</td>
<td>M</td>
<td>46.675369</td>
<td>8.066858</td>
<td>0.050</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.661984</td>
<td>8.053571</td>
<td>0.075</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.678983</td>
<td>8.069018</td>
<td>0.025</td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>Klausenpass</td>
<td>M</td>
<td>46.874728</td>
<td>8.833698</td>
<td>0.175</td>
<td>0.825</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.872991</td>
<td>8.828549</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.875378</td>
<td>8.840825</td>
<td>0.025</td>
<td>0.975</td>
</tr>
<tr>
<td>Central eastern Alps</td>
<td>Albula</td>
<td>M</td>
<td>46.586263</td>
<td>9.855319</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.583082</td>
<td>9.842789</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.579764</td>
<td>9.822555</td>
<td>0.175</td>
<td>0.825</td>
</tr>
<tr>
<td></td>
<td>Samnaun</td>
<td>M</td>
<td>46.924740</td>
<td>10.362366</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.929639</td>
<td>10.363697</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.914328</td>
<td>10.357346</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Central western Alps</td>
<td>Täsch</td>
<td>M</td>
<td>46.059497</td>
<td>7.813653</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.058301</td>
<td>7.813489</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.051594</td>
<td>7.828250</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Bachalp</td>
<td>M</td>
<td>46.357452</td>
<td>7.689186</td>
<td>0.225</td>
<td>0.775</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.368931</td>
<td>7.668992</td>
<td>0.250</td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.345633</td>
<td>7.646899</td>
<td>0.850</td>
<td>0.150</td>
</tr>
<tr>
<td>Southern Alps</td>
<td>Piora</td>
<td>M</td>
<td>46.545169</td>
<td>8.726365</td>
<td>0.050</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.539683</td>
<td>8.745035</td>
<td>0.825</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.530891</td>
<td>8.749164</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>San Bernardino</td>
<td>M</td>
<td>46.475167</td>
<td>9.145803</td>
<td>0.225</td>
<td>0.775</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>46.498959</td>
<td>9.170696</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RS</td>
<td>46.486927</td>
<td>9.157168</td>
<td>1.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Chapter III

Using the 454 pyrosequencing-based technique in the development of nuclear microsatellite loci in the alpine plant *Arabis alpina* (Brassicaceae)

Dominique Buehler, René Graf, Rolf Holderegger, and Felix Gugerli

*Published in American Journal of Botany*
ABSTRACT

• *Premise of the study:* Polymorphic microsatellite markers were developed for the inbred alpine perennial plant *Arabis alpina* to infer life-history parameters and measure patterns of contemporary gene flow within populations.

• *Methods and Results:* Using the 454 pyrosequencing technique, 19 microsatellite primer sets were developed for *A. alpina*. The primer sets were tested on 60 individuals sampled from three sub-populations in the Swiss Alps. The primers amplified di- and trinucleotide repeats with two to five alleles per locus.

• *Conclusions:* Previous attempts to isolate microsatellite loci in *A. alpina* using enrichment libraries and cross-amplification were difficult and produced an insufficient number of polymorphic microsatellite loci. In contrast, next-generation sequencing technology was successful in identifying microsatellite repeats in *A. alpina*. These newly developed microsatellite primers will be useful to further develop *A. alpina* into a model species for eco-genomic studies.

**Keywords:** alpine; *Arabis alpina*; gene flow; inbreeding; microsatellites.
INTRODUCTION
In species for which prior genomic data are lacking, microsatellite loci are traditionally developed using enrichment or selective hybridization or by cross-amplifying known microsatellite primers developed for related taxa. These methods can be time-consuming and potentially lead to a low number of polymorphic microsatellite loci (Zane et al., 2002; Castoe et al., 2010). This could particularly be the case in species that are inbred, such as our study species, Arabis alpina L. The low success rate of transferring microsatellite loci in highly selfing species with small effective population sizes is often attributed to the accumulation of mutations (reviewed in Barbará et al., 2007), leading to primer mismatch or unspecific polymerase chain reaction (PCR) amplification. For species in which microsatellite development is considered difficult, next-generation sequencing provides a tremendous advantage over conventional techniques by rapidly supplying a substantial number of potential microsatellite sequences (Csencsics et al., 2010).

The alpine rock-cress, A. alpina L., is a diploid alpine rosette plant with an extensive arctic-alpine distribution. Its growth conditions reflect varying low-competition habitats between 400 and 3200 m a.s.l. (Schultze-Motel, 1986). Arabis alpina is significantly inbred in the European Alps and is at least partially selfed (Ansell et al., 2008). Arabis alpina is a wild relative of the model plant Arabidopsis thaliana (L.) Heynh, both occurring in the Brassicaceae family and classified in the tribe Arabideae. Therefore, it has been suggested to develop A. alpina into a model species due its wide distribution and the availability of rich genetic resources from related species (including A. thaliana and A. lyrata L.; Ansell et al., 2008). Despite the potential of A. alpina as a model species for eco-genomic research, informative codominant genetic markers for high-throughput genotyping are not available to date. Here we report 19 nuclear microsatellite loci for A. alpina developed using next-generation sequencing, which will facilitate studies on the evolution and ecology of this species.

METHODS AND RESULTS
We selected one sample from Pizol, Switzerland (2 226 m a.s.l.; 46.97981 ° N, 9.43640 ° E) to be shotgun sequenced (1/16th run) using a Roche 454 Genome Sequencer FLX with the Titanium Sequencing kit XLR 70 at Microsynth AG (Balgach, Switzerland). The 454 sequencing technique is described in detail in Margulies et al. (2005). Total genomic DNA was extracted
from leaf tissue using the DNeasy Plant Mini Kit (Qiagen, Hombrechtikon, Switzerland) according to the manufacturer’s protocol, and DNA quality was checked by agarose gel electrophoresis. We obtained 40,367 reads with an average read length of 354 bp and a total amount of 14,307,338 bases. We screened all unassembled sequences in fasta files using MSATCOMMANDER 0.8.2 (Faircloth, 2008) accepting dinucleotide repeats of \( \geq 6 \), trinucleotide repeats of \( \geq 4 \), and tetranucleotide repeats of \( \geq 4 \). The program allows for direct primer design using PRIMER 3 (Rozen and Skaletsky, 2000) by searching for microsatellite repeats and primer annealing sites in the flanking regions. A total of 341 repeats were found in the 40,361 reads screened, consisting of 303 dinucleotide, 30 trinucleotide, and eight tetranucleotide repeats. Primers were successfully designed for a total of 55 repeats including 42 dinucleotide, 10 trinucleotide, and three tetranucleotide repeats. We roughly screened the sequences and primers for suitability and searched for possible duplicate sequences using CLC SEQUENCE VIEWER 6.0.2 (CLC bio, Aarhus, Denmark). This reduced the number of potential microsatellite loci to 34 comprising of 28 dinucleotide, five trinucleotide, and three tetranucleotide repeats worth further testing. We used the BLAST algorithm to query the NCBI nucleotide collection to identify putative homologue sequences in Arabidopsis.

We tested 20 samples in each of three subpopulations of A. alpina occurring in Urnerboden, Switzerland (subpopulation 1: 46.89346 ° N, 8.89841 ° E, 1,520 m a.s.l.; subpopulation 2: 46.8960 ° N, 8.9032 ° E, 1,350 m a.s.l.; subpopulation 3: 46.8941 ° N, 8.8979 ° E, 1,760 m a.s.l.) for microsatellite amplification. PCRs were carried out in a final volume of 10 \( \mu \)L containing 1 \( \mu \)L of genomic DNA, 4.6 \( \mu \)L 2 \( \times \) Master Mix (Multiplex PCR kit; Qiagen, Hombrechtikon, Switzerland), and 0.1 \( \mu \)L of each forward and reverse primer (5 \( \mu \)M). Forward primers were labeled with one of the following fluorescent dyes: NED, HEX, or FAM (Applied Biosystems, Foster City, California, USA). We amplified loci on Veriti thermocyclers (Applied Biosystems) with initial 15 min denaturation at 94 ° C, followed by 29 cycles of denaturing at 94 ° C for 30 s, annealing at 57 ° C for 90 s, elongating at 72 ° C for 1 min, and a final extension at 72 ° C for 30 min. PCR products were run with ROX 400 HD as size standard on a 3130xl Genetic Analyzer (Applied Biosystems), and electropherograms were analyzed using GENEMAPPER 3.7 (Applied Biosystems). We targeted only nuclear microsatellite loci, and therefore we discarded loci that were not heterozygous in at least one tested individual. Sequences of the microsatellite loci as they appear in the original sample were
Microsatellites in *Arabis alpina*

deposited in GenBank (Table 1). For all polymorphic loci providing clear electropherograms, we recorded the overall number of alleles per locus, calculated expected and observed heterozygosity (*Ho* and *He*) as well as inbreeding coefficients (*FIS*) and tested for linkage disequilibrium between pairs of microsatellite loci using FSTAT 2.9.3.2 (Goudet, 1995) within each subpopulation.

We identified 19 of the 34 microsatellite loci (16 dinucleotide and three trinucleotide repeats) as polymorphic, which generated consistent amplification products in the expected size range. The loci contained two to five alleles in the 60 individuals tested with *He* and *Ho* ranging from 0.10 to 0.75 and from 0 to 0.65, respectively (Table 2). As expected for a highly inbred species, *FIS* was relatively high for most loci (mean *FIS* in the three subpopulations were 0.35, 0.57, and 0.54, respectively; Table 2). These values are similar to the *FIS* value of 0.59 detected by Ansell et al. (2008) using allozymes. Significant linkage disequilibrium was detected in only four pairwise comparisons (loci 5GTC and A4JW7, DJSE and BWF1, 5GTC and 7PJQ, and between 5GTC and DEET), potentially also reflecting inbreeding effects.

**CONCLUSIONS**

Here, we provide the first set of nuclear microsatellite loci for *A. alpina* and show the utility of next-generation sequencing in establishing microsatellite primers in highly inbred species. Earlier attempts to establish microsatellite loci using an enrichment approach and cross-amplification of microsatellite loci from Brassicaceae species were found to be difficult. These approaches yielded only a low number of successfully amplified microsatellite loci, which ultimately showed limited or, in most cases, no polymorphisms. Thus, we used 454 pyrosequencing data to obtain a large number of potential microsatellite sequences. Using this technology, we easily established 19 polymorphic loci with respective primer sets for multiplex-PCR amplification. These microsatellite loci will be useful to investigate the genetic structure, contemporary gene flow, and realized mating patterns in *A. alpina*. Obtaining more information about this species’ life history and population biology will provide baseline knowledge for the further development of *A. alpina* into an eco-genomic model species.
ACKNOWLEDGMENTS
The authors thank D. Zulliger for comments on the manuscript and the Genetic Diversity Centre, ETH Zurich, for technical assistance. This study was supported by the CCES-BIOCHANGE project of the ETH-domain.
LITERATURE CITED
Table 1. Characteristics of 19 nuclear microsatellite loci developed for *Arabis alpina*. For each locus, the forward and reverse primer sequences, repeat motif, size of the original fragment (bp), annealing temperature when run individually (*T*<sub>a</sub>), GenBank accession number and *Arabidopsis* hit in GenBank are shown.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Primers forward/reverse</th>
<th>Repeat motif</th>
<th>Size (bp)</th>
<th><em>T</em>&lt;sub&gt;a&lt;/sub&gt; [°C]</th>
<th>GenBank Accession no.</th>
<th><em>Arabidopsis</em> hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>4MDH</td>
<td>F: CCAGCCAGAGCACTAAGC</td>
<td>(CT)&lt;sub&gt;10&lt;/sub&gt;</td>
<td>172</td>
<td>59</td>
<td>HQ599541</td>
<td>AC025294.14</td>
</tr>
<tr>
<td></td>
<td>R: TGTATTTCCATCAACTCTGGGAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9VSH</td>
<td>F: GTTGATGCTTTAGCTGGAACC</td>
<td>(CT)&lt;sub&gt;4&lt;/sub&gt; (GT)&lt;sub&gt;9&lt;/sub&gt;</td>
<td>208</td>
<td>59</td>
<td>HQ599542</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: CGATGAGAACGCCTCAAGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJ5E</td>
<td>F: GACCCATCCCTGGACCC</td>
<td>(GT)&lt;sub&gt;12&lt;/sub&gt;</td>
<td>262</td>
<td>59</td>
<td>HQ599543</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: ACAATCTTTGTTGCAAACGTGC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5GTC</td>
<td>F: TGGTGTTAGGTCTCGGC</td>
<td>(AC)&lt;sub&gt;14&lt;/sub&gt;</td>
<td>213</td>
<td>59</td>
<td>HQ599544</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: CGATGAGAACGCCAGAAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A4JW7</td>
<td>F: ACGGCTGAAAGATTTGGTGAG</td>
<td>(GTT)&lt;sub&gt;9&lt;/sub&gt;</td>
<td>276</td>
<td>59</td>
<td>HQ599545</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: TTGGTCCACGTCTTTGGCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3JUY</td>
<td>F: ACCCACAAGTTCGATCTTC</td>
<td>(TC)&lt;sub&gt;9&lt;/sub&gt;</td>
<td>183</td>
<td>59</td>
<td>HQ599546</td>
<td>AB025604.1</td>
</tr>
<tr>
<td></td>
<td>R: AGGTCCACGGAAGATAGCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1T8T</td>
<td>F: TGCTACACCGGGCAAGATG</td>
<td>(AAC)&lt;sub&gt;8&lt;/sub&gt;</td>
<td>226</td>
<td>60</td>
<td>HQ599547</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: CCGACATGTTGTTCTTTGGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3XGR</td>
<td>F: ACCGGGACCTTTGTTCC</td>
<td>(AG)&lt;sub&gt;8&lt;/sub&gt;</td>
<td>254</td>
<td>59</td>
<td>HQ599548</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: GATGCGTCCTGACAGGCGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEET</td>
<td>F: TCATGTCGCCTCATAATTTG</td>
<td>(AT)&lt;sub&gt;12&lt;/sub&gt;</td>
<td>304</td>
<td>60</td>
<td>HQ599549</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: TGGCGGGGAACAGAGAAGCTG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A25GM</td>
<td>F: AGATTCTGGGTTCTGTAATGG</td>
<td>(GT)&lt;sub&gt;8&lt;/sub&gt;</td>
<td>220</td>
<td>59</td>
<td>HQ599550</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: TGTTTCATATAGTGCTTCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6U3A</td>
<td>F: ATCGAAACGGTAAGTGGG</td>
<td>(AT)&lt;sub&gt;8&lt;/sub&gt;</td>
<td>163</td>
<td>59</td>
<td>HQ599551</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: CACGGATATGTTAAGACCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3Q19</td>
<td>F: TGCCCTACGTGGAAGTCCTTC</td>
<td>(AG)&lt;sub&gt;8&lt;/sub&gt;</td>
<td>239</td>
<td>59</td>
<td>HQ599552</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: CACCAGTGAATGTTAAGACCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BWF1</td>
<td>F: TGTGCGGAGTTCGGGCACTC</td>
<td>(AT)&lt;sub&gt;10&lt;/sub&gt;</td>
<td>317</td>
<td>59</td>
<td>HQ599553</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: CTAACCCATGAAATACGGACTGACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7PJQ</td>
<td>F: CGACCGTCAATGTAGTTCCTC</td>
<td>(AC)&lt;sub&gt;4&lt;/sub&gt; (AT)&lt;sub&gt;8&lt;/sub&gt;</td>
<td>223</td>
<td>59</td>
<td>HQ599554</td>
<td>AB005232.1</td>
</tr>
<tr>
<td></td>
<td>R: TGATTCTGCAAGTGTTCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4GHI</td>
<td>F: ACATACGATAAGTTACAGTGG</td>
<td>(AT)&lt;sub&gt;10&lt;/sub&gt; (CA)&lt;sub&gt;5&lt;/sub&gt;</td>
<td>293</td>
<td>60</td>
<td>HQ599555</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: TTTGGTGACTGCTGAGGAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A8E90</td>
<td>F: AAGCGCTATAGAACGTTTGC</td>
<td>(TAA)&lt;sub&gt;7&lt;/sub&gt;</td>
<td>214</td>
<td>59</td>
<td>HQ599556</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: CACTTCCTACTTTTAGTTGGAATGACCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A93Q</td>
<td>F: GGGAATGGAAGGGAAACCCTC</td>
<td>(AG)&lt;sub&gt;8&lt;/sub&gt;</td>
<td>336</td>
<td>59</td>
<td>HQ599557</td>
<td>AC005167.3</td>
</tr>
<tr>
<td></td>
<td>R: CCTAAGCAACGGCAATTTCATC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBS4</td>
<td>F: TGTAGAGGTCATGCGGGAAAG</td>
<td>(AT)&lt;sub&gt;8&lt;/sub&gt;</td>
<td>201</td>
<td>59</td>
<td>HQ599558</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: CACACTCAAGAAACCCAATGCTG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79PO</td>
<td>F: GCCAAACTCTCGGTTCTCCAC</td>
<td>(AT)&lt;sub&gt;10&lt;/sub&gt;</td>
<td>251</td>
<td>59</td>
<td>HQ599559</td>
<td>No hit</td>
</tr>
<tr>
<td></td>
<td>R: ACTAGGGAGCCCTTTTGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Nuclear microsatellite loci screened in three Swiss subpopulations of *Arabis alpina*. For each locus in each subpopulation, the number of alleles observed ($N_A$), observed heterozygosity ($H$), expected heterozygosity ($H_e$), and inbreeding coefficient ($F_{IS}$) are shown. Sample sizes per population are given in parentheses.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Sub-population 1 (N = 20)</th>
<th>Sub-population 2 (N = 20)</th>
<th>Sub-population 3 (N = 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N_A$</td>
<td>$H_o$</td>
<td>$H_e$</td>
</tr>
<tr>
<td>3JUY</td>
<td>2</td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>4MDH</td>
<td>2</td>
<td>0.15</td>
<td>0.36</td>
</tr>
<tr>
<td>5GTC</td>
<td>3</td>
<td>0.60</td>
<td>0.67</td>
</tr>
<tr>
<td>9VSH</td>
<td>4</td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>A1T8T</td>
<td>2</td>
<td>0.05</td>
<td>0.23</td>
</tr>
<tr>
<td>A4JW7</td>
<td>3</td>
<td>0.25</td>
<td>0.62</td>
</tr>
<tr>
<td>DJSE</td>
<td>3</td>
<td>0.10</td>
<td>0.52</td>
</tr>
<tr>
<td>3QI9</td>
<td>3</td>
<td>0.05</td>
<td>0.49</td>
</tr>
<tr>
<td>3X6R</td>
<td>2</td>
<td>0.20</td>
<td>0.39</td>
</tr>
<tr>
<td>6U3A</td>
<td>3</td>
<td>0.05</td>
<td>0.23</td>
</tr>
<tr>
<td>A25GM</td>
<td>4</td>
<td>0.15</td>
<td>0.45</td>
</tr>
<tr>
<td>DEET</td>
<td>4</td>
<td>0.20</td>
<td>0.59</td>
</tr>
<tr>
<td>46HI</td>
<td>3</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>7PJQ</td>
<td>4</td>
<td>0.60</td>
<td>0.63</td>
</tr>
<tr>
<td>BWF1</td>
<td>4</td>
<td>0.25</td>
<td>0.70</td>
</tr>
<tr>
<td>79PO</td>
<td>4</td>
<td>0.20</td>
<td>0.63</td>
</tr>
<tr>
<td>A8E90</td>
<td>4</td>
<td>0</td>
<td>0.66</td>
</tr>
<tr>
<td>A93Q</td>
<td>2</td>
<td>0.20</td>
<td>0.33</td>
</tr>
<tr>
<td>EB54</td>
<td>5</td>
<td>0.20</td>
<td>0.48</td>
</tr>
<tr>
<td>Mean</td>
<td>3.21</td>
<td>0.19</td>
<td>0.45</td>
</tr>
</tbody>
</table>
Chapter IV

Contemporary gene flow and mating system of *Arabis alpina* in an alpine landscape

Buehler D., Graf, R., Holderegger, R., Gugerli, F

*Published in Annals of Botany*
ABSTRACT

• **Background and Aims** Gene flow is important in counteracting the divergence of populations but also in spreading genes across landscapes. However, contemporary gene flow is not well understood across alpine landscapes. The aim of this paper was to estimate contemporary gene flow through pollen and to examine the mating system in the alpine perennial plant, *Arabis alpina*.

• **Methods** An entire sub-alpine to alpine landscape of 2 km$^2$ was exhaustively sampled in the Swiss Alps. We used 18 nuclear microsatellites to genotype 595 individuals and 499 offspring from 49 maternal plants. Contemporary gene flow by pollen was estimated from paternity analysis, matching the genotypes of maternal plants and offspring to the pool of likely father plants. Realized mating patterns and genetic structure were also estimated.

• **Key Results** Paternity analysis revealed several long-distance gene flow events (≤ 1 km). However, most outcrossing pollen was dispersed close to the mother plants and 84% of all offspring were selfed. Individuals that were spatially close were more related than by chance and were also more likely to be connected by pollen dispersal.

• **Conclusions** In the alpine landscape studied, genetic structure occurred on small spatial scales as expected for alpine plants. However, gene flow may also cover large distances. This makes it plausible for alpine plants to spread beneficial alleles at least via pollen across the landscape at a short timescale. Thus, gene flow potentially facilitates rapid adaptation in *A. alpina* likely required under on-going climate change.

**Keywords:** alpine, *Arabis alpina* L., contemporary gene flow, genetic structure, mating system, paternity analysis, pollen dispersal, spatial autocorrelation
INTRODUCTION

Gene flow, the transfer of genes among populations, is a key factor in determining genetic differentiation of populations (Mayr, 1963) and in spreading genes across a landscape (Wright, 1984). In heterogeneous environments, populations of sessile organisms like plants may be naturally isolated. Fragmented populations are more likely to be subject to genetic erosion and loss of genetic diversity owing to genetic drift (Ellstrand, 1992; Frankham et al., 2002). Thus, gene flow is important to maintain a network of connected local populations, thereby increasing a species’ chances of survival.

In alpine ecosystems, the heterogeneity of the landscape and the resulting spatio-temporal isolation of plant populations (Till-Bottraud and Gaudeul, 2002; Körner, 2003) have resulted in contrasting hypotheses of the amount of gene flow. On the one hand, gene flow is considered a rare event because of natural fragmentation of populations which results in genetic divergence and substantial genetic structure of populations. At a large spatial scale, the genetic structure of populations of alpine species has mainly resulted from historical factors such as the location of glacial refugia and subsequent recolonization along post-glacial expansion routes (Körner, 2003; Schönswetter et al., 2005; Alvarez et al., 2009; Thiel-Egenter et al., 2011). These historical processes have left distinct imprints in the genetic structure of alpine species which are still detectable today (Schönswetter et al., 2005; Thiel-Egenter et al., 2011). Even at a smaller spatial scale, the alpine landscape is characterized by topographic, edaphic, and microclimatic heterogeneity which potentially causes population divergence and local adaptation in alpine plants (Körner, 2003). These environmental gradients maintain population genetic structure (Till-Bottraud and Gaudeul, 2002; Körner, 2003) as a consequence of limited gene flow by pollen or seed and mating patterns favoring selfing and clonality. This assumption is supported by the low abundance and short flying distances of pollinators observed in the alpine landscape (Bingham and Orthner, 1998). In addition, clonality and selfing have been shown to increase at higher altitudes, impeding the dispersal of genetic variation but ensuring local population survival (Schröter 1926; Bliss 1962; Bliss 1971; García-Camacho and Totland, 2009).

In alpine plants, low gene flow and the resulting lack of functional connectivity has rarely been proven by empirical data. Recent studies in alpine plants report restricted (Ohsawa and Ide, 2008; Gonzalo-Turpin and Hazard, 2009) to substantial long-distance gene flow events (Raffl et al., 2008). Therefore, it might be possible that gene flow facilitates the dispersal of genes more
Contemporary gene flow patterns in *Arabis alpina*

frequently and at larger distances in alpine landscapes than expected. This latter assumption is in agreement with evidence from non-alpine study organisms suggesting that gene flow occurs over large distances and at evolutionary significant rates (Ellstrand, 1992, 2003). Especially gene flow via pollen has been shown to disperse large distances in non-alpine studies (e.g. Watrud *et al.*, 2004; Kamm *et al.*, 2009). In most studies on alpine species, gene flow was indirectly studied from pollinator foraging distances or directly estimated from pollen movement measured by pollen dye (Ellstrand, 1992). Also, gene flow is estimated from population structure based on genetic differentiation such as Wright’s $F_{ST}$ (Wright, 1984). The parameter $F_{ST}$ gives a historical estimate of gene flow based on pollen and seed dispersal which is insensitive to small changes in allele frequencies in the contemporary landscape (Sork *et al.*, 1999). Therefore, this parameter is not suited to study current patterns of gene flow (Whitlock and McCauley, 1999). In contrast, pollinator foraging distance and pollen movement gives a contemporary estimate of gene flow by pollen, but severely underestimates the real amount of gene movement (Ellstrand, 1992).

To study contemporary gene flow by pollen, segregating molecular markers among parental and offspring populations are suitable to be analyzed. The genotype of the mother plant and offspring are compared with a pool of all potential father plants within a population to detect the most likely pollen parent through paternity analysis (Sork *et al.*, 1999; Holderegger *et al.*, 2010). So far, studies on contemporary gene flow in herbs across natural landscapes are rare, and in alpine plants, studies that have been conducted with paternity analysis or by pollinator observation are limited to very small scales in field sites (e.g. Schmitt, 1980; Hirao *et al.*, 2006; Brunet and Holmquist, 2009; Stöcklin *et al.*, 2009).

Investigating contemporary gene flow is especially relevant in alpine ecosystems because alpine plants are particularly vulnerable to climate change (Theurillat and Guisan, 2001; Till-Bottraud and Gaudeul, 2002; Byars *et al.*, 2007). In alpine organisms, climate change has resulted in specific biological effects such as range contractions and local extinctions (Hughes, 2000; McCarthy, 2001; Parmesan, 2006). By measuring contemporary gene flow, the dynamics of gene movement in environmentally heterogeneous alpine landscapes can be estimated (Sork *et al.*, 1999). This will help in inferring possible changes caused by climate change in present-day processes such as effective mating and local adaptation.

This is the first study to investigate contemporary gene flow by pollen in an alpine plant and in a completely sampled landscape. We chose the alpine perennial plant, *Arabis alpina* L.
Contemporary gene flow patterns in *Arabis alpina*

(Brassicaceae) to describe current gene flow patterns with paternity analysis because of its recent development as an evolutionary model species for alpine plants (Ansell et al., 2008; Poncet et al., 2010; Tedder et al., in press). The objectives of our study were to (1) investigate the extent to which individuals are functionally connected by current pollen transfer, (2) estimate the realized mating pattern, and (3) determine the fine-scale spatial genetic structure of this species in an alpine landscape of approximately 2 km$^2$. For this, we sampled all individuals of *A. alpina* in a sub-alpine to alpine landscape and genotyped adult individuals and their offspring/seeds at 18 nuclear microsatellites.

**MATERIALS AND METHODS**

**Study species**

The alpine perennial *Arabis alpina* (Brassicaceae) is a pioneer plant species with a wide distribution in the northern hemisphere ranging from eastern Canada to the Ural Mountains and tropical East Africa (Koch et al., 2006). In Central Europe, *A. alpina* occurs in the montane to alpine zones in calcareous open habitats (Schultze-Motel, 1986). It is found on scree fields, along small streams and wells, in moist ravines and on humus-rich rock floors. *Arabis alpina* is described as diploid (2n=16) and reproduces sexually as well as asexually by stoloniferous growth (Schultze-Motel, 1986). It has small white flowers frequently visited by insects from various groups and its seeds are about 1 mm in size with wings extending around the edges, potentially facilitating wind dispersal (Schultze-Motel, 1986). Previous studies showed that *A. alpina* is self-compatible (Tedder et al., in press) and highly selfing (Ansell et al., 2008; Buehler et al., 2011)

**Study area and sampling**

Our study area was located on a sub-alpine to alpine landscape of approximately 2 km$^2$ in Urnerboden, Switzerland (Fig. 1). This landscape encompasses a riverbank, a small ravine and a scree field. Along the riverbank at around 1350 m.a.s.l., *A. alpina* occurs frequently over a stretch of about 700 m. The small ravine is located along a mountain side extending from 1350 m to 1700 m.a.s.l., over a length of about 1 km. In the rock/scree field at 1750 m.a.s.l., *A. alpina* occurs on dry scree as well as on nutrient-rich alpine pastures surrounding the scree field. In this landscape, *A. alpina* grows as scattered individuals or in small groups with varied degrees of spatial isolation. We exhaustively mapped and sampled all *A. alpina* plants in summer 2008. Fresh leaf material was
collected from 595 adult plants and was immediately stored in silica gel. Open-pollinated seeds were collected from all fruiting adult plants. The collected seeds were stored at room temperature until DNA extraction. We mapped the location of individuals and the relative distance to neighboring plants by hand and recorded coordinates every 10 m as well as separately for every mother plant from which we sampled seeds using a hand-held GPS receiver. For paternity analysis, we randomly selected 49 mother plants, we genotyped all adult plants and ten seeds per mother plant (N = 499 seeds).

**DNA extraction and microsatellite genotyping**

To genotype adult plants, DNA was isolated from 10 mg of lyophilized material grinded using a disruptor mill (Retsch, Haan, Germany). Total genomic DNA was extracted following the DNeasy 96 plant kit protocol (Qiagen, Hombrechtikon, Switzerland). For paternity analysis, embryos and cotyledons were excised from seeds after placing them in H2O at room temperature for 12 hours. Disruption and DNA extraction of the seeds were done as described for adult plants.

We used 18 of the 19 polymorphic nuclear simple sequence repeats (nSSR) markers (excluding locus 3X6R) described in Buehler *et al.* (2011). nSSr loci were amplified in four multiplex-PCRs (Table 1). Multiplex PCRs with fluorescent dye-labeled primers were performed in 10-µL volumes containing 1-2 ng DNA, 4.6 µL 2x Master Mix (Multiplex PCR Kit, Qiagen) and 0.1 µL of each primer (5 µM). PCRs were performed on Veriti thermocyclers (Applied Biosystems, Foster City, USA). Initial denaturation was set at 94 ºC for 15 min, and was followed by 27–29 cycles (multiplex 1 & 3: 28 cycles; multiplex 2: 29 cycles; multiplex 4: 27 cycles) at 94 ºC for 30 s, 57 ºC for 90 s, 72 ºC for 30 s, ending with a final extension at 72 ºC for 30 min. PCR products were run with ROX 400 HD as internal size standard (Applied Biosystems, Foster City, USA) on a 3130xl Genetic Analyzer (Applied Biosystems, Foster City, USA), and electropherograms were analyzed using GENEMAPPER 3.7 (Applied Biosystems, Foster City, USA).

**Genotypic diversity and clonal structure**

The average number of alleles per locus, expected (\(H_E\)) and observed (\(H_O\)) heterozygosity, and the global inbreeding coefficient (\(F_{IS}\)) were calculated separately for adult plants and offspring using GENEPOP 4.0.10 (Raymond and Rousset, 1995; Rousset, 2008) and FSTAT 2.9.3 (Goudet, 1995). Clonal structure was determined using GENCLONE 2.0 (Arnaud-Haond and Belkhir, 2007).
Paternity analysis

Paternity analysis was carried out with 49 mother plants and ten open-pollinated offspring per mother plant using CERVUS 3.0.3 (Marshall et al., 1998; Kalinowski et al., 2007). CERVUS uses a likelihood-based approach and assigns paternity according to the highest logarithm of the likelihood (LOD score). LOD scores are calculated by determining the likelihood of assignment of a parent relative to the likelihood of arbitrary parents. We applied the following simulation parameters to find the confidence level of paternity analysis assignment: 10 000 simulated mating events; 1000 candidate father plants; eight as the minimum number of loci; 0.8 as the proportion of candidate parents sampled; genotyping error rate of zero (Slate et al., 2000; Slavov et al., 2005); 0.66 as the rate of inbreeding (F<sub>IS</sub> value given by the GENEPOP analysis above); all adult plants treated as candidate father plants. As in Rathmacher et al. (2010), allele frequencies used in the simulation step were based on the A. alpina adult plants. We subsequently added alleles private to the offspring as P = 0.0001 to the frequency data file. In the paternity analysis, we used 95% as strict and 80% as relaxed confidence levels as recommended by Marshall et al. (1998). Effective pollen dispersal distances were calculated as the straight-line distance from the mother plant to the most likely pollen parent. Curve fitting estimation for the distribution of pollen dispersal distances was done using TABLECURVE 2D (Systat, Erkrath, Germany).

Mating system

To characterize the effective mating system in our study population of A. alpina, the mother plants and the open-pollinated offspring were analyzed with MLTR v3.3 (Ritland, 2002). The program generates the multi- and single-locus outcrossing rates (t<sub>m</sub> and t<sub>s</sub>, respectively), biparental inbreeding (t<sub>m - t</sub>s), maternal inbreeding (F), correlated selfing rate (r<sub>s</sub>), and the multilocus paternity correlation (r<sub>p</sub>). The correlated selfing rate indicates among-family variation in mating system; r<sub>s</sub> = 1 signifies that siblings are all either selfed or outcrossed and r<sub>s</sub> = 0 indicates that selfing rates do not vary among families. The multilocus paternity correlation is the proportion of full-sibs among outcrossed sibs and indicates whether offspring are the result of single or multiple paternities. The reciprocal of r<sub>p</sub> (N<sub>cp</sub> = 1 / r<sub>p</sub>) gives the effective number of pollen donors (Smouse et al., 2001; Fernández-Manjarres and Sork, 2005). All seeds belonging to a mother plant were grouped as a family and default settings for the estimated parameters were used in MLTR for the analysis of families (i.e. t =
Contemporary gene flow patterns in *Arabis alpina*

0.9, \( r_t = 0.1, r_p = 0.1, F = 0.1 \). We used 1000 bootstraps to derive confidence intervals with resampling done over families.

**Spatial genetic structure**

To estimate genetic structure of the adult plants in our study landscape, we used spatial autocorrelation analysis (Smouse and Peakall, 1999) as implemented in GENALEX 6.41 (Peakall and Smouse, 2006). Pairwise spatial autocorrelation coefficients \( \rho \) were computed as a multilocus average for diploid individuals and codominant markers. Correlograms were generated with eight size classes of even sample size: 0-30 m, 30-60 m, 60-120 m, 120-240 m, 240-480 m, 480-960 m, 960-1920 m for a dataset with all individuals and a dataset excluding clonal individuals. A second analysis was performed to test for spatial genetic structure for the dataset with all individuals at distances < 30 m, with size classes of 5 m. We used 1000 bootstraps to estimate 95% confidence intervals for the significance of \( \rho \) against no spatial structuring, i.e. random distribution.

**RESULTS**

*Genotypic diversity and clonal structure*

Over the entire landscape, adult plants and offspring of *A. alpina* showed high inbreeding coefficients \( F_{IS} = 0.66 \) and \( F_{IS} = 0.76 \), respectively), while mean genetic diversity was identical for both categories \( H_E = 0.54 \); Table 1). The number of alleles per locus ranged from three to 12, and the offspring had five private alleles. Only 28 clonal groups of adult plants were determined. There were 16 clonal groups with individuals located within 5 m and 12 clonal groups with individuals located at distances of 40 to 330 m between them. Nineteen clonal groups consisted of only two plants, while nine clonal groups consisted of three to seven plants.

*Paternity analysis and pollen flow distance*

CERVUS analysis resulted in an exclusion probability of 0.969 and a clear assignment of 249 of the 499 offspring analyzed (51%) to one most-likely father plant. The remaining 240 offspring (49%) were unassigned to a potential father plant. Of these, 51% of offspring had several potential father plants, whereas 49% of offspring had no potential father plants assigned in the study area. Of the assigned offspring, 86 were assigned with a confidence level of 95%, and 163 were assigned at an 80% confidence level. Ten offspring were typed at fewer than eight loci and were excluded from the
analysis. An overall selfing rate of 84%, i.e. 209 offspring resulting from selfing, was detected. Outcrossing events were rare, as only 39 of the assigned offspring (16%) were outcrossed. Eighteen of the 41 mother plants produced only selfed offspring and at most, mother plants had one to three outcrossed offspring.

The distance of effective pollen dispersal ranged from ≤ 5 m to about 1000 m (Figs. 1, 2). Altogether, 36% of the effective outcrossed pollen travelled ≤ 5 m and about 51% was dispersed to ≤ 20 m. Still, 6% of the effective pollen was dispersed at a maximum distance of 1000 m. Pollen dispersal occurred at elevation changes up to 350 m (Fig. 1). The curve best fitting the empirical data on pollen flow was a general non-linear power equation; pollen dispersal distance = 1.548 + 1672.60x^{-3.045} \ (R^2 = 0.804; \ P=0.001, \ Fig. 2).

**Mating system**

Outcrossing rates were considerably low ($t_m = 0.315; \ t_s = 0.118$) indicating a high degree of selfing in *A. alpina* (Table 2). Biparental inbreeding was significant ($t_m - t_s = 0.197$) and suggests frequent mating among relatives. The estimate for inbreeding in the mother plants was also high ($F = 0.674$). The correlated selfing rate $r_s$ was 0.305, and the multilocus paternity correlation $r_p$ was 0.712. Hence, a large number of offspring were the result of single paternity. The number of effective pollen donors ($N_{ep} = 1/r_p$) was 1.4 individuals.

**Spatial genetic structure**

In a correlogram showing autocorrelation coefficients as a function of distance classes and a corresponding 95% confidence interval, the distance class at which the estimate of $r$ is no longer significant provides an approximation of the extent of detectable positive spatial genetic structure (Peakall *et al.*, 2003). In the dataset considering all samples, the intercept occurred at 261 m indicating that in distance classes 0-240 m individuals were significantly more genetically related than expected from random distribution (Fig. 3a). The correlogram for the spatial distance class for < 30 m showed that individuals located at 5 m distance were substantially genetically related (Fig. 3b). The correlogram excluding clonal individuals gave similar results, supporting that clones did not affect spatial structuring (results not shown).
DISCUSSION

A considerable number of studies has been published describing historical gene flow in alpine plants as inferred from population genetic structure at large spatial scales or by direct observation of pollinators at small spatial scales (e.g. Hirao et al., 2006; Ohsawa and Ide, 2008; Paun et al., 2008; Gonzalo-Turpin and Hazard, 2009; Stöcklin et al., 2009; Meirmans et al. 2011). However, studies describing contemporary gene flow at the landscape scale are absent to our knowledge. The aim of this study was to estimate contemporary gene flow by pollen using paternity analysis and to examine the mating system in the alpine plant *Arabis alpina*. As we investigated gene flow by pollen and not by seed, we cannot infer current plant dispersal or migration. We detected a high selfing rate and a low number of outcrossing events in *A. alpina*. However, maximum gene flow distances were larger (≤ 1 km) than previously inferred for alpine plants (e.g. Stöcklin et al., 2009), and gene flow from outside the study system also occurred. Spatial genetic structure, indicative of gene flow within a given area, showed that individuals growing in close vicinity were genetically more related than expected from random distribution. Clonally propagated individuals were rare and did not considerably affect the spatial genetic structure of *A. alpina* in an alpine landscape.

*Long distance gene flow and selfing*

The maximum pollen dispersal distances found in the present study were much larger (≤ 1 km) than those detected in other studies on alpine plants quantifying contemporary pollen flow from pollinators (e.g. Schmitt, 1980; Hirao et al., 2006; Stöcklin et al., 2009) or from paternity analysis (Brunet and Holmquist, 2009). For instance, in the alpine plants *Epilobium fleischeri*, *Geum reptans* and *Campanula thyrsoides*, pollen was measured by the transport of fluorescent powder from flower to flower. This resulted in a maximum distance of pollen dispersal of only 40 m (Stöcklin et al., 2009). In another study, contemporary gene flow in the alpine plant *Aquilegia coerulea* was measured across patches using paternity analysis. The results of this study showed that pollinators frequently dispersed pollen among patches located 35 - 150 m apart (Brunet and Holmquist, 2009). The long-distance dispersal detected in our study in part reflects both the size of our study landscape (2 km²) and the method used to quantify pollen flow. This suggests that pollen flow measurements on small spatial scales such as distinct patches or pollen flow inferred from pollinator observation may only give limited information. In addition, we detected a high percentage of offspring (49%) resulting from either gene flow from unsampled individuals within the landscape.
(cryptic gene flow) or pollen immigration via gene flow from outside the study landscape (pollen inflow). Contemporary pollen immigration in natural populations of trees usually accounts for approximately one third of pollen flow (Hoebee et al., 2007). In our case, cryptic gene flow seemed to be possible as a result of working with a small herb. Some individuals might have been left unsampled causing undetected gene flow within the landscape. In addition, it is possible that pollen inflow originated from scree fields at higher elevations which were not sampled in our study or from ravines occurring at approximately 2 km distance.

Although, pollen flow in our study occurred across larger distances than in other studies, long-distance pollen dispersal was rare. Most outcrossing pollen dispersed within a radius of 5 m to the mother plant. This is reflected in general findings of other studies, which describe pollen flow as a numerically frequent event close to the source plant and decreasing with distance (Ellstrand, 1992; Kamm et al., 2009). For instance, in the study by Kamm et al. (2009), gene flow by pollen in the insect-pollinated tree Sorbus domestica was shown to reach distances up to 16 km, but most gene flow occurred within 200 m of the mother plant. The dispersal of pollen in alpine areas is generally believed to be directly limited by pollinator abundance, pollinator flight distances, difference in plant phenology and indirectly by low nutrient availability (Hirao and Kudo, 2004; García-Camacho and Totland, 2009). Pollinator abundance is low in alpine areas and pollinator activity depends on the frequency of adverse weather conditions and the seasonal phenology of plants (Hirao et al., 2006). However in our study, long-distance pollen flow was not restricted by differences in elevation (Fig. 1), suggesting that asynchronous flowering patterns might only partially have been responsible for restricting gene flow in A. alpina. Frequent pollen dispersal over short distances as inferred from pollinator behavior has been detected in other alpine plants (e.g. Schmitt, 1980; Stöcklin et al., 2009). Therefore, a combination of the above-described factors could explain the low number of outcrossing events and pollen dispersal over larger distances.

A substantial number of offspring (84%) in the paternity analysis resulted from selfing. This corresponds to the generally held view that alpine plants are often selfed, especially with increasing altitude (Bliss 1971). Previous studies have shown that A. alpina is self-compatible and highly selfing in the European Alps (Titz, 1971; Ansell et al., 2008; Buehler et al., 2011; Tedder et al., in press). As a result of pollen limitation, selfing could be a reproductive assurance mechanism (García-Camacho and Totland, 2009) allowing A. alpina to maintain population growth.
Realized mating patterns and spatial genetic structure

Arabis alpina populations in the Central Alps, which included our study landscape, are characterized by higher inbreeding than other alpine populations (Ansell et al., 2008). This suggests a regional change in breeding system (Ansell et al., 2008). In addition, the Central Alps show distinct phylogeographic patterns indicating independent origins and/or local glacial refugia (Ehrich et al., 2007; Ansell et al., 2008; Alvarez et al. 2009). In our investigation of realized mating patterns in A. alpina, sexual reproduction occurred mostly by selfing. Mother plants were shown to have low outcrossing rates, which often resulted in offspring from single paternity. Due to low number of effective pollen donors, inferred pollen availability was low. Another study conducted by Tedder et al. (in press) showed that populations in the Central Alps lack an inbreeding avoidance mechanism generally present in Brassicaceae (Bateman, 1954). This could suggest that in our landscape, self-compatible plants colonized the area, or a loss of self-incompatibility has occurred. Thus, inbreeding may be facilitated by the local mating system and cause higher divergence of A. alpina populations in the Central Alps.

Spatial genetic structure is expected to be prominent in alpine plant populations as a result of habitat fragmentation and spatial isolation (Till-Bottraud and Gaudeul, 2002). In our dataset, we detected that up to a distance of about 260 m, individuals were genetically more similar than expected under random distribution (Fig 3). Spatial genetic structure was not affected by clonality, since few clonally propagated individuals were found in the landscape, and these did not visibly affect the correlogram. A closer look at the genetic structure found for A. alpina showed that individuals located within 5 m were highly related, indicating that a large fraction of gene dispersal frequently occurs over such short distances. This assumption was congruent with the result of the paternity analysis showing that the highest frequency of pollen dispersal occurred over < 5 m from the source (Fig. 2). Therefore, the genetic structure in the landscape mirrors the contemporary pollen dispersal and mating patterns in A. alpina.

Conclusions and implications for climate change

Alpine landscapes are especially prone to be affected by climate change (IPCC, 2007), and alpine organisms are showing first signs of range contraction or distribution changes (Hughes, 2000; McCarthy, 2001; Parmesan, 2006). Various studies have shown that alpine plants usually possess ample variation at neutral genetic loci (Stöcklin et al., 2009). Alpine plants may, thus, harbor
genetic variation suitable for adaptation to changing environmental and climatic conditions. However, the question whether they are able to adapt at short timescales remains open (Till-Bottraud and Gaudeul, 2002; Thuiller, 2007). In any case, the survival of alpine species under climate change does not only depend on suitable standing genetic variation, but also on an organism’s ability to spread beneficial genes to those areas where they match future environmental conditions (Byars et al., 2009). This is strongly influenced by outcrossing gene flow and the mating system of alpine plants.

The present work challenges the common findings that gene flow only occurs across short distances in alpine areas. To make general conclusions about gene flow in *Arabis alpina* and in alpine plants additional landscapes or alpine species, representing a variety of mating and dispersal strategies, should be analyzed. In predominantly outcrossing plants, gene flow is expected to occur more frequently than shown here for the highly selfing plant *A. alpina*. However, in this study, we used molecular methods and conducted a landscape-scale analysis, which showed that contemporary gene flow can reach large distances in *A. alpina*. In concert with other gene flow studies, we detected that most gene flow was dispersed close to the source. In addition, spatial structuring occurred at small spatial scales as expected for alpine plants, a likely result of the small dispersal distances and the high selfing rate of *A. alpina*. Nevertheless, even few long-distance gene flow events have been shown to spread advantageous alleles among populations (Morjan and Rieseberg, 2004). By combining molecular methods with a landscape scale analysis, we can conclude that gene flow has the potential to spread beneficial alleles in *A. alpina* at a short timescale, which might help this species to rapidly adapt to climate change.

**FUNDING**

This work was supported by the Competence Center Environment and Sustainability BIOCHANGE project of the ETH-domain.

**ACKNOWLEDGMENTS**

We thank Simone Prospero and Sarah Bryner for comments on the manuscript and Thomas Wuest for software assistance. Microsatellite analysis was carried out at the Genetic Diversity Centre, ETH Zurich.
LITERATURE CITED


Contemporary gene flow patterns in *Arabis alpina*


Contemporary gene flow patterns in *Arabis alpina*


**IPCC. 2007.** *Climate change 2007: the physical science basis, contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Edited by Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL. Cambridge: Cambridge University Press.


Contemporary gene flow patterns in *Arabis alpina*


**Schröter C. 1926.** *Das Pflanzenleben der Alpen.* Zürich, Switzerland: Raustein.


Fig. 1. Distribution of *Arabis alpina* in Urnerboden Switzerland. (A) Spatial clusters of individual *A. alpina* plants are represented by squares. Patterns of current gene flow by pollen of *A. alpina* are represented by lines connecting mating partners (mother plants are represented by circles and father plants are represented by triangles) as detected in genetic paternity analysis. Thin lines represent one mating event and thick lines represent two mating events. (B) Long-distance gene flow events (> 100 m) across the landscape, and close ups of the scree field (C), ravine (D), and river (E) are shown for short distance gene flow events (< 100 m).
Contemporary gene flow patterns in *Arabis alpina*

**Fig. 2.** Frequency distribution of distances of current gene flow by pollen in *Arabis alpina* as detected in genetic paternity analysis. The dashed line represents the best fitting curve ($y = 1.548 + 1672.60x^{-3.045}$; $R^2 = 0.804$, $P=0.001$).

**Fig. 3.** Analysis of spatial genetic structure of *Arabis alpina* in a sub-alpine to alpine landscape using spatial autocorrelation analysis. The correlograms show average correlation coefficients between pairs of individuals plotted against distance classes: (A) for eight distance classes across the entire range of the study area, and (B) for a section of the correlogram with 5 m distance classes up to 30 m. Dashed lines delimit 95% confidence intervals under the hypothesis of randomly distributed genotypes.
TABLE 1. Characteristics of 18 nuclear microsatellite loci for Arabis alpina in a subalpine to alpine landscape. Shown are the number of alleles ($N_A$), expected heterozygosity ($H_E$), observed heterozygosity ($H_O$) and inbreeding coefficient ($F_{IS}$) for each locus in the adult plants and offspring. Standard errors are shown in parentheses. Furthermore, for each locus the fluorescent labeling (Labels) and the multiplex PCR in which the locus were amplified (Multiplex) are given.

<table>
<thead>
<tr>
<th>Locus</th>
<th>$N_A$</th>
<th>$H_E$</th>
<th>$H_O$</th>
<th>$F_{IS}$</th>
<th>$N_A$</th>
<th>$H_E$</th>
<th>$H_O$</th>
<th>$F_{IS}$</th>
<th>Labels</th>
<th>Multiplex</th>
</tr>
</thead>
<tbody>
<tr>
<td>3JUY</td>
<td>6</td>
<td>0.74</td>
<td>0.22</td>
<td>0.79</td>
<td>6</td>
<td>0.68</td>
<td>0.23</td>
<td>0.67</td>
<td>NED</td>
<td>1</td>
</tr>
<tr>
<td>4MDH</td>
<td>5</td>
<td>0.48</td>
<td>0.13</td>
<td>0.73</td>
<td>4</td>
<td>0.41</td>
<td>0.07</td>
<td>0.83</td>
<td>FAM</td>
<td>1</td>
</tr>
<tr>
<td>5GTC</td>
<td>10</td>
<td>0.70</td>
<td>0.26</td>
<td>0.63</td>
<td>6</td>
<td>0.54</td>
<td>0.15</td>
<td>0.72</td>
<td>HEX</td>
<td>1</td>
</tr>
<tr>
<td>5VSH</td>
<td>12</td>
<td>0.45</td>
<td>0.18</td>
<td>0.61</td>
<td>8</td>
<td>0.44</td>
<td>0.13</td>
<td>0.71</td>
<td>FAM</td>
<td>1</td>
</tr>
<tr>
<td>A1T8T</td>
<td>4</td>
<td>0.64</td>
<td>0.24</td>
<td>0.62</td>
<td>3</td>
<td>0.65</td>
<td>0.15</td>
<td>0.77</td>
<td>NED</td>
<td>1</td>
</tr>
<tr>
<td>A4W7</td>
<td>5</td>
<td>0.50</td>
<td>0.15</td>
<td>0.71</td>
<td>5</td>
<td>0.58</td>
<td>0.13</td>
<td>0.77</td>
<td>HEX</td>
<td>1</td>
</tr>
<tr>
<td>5JSE</td>
<td>4</td>
<td>0.55</td>
<td>0.20</td>
<td>0.64</td>
<td>4</td>
<td>0.57</td>
<td>0.09</td>
<td>0.84</td>
<td>FAM</td>
<td>1</td>
</tr>
<tr>
<td>5Q19</td>
<td>6</td>
<td>0.40</td>
<td>0.16</td>
<td>0.61</td>
<td>5</td>
<td>0.47</td>
<td>0.14</td>
<td>0.69</td>
<td>NED</td>
<td>2</td>
</tr>
<tr>
<td>6U3A</td>
<td>5</td>
<td>0.41</td>
<td>0.14</td>
<td>0.66</td>
<td>5</td>
<td>0.53</td>
<td>0.06</td>
<td>0.88</td>
<td>NED</td>
<td>2</td>
</tr>
<tr>
<td>A25GM</td>
<td>6</td>
<td>0.71</td>
<td>0.22</td>
<td>0.69</td>
<td>5</td>
<td>0.67</td>
<td>0.23</td>
<td>0.65</td>
<td>HEX</td>
<td>2</td>
</tr>
<tr>
<td>DEET</td>
<td>7</td>
<td>0.45</td>
<td>0.18</td>
<td>0.59</td>
<td>6</td>
<td>0.59</td>
<td>0.08</td>
<td>0.86</td>
<td>FAM</td>
<td>2</td>
</tr>
<tr>
<td>46HI</td>
<td>5</td>
<td>0.49</td>
<td>0.19</td>
<td>0.61</td>
<td>5</td>
<td>0.38</td>
<td>0.09</td>
<td>0.76</td>
<td>HEX</td>
<td>3</td>
</tr>
<tr>
<td>7PQ</td>
<td>7</td>
<td>0.64</td>
<td>0.22</td>
<td>0.66</td>
<td>4</td>
<td>0.63</td>
<td>0.12</td>
<td>0.81</td>
<td>HEX</td>
<td>3</td>
</tr>
<tr>
<td>BWF1</td>
<td>7</td>
<td>0.45</td>
<td>0.12</td>
<td>0.74</td>
<td>7</td>
<td>0.34</td>
<td>0.10</td>
<td>0.72</td>
<td>FAM</td>
<td>3</td>
</tr>
<tr>
<td>7PQ</td>
<td>10</td>
<td>0.73</td>
<td>0.25</td>
<td>0.66</td>
<td>9</td>
<td>0.76</td>
<td>0.17</td>
<td>0.77</td>
<td>HEX</td>
<td>4</td>
</tr>
<tr>
<td>A8E90</td>
<td>4</td>
<td>0.67</td>
<td>0.17</td>
<td>0.75</td>
<td>5</td>
<td>0.69</td>
<td>0.13</td>
<td>0.81</td>
<td>FAM</td>
<td>4</td>
</tr>
<tr>
<td>A93Q</td>
<td>3</td>
<td>0.18</td>
<td>0.06</td>
<td>0.65</td>
<td>2</td>
<td>0.30</td>
<td>0.06</td>
<td>0.79</td>
<td>FAM</td>
<td>4</td>
</tr>
<tr>
<td>EB54</td>
<td>5</td>
<td>0.61</td>
<td>0.20</td>
<td>0.66</td>
<td>6</td>
<td>0.54</td>
<td>0.17</td>
<td>0.69</td>
<td>HEX</td>
<td>4</td>
</tr>
</tbody>
</table>

Mean 6.17 (2.38) 0.54 (0.15) 0.18 (0.05) 0.66 (0.05) 5.28 (1.67) 0.54 (0.13) 0.13 (0.05) 0.76 (0.07)
TABLE 2. Realized mating pattern analysis of Arabis alpina on a sub-alpine to alpine landscape. The parameters given are the parental inbreeding coefficient (F), multi-locus (t<sub>m</sub>) and single-locus (t<sub>s</sub>) outcrossing rates, biparental inbreeding (t<sub>m</sub> - t<sub>s</sub>), correlated selfing rate (r<sub>s</sub>), multilocus correlation of paternity (r<sub>p</sub>) and the number of effective pollen donors (N<sub>ep</sub>). Standard errors are shown in parentheses.

<table>
<thead>
<tr>
<th>Mating system parameters</th>
<th>Value (S.E.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parental inbreeding, F</td>
<td>0.674 (0.050)</td>
</tr>
<tr>
<td>Multilocus outcrossing rate, t&lt;sub&gt;m&lt;/sub&gt;</td>
<td>0.315 (0.035)</td>
</tr>
<tr>
<td>Single-locus outcrossing rate, t&lt;sub&gt;s&lt;/sub&gt;</td>
<td>0.118 (0.019)</td>
</tr>
<tr>
<td>Biparental inbreeding, t&lt;sub&gt;m&lt;/sub&gt;-t&lt;sub&gt;s&lt;/sub&gt;</td>
<td>0.197 (0.025)</td>
</tr>
<tr>
<td>Correlated selfing rate, r&lt;sub&gt;s&lt;/sub&gt;</td>
<td>0.305 (0.062)</td>
</tr>
<tr>
<td>Multilocus paternity correlation, r&lt;sub&gt;p&lt;/sub&gt;</td>
<td>0.712 (0.059)</td>
</tr>
<tr>
<td>Number of effective pollen donors, N&lt;sub&gt;ep&lt;/sub&gt;</td>
<td>1.400</td>
</tr>
</tbody>
</table>
Chapter V

Landscape genetics of plants

Rolf Holderegger, Dominique Buehler, Felix Gugerli and Stéphanie Manel

Published in Trends in Plant Sciences
I worked on the conceptualization of the work and writing
Abstract

Landscape genetics is the amalgamation of landscape ecology and population genetics to help with understanding micro evolutionary processes such as gene flow and adaptation. In this review, we examine why landscape genetics of plants lags behind that of animals, both in number of studies and consideration of landscape elements. The classical landscape distance/resistance approach to study gene flow is challenging in plants, whereas boundary detection and the assessment of contemporary gene flow are more feasible. By contrast, the new field of landscape genetics of adaptive genetic variation, establishing the relationship between adaptive genomic regions and environmental factors in natural populations, is prominent in plant studies. Landscape genetics is ideally suited to study processes such as migration and adaptation under global change.
The booming field of landscape genetics

Landscape genetics is the amalgamation of landscape ecology and population genetics to help with understanding micro evolutionary processes such as gene flow and adaptation on the scale of natural landscapes [1]. This field investigates how landscape elements and environmental factors influence the spatial distribution of genetic variation. For instance, landscape genetics assesses how landscape elements such as forests or open fields affect gene flow (see Glossary) in species inhabiting semi-natural habitat remnants in an otherwise unsuitable and intensively used landscape; a question of interest in conservation management. Similarly, landscape genetics analyzes whether non-crop strips provide effective barriers to gene flow between organic and genetically modified crops; a question of practical importance in agriculture. Landscape genetics also deals with how environmental factors such as temperature or precipitation affect adaptive genetic variation; relevant information in the context of climate change [2, 3]. Landscape genetics still suffers from a lack of theoretical foundations and expectations [4–6], but empirical studies are typically characterized by including geo-referenced individuals or populations genotyped at multiple loci and at least one landscape or environmental variable of interest (in addition to geographical distance) measured at or in between sampling locations [7]. Depending on the landscape or environmental variables assessed, the study area might be large (thousands of km$^2$), for example, when studying the adaptive response of a dominant forest tree to latitude, or small (hundreds of m$^2$), for example, when assessing pollen dispersal in a forest herb.

Given its appeal to both basic and applied sciences, landscape genetics has received increasing attention in recent years. The field currently deals with two main topics. First, landscape and environment are evaluated by considering their effects on migration, dispersal and gene flow, which are measured in terms of neutral genetic variation [7]. Second, landscape genetics has started to explore the interaction between environment and adaptive genetic variation in natural populations and individuals, a new field often referred to as landscape genomics [8–10].

Landscape genetics of plants

A recent survey [7] showed that the majority of landscape genetic studies have dealt with animals and only seldom with plants (~18%). Furthermore, studies on animals and plants differ
Landscape genetics in plants

in study design and analytical approaches. Why are there such differences in landscape genetic studies on plants and animals?

When dealing with gene flow, landscape genetics considers the landscape between sampling locations. When studying adaptive genetic variation, landscape genetics deals with the particular environment at sampling locations [11]. These two situations make plants either less or more amenable to landscape genetic analysis than animals. Because sessile plants directly respond to the environment at their growing site, the study of adaptive genetic variation in plants is straightforward. By contrast, corresponding animal studies must account for the environment of the entire home range or of all resource sites of individuals or populations. However, when assessing migration, dispersal and gene flow, studying plants is more complex than investigating animals. In animals, individuals disperse to other locations and provide gene flow when mating. In vascular plants, gene flow mainly happens through two processes, the dispersal of diploid embryos in seeds and of haploid male gametes in pollen. In wind-pollinated and -dispersed plants, an abiotic factor acts as the primary pollen and seed vector. In insect-pollinated and animal-dispersed plants, animals act as dispersal vectors of propagules. Here, it is a moving animal that reacts to the landscape and not the plant itself. Landscape genetic studies investigating gene flow in plants thus deal with the problem that two propagule types are dispersed by particular vectors. For instance, the fragmentation effect of roads on plants is not direct, but induced by the indirect effects of roads on pollinators or animal seed dispersers, whereas wind-pollinated and -dispersed plants might not be affected by roads [12].

Published reviews on landscape genetics have focused on animals, and many plant studies have not included landscape elements, apart from geographical distance. Given the characteristics of plants and the present shortage of empirical studies, it seems relevant to provide an overview focusing on plants and stressing associated benefits and limitations of common landscape genetic approaches. In this review, we will first summarize and discuss the main approaches currently used to study gene flow at the landscape scale: the landscape distance/resistance approach, the overlay technique and the assessment of contemporary gene flow. We will then examine landscape genetic approaches exploring adaptively relevant genetic variation [13, 14].
Gene flow on the landscape scale

Landscape distance/resistance
This classical approach correlates a matrix of genetic distances as indirect measurements of gene flow [15] with matrices of landscape distances/resistances and geographic distances (Figure 1a). The genetic matrix consists of pairwise genetic distances among all pairs of individuals or populations studied. Various estimators can be used for this purpose, for example, genetic Chord distance ($d_c$), Nei’s distance ($d_n$) or population differentiation ($F_{ST}$) [7, 16]. The geographic distance matrix contains the straight line distances among all sampling locations. By contrast, the landscape distance matrix varies and can contain the length, area or percentage of cover of landscape elements, such as ditches, wetlands or woodlands in corridors of a certain width between pairs of sampling locations. It can also be a 0/1 matrix, for example, when some sampling locations are separated by a river and others are not. In a more sophisticated design, land cover/land use and topography are taken from existing geographic information system (GIS) data or from field surveys, and a level of resistance to movement is given to each raster cell containing a particular landscape feature. For instance, forests might hinder gene flow whereas open fields have no effect, and a high resistance is thus assigned to forested grid cells, whereas open-field cells receive low resistance values. Using GIS technology, the length of the shortest path connecting two sampling locations is determined by maximizing movement through low-resistance cells [17]. Such least-cost paths form the entries for a landscape resistance matrix. Resistance assignment relies on expert knowledge or a priori ecological information on study organisms [18], and several alternative landscape resistance models are generally tested per study.

For statistical analysis, the genetic distance matrix is either correlated to the geographic distance and landscape distance/resistance matrices separately using a Mantel test, or the effect of geographic distance is first partialed out before estimating correlation with landscape distance/resistance in a partial Mantel test [19] (Figure 1a). This procedure enables the effects of geographic distance (i.e. isolation by distance [20]) and landscape elements on gene flow to be disentangled [21]. However, partial Mantel tests have been criticized because of their permutation procedure (e.g. [22]), and a variety of alternative approaches have been suggested [23, 24].
The landscape resistance approach and especially least cost path analysis have been popular in animal studies, but there are few such studies in plants [7] (but see [25, 26]). The approach is intuitively appealing in animals, but less so in plants because pollen and seed dispersal mechanisms depend on (multiple) biotic or abiotic vectors. In addition, genetic distances in plants do not account for differences in seed and pollen dispersal [16].

Given these shortcomings of the landscape distance/resistance approach in plants, it is not surprising that it has mainly been applied in special situations. Mantel and partial Mantel tests of genetic distances with geographic distances along coasts or streams have been used in studies on wild sugar beet (*Beta vulgaris* ssp. *maritima* [27]) or Japanese primrose (*Primula sieboldii* [28]). Similarly, the effect of differences in flowering time among individuals on genetic distances has been studied with partial Mantel tests in several alpine snowbed plants [29, 30]. In fact, researchers can still profit from the landscape distance/resistance approach in plants when analyzing the influence of major landscape elements on gene flow in a general way (Box 1). For instance, forests might act as barriers to gene flow in a wind-pollinated and wind-dispersed meadow herb. The occurrence of large forests between populations or individuals should generally exhibit a negative effect on gene flow and effectively increase genetic distances. Researchers could also simultaneously determine the effects of several landscape distances between pairs of sampling locations (including geographic distance) on genetic distances in multiple linear regression with permutation-based significance testing using software such as PERMUTE [31] or BLOSSOM [32]. When testing several models of landscape distance/resistance, significance values should be adjusted because of multiple testing, and model performance has to be evaluated (Box 2) [33], which has rarely been done in landscape genetics (but see [34, 35]).

**Overlay technique**

The overlay approach of landscape genetics [7] identifies population groups, barriers, genetic discontinuities or isolines. These genetic structures are overlaid onto maps of selected landscape elements such as topography or land cover/land use to search for geographical coincidences of group boundaries, barriers, genetic discontinuities or isolines with landscape elements (Figure 1b).
Various statistical approaches can be used to form genetic groups of populations or individuals to be used in overlays. Bayesian clustering [36] based on Hardy-Weinberg and linkage equilibrium as implemented in STRUCTURE [37], BAPS-5 [38], TESS [39] or GENELAND [40] are widely used for this purpose. These programs can also consider coordinates of sampling locations [38–40]. Alternatively, non-Bayesian S-AMOVA can be applied to form population groups. This method maximizes genetic differentiation ($F_{CT}$) among groups [41]. For instance, no or only weak clustering was detected in the alpine blue thistle (*Eryngium alpinum*) [42] and the rainforest Anguama tree (*Aucoumea klaineana*) [43] despite spatial distribution gaps, and grouping according to river catchments was inferred in the Chinese maidenhair fern (*Adiantum reniforme* var. *sinensis*) [44]. Other methods search for areas of strong changes in allele frequencies, such as the Monmonier algorithm implemented in BARRIER [45] or ALLELES IN SPACE [46] to detect genetic barriers among populations (e.g. [26]). Individual- or population-based wombling also identifies genetic discontinuities [1, 47, 48]. Finally, interpolation such as kriging from principal component analysis (PCA) axis loadings determines genetic isolines similar to contour lines in topographical maps (Figure 1b). For instance, kriging helped in visualizing the small-scale genetic structure in snapdragon (*Antirrhinum microphyllum*) [49]. Membership coefficients estimated from Bayesian clustering methods can also be interpolated, providing clustering surface maps (e.g. [25]).

The major drawback of the common overlay approach is that spatial coincidence of landscape elements with genetic discontinuities, barriers or isolines is simply based on subjective visual inspection. No statistical procedure is usually involved in this step. Overlays are thus of exploratory nature and prone to false inference. However, subjectivity in analysis could be avoided by applying boundary overlap statistics [50] and [51]. Such statistics have, for example, been used to study the spatial coincidence between heterogeneity of forest structures and boundaries of bird territories [52]. To our knowledge, boundary overlap statistics have rarely been used in landscape genetics (Box 1).

Despite its appeal for studying sessile organisms, the overlay technique has not been popular in plants and should therefore be further explored, especially in connection with appropriate significance testing. In fact, most plant studies have dealt with phylogeographic patterns [53], and only a few have overlaid inferred genetic patterns onto land-cover/land-use or
topographical maps. By contrast, obvious effects of various landscape elements on gene flow have been identified with overlays in animal studies (e.g. [54, 55]).

Assessment of contemporary gene flow
Landscapes throughout the world are changing at an unprecedented speed [2]. This poses a problem to landscape genetics because the landscape distance/resistance and (most) overlay approaches rely on historical measurements of gene flow (i.e. genetic distance and differentiation [15]). However, when landscapes are changing rapidly, historical gene flow measures tend to reflect the historical rather than the contemporary landscape [56, 57]. In such situations, researchers would like to assess contemporary migration and gene flow [6]. To do so, two main approaches are currently available: parentage analysis [58] and assignment tests [36] (Figure 1c).

Parentage analysis, and its variants maternity and paternity analysis, has widely been used in plants, especially trees [59]. Paternity analysis infers contemporary pollen flow from the genetic analysis of open-pollinated offspring of known mothers by using maximum likelihood methods (e.g. CERVUS [60]) or Bayesian inference (MASTERBAYES [61]) to identify the most probable fathers [58]. Similarly, maternity analysis uses maternal seed coat tissue [62] or uniparentally inherited organellar DNA markers (e.g. chloroplast DNA [16]) to detect the most probable mothers of seedlings or trapped seed. Parentage analysis asks for complete sampling of all potential parents in a study area, which logistically limits the study range. This is a major shortcoming of studies using parentage analysis, and therefore landscape-scale studies are rare (but see [63]).

From studies on contemporary gene flow, we have learned that pollen and seed dispersal are more frequent and occur over greater distances than expected from ecological investigations [6]. However, the effect of landscapes has rarely been considered, and most studies simply compared contemporary gene flow patterns in fragmented versus non-fragmented situations (e.g. [64–66]). Researchers have started to assess landscape effects by applying multiple linear regression with permutation testing of the frequency of mating events among pairs of individuals and various landscape elements [31, 32] (Figure 1c). We have also compared realized mating patterns with a null model of saturated mating among all individuals studied in a regional population of the service tree (Sorbus domestica [67]). One problem with the landscape genetic analysis of data from parentage analysis is that no mating between particular individuals either
reflects a real lack of mating or simply a low detection probability because of insufficient sample size in terms of parents and offspring studied.

Alternatively, contemporary migration and gene flow are inferred through assignment tests, which are often used in animals [36]. Assignment tests can discriminate between first generation migrants, migrants during the last generation and recent migrants (during the last few generations) using Bayesian software such as GENECLASS [68] or BAYESASS [69]. Assignment tests on plants are rare, probably because it is difficult to disentangle gene flow by seed and pollen. However, by using particular settings in GENECLASS [68], He et al. [70] restricted assignment to contemporary seed dispersal. Two major caveats to be considered when using assignment tests are that although assignment tests enable individual migrants to be identified, migration rates and populations of origin can only be appropriately identified if all populations in a landscape were included in sampling and analysis.

Several studies have detected that contemporary seed dispersal occurs over large stretches of unsuitable habitat, across inhospitable mountain ridges or within river catchments [70–72]. Again, landscape effects have seldom been combined with assignment tests, although dedicated software (BMIR [73]) is available for correlating directional contemporary migration rates based on Bayesian inference with landscape distance data in a multiple linear regression framework. It is evident that additional statistical tools have to be developed to analyze the full breadth of contemporary gene flow and migration in a landscape context (Box 1).

**Landscape genetics of adaptive genetic variation**

A popular route in landscape genetics of plants has been to correlate population genetic diversity with environmental factors at habitat patches. Researchers have studied whether genetic diversity was related to local soil type, humidity, vegetation structure or management type (e.g. [74–76]). These studies were based on the analysis of neutral markers. However, neutral genetic diversity does not directly relate to adaptive genetic variation [77, 78] and mainly reflects local population size [79]. Neutral genetic diversity is only indirectly affected by local environmental factors if these factors influence processes such as gene flow or mating [3]. Therefore, if researchers want to establish the relationship between genetic diversity and environmental factors, they should specifically assess adaptive genetic variation [2].
The first step in analyzing adaptive genetic variation is to identify genomic regions bearing signs of selection. Among various methodological approaches [80], genome scans have been the preferred method in non-model organisms for about a decade. In this analysis, loci showing a higher genetic differentiation among populations ($F_{ST}$) than expected under neutrality (i.e. outlier loci) are identified out of a large number of loci studied across a genome [81–89]. This population genomic approach has also been applied to plants (e.g. [84] and [85]). Selected environmental factors are only \textit{a posteriori} correlated with allele frequencies at outlier loci to infer potential selective pressures [9].

By contrast, the landscape genetic approach [8] directly uses environmental data to pinpoint molecular markers linked to or located within genomic regions under selection [9]. Accordingly, many samples are collected along environmental gradients (Figure 2), and large genome scans with hundreds to thousands of amplified fragment polymorphisms (AFLPs) or single-nucleotide polymorphisms (SNPs) are performed [16, 86]. Allele occurrence in individuals or allele frequency in populations is subsequently correlated with local environmental conditions, for example, estimates of temperature, precipitation, slope, altitude or habitat type [9, 10]. Several statistical methods are used for this purpose, such as simple linear regression [13] or generalized linear models, for example the logistic regression implemented in the spatial analysis method (SAM [87]) (Box 2). Molecular markers significantly correlated with environmental factors having substantial effect size are seen as linked to genomic regions influenced by these factors. After molecular characterization, identified markers can be used for cross-validation in other landscapes or in experiments aiming at verifying their adaptive or molecular functionality (Boxes 1 and 2) [2, 9]. Plants are particularly more amenable to corresponding experimentation than (most) animals. In the model plant thale cress (\textit{Arabidopsis thaliana}), a large SNP genome scan and an in detail genomic analysis found that one particular allele at locus \textit{ACD6} underpins resistance to microbial infection and herbivory in natural populations and therefore provides large fitness advantages under high pathogen and herbivory pressure, despite severely reducing vegetative growth [88].

Only a few landscape genetic studies on the adaptive genetic variation of plants are available so far (Box 2), and the methodological foundations of the field are not yet fully explored. In particular, researchers have only started to deal with the problem of spatial genetic structure [10, 89], potentially interfering with landscape genetic analysis and leading to the
detection of molecular markers falsely found to be linked to genomic regions under selection (Boxes 1 and 3). Likewise, most studies still lack any indication of the functionality of the outlier loci identified, be it based on experimental or molecular evidence.

**Perspectives**

Landscape genetics of plants is a largely under-explored field. One reason for this is that applying the classical landscape distance/resistance approach to infer landscape effects on gene flow is less amenable to studies of plants than to animals, owing to the particular means of gene dispersal through pollen and seed. Also the overlay technique has rarely been used in plants in a true landscape genetic setting with sound statistical analysis. Many studies on contemporary gene flow were small-scale and did not consider landscape effects. Therefore, significant progress can be achieved by applying existing methodology to plants on adequate spatial scales. However, it is apparent that the development of new statistical tools is necessary to analyze genetic data in concert with landscape and environmental data [90]. Interesting recent advances include the incorporation of graph theory in landscape genetics [91], which considers landscape and environmental data at and between sampling locations. Various types of hierarchical Bayesian models are also gaining increasing popularity in ecology and genetics [11]. A conceptual shortcoming of many landscape genetic studies is the virtual lack of replication at the landscape level (but see [75, 92]) and of multi-species studies (but see [91, 93]). Replicated landscape genetic analyses require particularly large genetic sample sizes and hence should profit from the current exponential increase in sequencing and genotyping capacity [94]. To deal with the corresponding huge genetic data sets, bioinformatics already relies on machine learning techniques. These techniques enable the screening of large numbers of genetic markers such as SNPs, making them particularly relevant in genomics and the study of adaptive genetic variation.

Global change, that is the world-wide alteration of natural and traditionally used landscapes and the rapidly changing climate, demands profound knowledge about the migration ability of species as well as their potential to adapt to new, human-altered environments. Landscape genetics is ideally suited to provide such relevant real-world data [2, 9]. For this task to be achieved, researchers have to make full use of existing and newly developed methodological landscape genetic approaches, especially so in plants.
Acknowledgements

R.H., D.B. and F.G. thank the CCES ENHANCE and BIOCHANGE projects of the ETH domain, the AVE project of the Swiss National Science Foundation (CRSI33 127155/1) and the European ECOCHANGE project (FP6-036866) for financial support. S.M. was supported by the Institut Universitaire de France as a junior member. R.H. and S.M. also acknowledge support by the National Center for Ecological Analysis and Synthesis, funded by NSF (Grant DEB-0553768), the University of California at Santa Barbara, and the State of California. We also thank an anonymous referee for valuable comments on the manuscript.
References


Landscape genetics in plants


53 J.C. Avise, Molecular Markers, Natural History and Evolution, Chapman and Hall (1994).


Glossary

AFLPs
amplified fragment length polymorphisms. Dominant and anonymous DNA fingerprints.

Allele distribution model
a (usually non-spatial) statistical description of how allele frequencies at loci linked to genes under selection are influenced by distinct environmental factors.

Assignment test
statistical approach that assigns an individual to that sampled population from which its multilocus genotype is most likely to be derived.

Bayesian inference
statistical approach using prior data or information to estimate posterior probabilities of a hypothesis to be correct. For instance, prior information (or just a guess) on migration rates can be used in Bayesian assignment tests to infer contemporary gene flow.

Chord distance ($d_c$)
a measure of genetic similarity between individuals or populations based on allele frequencies and located on a sphere. Taking values between 0 and 1.

Gene flow
exchange of genes among populations or individuals.

Genome scan
genotyping of many samples at a large number of (potentially anonymous) molecular markers across the genome, used in outlier detection and the landscape genetic analysis of adaptive genetic variation.

Hardy-Weinberg equilibrium
population genetic principle stating that allele and genotype frequencies reach equilibrium and stay constant in random mating populations assuming large population size, no selection, no migration and no mutation.

Isolation by distance
spatial pattern describing decreasing genetic relatedness of populations or individuals with increasing geographic distance.

Kriging
geostatistical technique to interpolate the value of a parameter of interest at an unobserved geographic location from observations of this value at nearby locations.

Landscape
an area spatially heterogeneous in one or more biotic and abiotic factors of interest. From the human perspective, a landscape is perceived as a kilometers-wide environmental mosaic.

Landscape distance
distance-like measurements parameterizing the landscape between two localities, for example, geographical distance along a river, number of roads to be crossed, or percentage of forest cover in a corridor strip connecting two localities.
Landscape resistance
permeability values of different landscape elements, describing their resistance to migration and dispersal. From these resistance values, different types of cumulative resistances between localities can be calculated, for example, least cost paths.

Least cost path
length of a path minimizing the cumulative landscape resistance between two localities.

Mantel test
a permutation-based statistical test describing the correlation between two distance matrices. A partial Mantel simultaneously accounts for the effects of a third (or several) distance matrix (matrices).

Monmonier's algorithm
detects genetic boundaries by finding the path exhibiting the largest genetic distances among neighboring populations.

Nei's genetic distance ($d_n$)
measure of genetic similarity based on the probability that two randomly chosen alleles from different populations or individuals are identical. Taking values between 0 and 1.

Neutral molecular marker
molecular markers not affected by natural selection.

PCNMs
principal coordinates of neighbor matrices are a spectral representation of all spatial relationships among sampling locations and describe all spatial scales that can be accommodated by the sampling design. PCNMs are calculated from principal coordinate analysis (PCoA). In landscape genetics, they are used to account for spatial relationships among sampling locations and for unaccounted environmental factors.

Permutation test
type of statistical tests that rely on resampling of data for significance testing and not on theoretical probability distributions as in classical statistics. For instance, permutation tests are used to account for the non-independence of genetic, geographic and landscape distances among sampling locations in landscape genetics.

Population differentiation ($F_{ST}$)
different measurements of the amount of genetic variation found between populations. Most often used is Wright's $F_{ST}$, taking values between 0 and 1. Similar measurements can be calculated for genetic differentiation among groups of populations ($F_{CT}$).

SNPs
single-nucleotide polymorphisms. Bi-allelic, co-dominant molecular markers of known position in the genome. Increasingly used in landscape genetics.

Wombling
approach to search for areas across an interpolated allele frequency surface where the slopes of the surface are particularly steep. Used to infer genetic breaks or discontinuities.
Figure 1. Schematic summary of the three major analytical approaches that are currently used by landscape geneticists to study gene flow. (a–c) Individuals or populations (circles) are studied in a simple landscape consisting of two land cover types: meadows (open) and forests (hatched green). (a) In the landscape distance/resistance approach, a matrix of genetic distances between all pairs of individuals (e.g. Nei’s genetic distance $d_n$, Chord distance $d_c$ [17]) or between all pairs of populations (e.g. genetic differentiation $F_{ST}$ [16]) is correlated with a matrix of geographic distances between sampling locations (solid red straight lines) and landscape distances (landscape features such as percentage of forest between sampling locations) or landscape resistance (e.g. length of least-cost path [17] assuming a high resistance value of forests to gene flow; broken red lines) in Mantel or partial Mantel tests [7]. (b) The overlay technique uses
several methods to cluster individuals into groups (e.g. Bayesian clustering; groups indicated by open and red circles [5]), to detect genetic discontinuities or barriers (thick red line [36]) or to interpolate genetic distances among individuals or populations (e.g. kriging resulting in genetic isolines given in red [1]). These genetic groups, barriers, genetic discontinuities or isolines are overlaid on topographical or land cover/land use maps to search for spatial coincidences of these genetic structures with landscape elements. (c) Contemporary gene flow events or migrants (red arrows indicating direction and abundance) can be assessed by either parentage analysis (paternity or maternity analysis [58]) or assignment tests [36]. Any number of matrices of landscape distances between sampling locations connected by migration or gene flow is correlated with a matrix of the frequency of contemporary gene flow in multiple linear regression using permutation for significance testing. Note that in all of the above examples, forests are hindering gene flow.
Figure 2. Conceptual characteristics of landscape genetic studies assessing adaptive genetic variation. (a) A molecular marker linked to a genomic region under directional selection shows a change in allele frequency along an environmental gradient (solid red line). By contrast, neutral molecular markers show no such change on a small-spatial scale (hence, unrestricted gene flow; broken lines) because they are not affected by natural selection (modified from [96]). (b) If directional selection occurs on a larger spatial scale, geographic distance and environmental gradients often co-vary. In consequence, allele frequencies at neutral loci will randomly change with distance and, indirectly, along the environmental gradient because of restricted gene flow and genetic drift (i.e. isolation by distance [7]). Therefore, landscape genetic studies on adaptive genetic variation might falsely identify some neutral markers as bearing the signature of adaptive evolution, that is, spatial genetic structure is a nuisance factor in analysis (Box 3). (c) The effect presented in (b) is most prominent if spatial genetic structure due to phylogeography or population history has lead to strong changes in allele frequencies. Here, changes in allele frequencies of markers showing signs of adaptive evolution are expected to be more pronounced than those of neutral loci.
**Box 1. Five current hot topics in landscape genetics of plants**

(i) Perform landscape distance/resistance analysis to detect the influence of major landscape elements on gene flow by seed and pollen, that is correlate genetic distance with landscape elements (e.g. forests, mountain ridges) or abiotic factors (e.g. wind direction), and make use of statistical methods alternative to (partial) Mantel tests.

(ii) Expand the use of overlays in plants and incorporate boundary overlap statistics.

(iii) Combine estimates of contemporary gene flow by pollen or seed with landscape data in multiple regression analysis with permutation testing.

(iv) Evaluate the influence of spatial genetic structure and population history in outlier detection or allele distribution analyses.

(v) Describe and prove molecular function of identified outlier genomic region and provide empirical tests of the selective relevance of identified adaptive genetic markers in plants (e.g. transplant experiments).

**Box 2. Landscape genetic approach to identify molecular markers bearing the signature of natural selection**

Landscape genetics tries to identify molecular markers whose changes in allele frequencies are correlated with environmental factors potentially acting as selective pressures and enforcing directional natural selection (Figure 2a). Usually, many samples are taken along environmental gradients or in environmentally heterogeneous situations, and large genome scans with hundreds to thousands of co-dominant or dominant markers (AFLPs or SNPs [16]) are performed. Finally, allele presence/absence (individual-based analysis) or allele frequencies (population-based analysis) are correlated with environmental variables taken from geo-referenced databases or from field surveys [10]. Markers significantly correlated with environmental factors are considered to be linked to or to be located within genomic regions under selection, whereas uncorrelated molecular markers are considered as neutral, at least with respect to the particular set of environmental variables tested.

Various statistical methods are used to establish allele distribution models [9]. First, logistic regression relates allele occurrences with environmental variables. The spatial analysis method
(SAM [87]) offers a user-friendly framework to perform logistic regression in a landscape genetic approach. SAM has successfully been used in several animal studies [14, 97], but we are aware of only a single plant study that has applied SAM. Parisod and Joost [98] examined patterns of selection in populations of buckler mustard (*Biscutella laevigata*) characterized by different population histories. Logistic regression has also been used to correlate allele frequencies at an outlier locus associated with altitude and temperature in common beech (*Fagus sylvatica*) [99]. The logistic regression approach can easily be extended to more sophisticated generalized linear models [92]. Recently, we successfully applied multiple linear regression to identify AFLP fragments correlated with temperature and precipitation in the alpine rock cress (*Arabis alpina*) [13]. Polynomial transformation of environmental variables can be included in multiple linear regression, thus also tracking non-linear adaptive responses.

The correlative landscape genetic approach to adaptive genetic variation has several shortcomings. (i) Researchers can only identify correlations of molecular markers with those environmental factors that were included in the analysis. Molecular markers not correlated with these environmental factors are therefore not necessarily neutral in a general sense. For this reason, researchers might also wish to apply classical outlier locus detection to their data sets if population-based sampling is available [97]. (ii) Given that a potentially large number of statistical tests is applied in explorative data analysis, it is important to adjust significance values for type I error inflation due to multiple testing and/or using model evaluation criteria such as adjusted $R^2$ or the Akaike information criterion (AIC) values [33]. The use of adjusted $R^2$ and AIC is currently explored in landscape genetic studies of plants.
Box 3. Spatial genetic structure as a nuisance factor in landscape genetic studies on adaptive genetic variation

Spatial genetic structure caused by (i) restricted gene flow and leading to isolation by distance patterns (Figure 2b) or (ii) phylogeographic history, range expansion or population demography (e.g. bottlenecks; Figure 2c) might substantially interfere with both population genomic outlier analysis and landscape genetic approaches. Pronounced genetic structure is thus a nuisance parameter in landscape genetics of adaptive genetic variation [9] and [89]. Excoffier et al. [89] showed that ignoring hierarchical spatial genetic structure in classical outlier detection analysis results in the identification of numerous false outlier loci. This study highlights the need for adequate treatment of spatial genetic structure when searching for molecular markers linked to genes under selection.

The consideration of spatial genetic structure in landscape genetics of adaptive genetic variation has only just begun. So far, researchers have tended to apply Bayesian clustering to their samples and then use landscape genetic analysis with logistic or linear regression [13, 84] within each genetic cluster separately. Recently, more sophisticated approaches to account for spatial genetic structure have been introduced. For instance, mixed linear models allow controlling for population structure when populations are known [100]. Accordingly, we used generalized estimating equations (GEE), which take small-scale autocorrelation of samples into consideration [92], and we [13] applied principal coordinates of neighbor matrices (PCNMs [101]) to landscape genetic analysis using R [102]. PCNM values on large- and small-spatial scales can be introduced as additional factors in linear regression and account for the effects of different spatial scales and for the effects of un-accounted environmental factors. Despite such promising new tools, it is obvious that the issue of spatial genetic structure in landscape genomic research needs more attention in future analyses as well as the development of appropriate statistical methods.
Acknowledgments

If you have ever watched the movie Hangover, then you will remember the memorable quote about the wolf pack given by Alan on top of Cesar’s palace in Las Vegas:

“You guys might not know this, but I consider myself a bit of a loner. I tend to think of myself as a one-man wolf pack. But when my sister brought Doug home, I knew he was one of my own. And my wolf pack... it grew by one. So there... there were two of us in the wolf pack... I was alone first in the pack, and then Doug joined in later. And six months ago, when Doug introduced me to you guys, I thought, ”Wait a second, could it be?” And now I know for sure, I just added two more guys to my wolf pack.”

Just like Alan, I learned is that you might start off alone on your PhD but as time goes by the people supporting you grow…

so here I make a toast to my wolfpack...

Felix Gugerli for his wit, fast thinking and his help in guiding me along my journey. I really appreciated your detailed explanations of many complicated scientific facts and theories.

Rolf Holderegger for his leadership, for always having an open ear and putting the finishing touches on every manuscript.

Rene Graf for his relentless ways in helping me establish my microsatellites, his immense help in the field (for lending his hand when I had crazy ideas like planting seedlings out in a snow storm). Also thank you for keeping me grounded when I was on the brink of loosing myself in too many details, for being a friend and for introducing me to the art of mushroom gathering.

Sabine Brodbeck for her help in the lab and her extensive knowledge on anything from sports, hiking to field food.

Daniela Keller, Ivo Widmer, Leah Kamm, Mathias Koller, and Tsipe Aavik I enjoyed many great conversations with my office buddies. Thanks for all the laughs, the frustrating stories, and the accomplishments that we shared.

Sarah Bryner sometimes I needed to clear my head and the best cure was a nice long run. Thanks for being my running buddy, your encouraging comments on my manuscripts and your words of wisdom.

Simone Prospero thanks for reading my papers and generally making the days go by easier especially towards the end of the Thesis. I could always knock on your door, exchange stupid stories, and share a laugh.

Corine Schöbel and Sandra for helping with the German translation, apparently having to write German as my mother tongue on every official document is just formality.

Alex Boesch and Yolanda Zimmerman for their various help such as in sampling, carrying plants up a mountain and introducing me to wild camping. Most of all thank you for the laughs.
The Conservation biology FE at WSL there are too many of you to mention by name but I want to thank you all for the support I received.

Simone to my sister who probably doesn’t know it but deserves a medal for all the support she has given me. Stay the way you are.

Mom and Dad for their encouragement and making me explain what I do so many times that in the end I even understood it.

Tricia, Martina, Lauren and Rubi for being great roommates and friends, for being my window to the world outside and for their understanding when I had frustrated moments throughout the years.

Genetic Diversty Center for their technical help.

Biochange members I want to thank the BioChange team for their innovative input and their help in the field.

CCES for sponsoring my PhD work.
Curriculum Vitae

Dominique Buehler
Born May 31st, 1983 in Baden, Switzerland
Citizen of Sigriswil, Bern, Switzerland

2008-2011
Ph.D. thesis, Swiss Federal Institute of Forest, Snow and Landscape research WSL
Doctoral thesis on *Adaptive genetic variation and gene flow potential in the alpine plant, Arabis alpina*
Supervised by Dr. Felix Gugerli, WSL; Prof. Dr. Rolf Holderegger, WSL and ETH Zurich, and Prof. Dr. Alex Widmer, ETH Zurich

2006-2008
MSc. Environmental Science, University of Zurich, Switzerland
Master thesis on *Colonization and local adaptation of Carex acuta in an endangered lakeshore community at Lake Constance*
Supervision by Dr. Jasmin Joshi, University of Zurich; Dr. Markus Peintinger, University of Zurich; and Prof. Dr. Bernhard Schmid (University of Zurich)

2002-2006
B.Sc. Biology, University of North Carolina at Greensboro, USA
Research study on *Aging in honey bee workers*
Supervision by Prof. Dr. Olav Rueppell, University of North Carolina at Greensboro

1998-2002
Havelock High School, Havelock, NC, USA