Hybridization in the *Triticum -Aegilops* complex

Nils ARRIGO

PhD Thesis, April 17th 2009

University of Neuchâtel – Faculty of Science – Institute of Biology – Laboratory of Evolutionary Botany, Neuchâtel, Switzerland

Experts committee:

Roberto Guadagnuolo University of Neuchâtel, Switzerland

François Felber University of Neuchâtel, Switzerland

Philippe Küpfer University of Neuchâtel, Switzerland

Felix Kessler University of Neuchâtel, Switzerland

Beat Keller University of Zürich, Institut für Pflanzenbiologie, Switzerland

Jacques David SupAgro Montpellier, UMR Recherche Diversité et Adaptation des Plantes Cultivées, France

Université de Neuchâtel

FACULTE DES SCIENCES
Secrétariat-Décanat de la faculté
- Rue Emile-Argand 11
- CP 158
- CH-2009 Neuchâtel

IMPRIMATUR POUR LA THESE

Hybridization in the Triticum-Aegilops complex

Nils ARRIGO

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel,
sur le rapport des membres du jury

MM. P. Küpfer (directeur de thèse), F. Felber, R. Guadagnuolo, F. Kessler,
B. Keller (Zürich) et J. David (Montpellier F)

autorise l'impression de la présente thèse.

Neuchâtel, le 23 avril 2009

Le doyen :
F. Kessler

UNIVERSITE DE NEUCHATEL
FACULTE DES SCIENCES
Secrétariat - décanat de la faculté
Rue Emile-Argand 11 - CP 158
CH-2009 Neuchâtel
Felix Kessler

■ Téléphone : +41 32 718 21 00 ■ E-mail : secretariat.sciences@unine.ch ■ www.unine.ch/sciences

Key words transgene, risk assessment, crop-to-wild gene flow, Mediterranean basin, human activities, trade routes, agriculture, sympatry, introgression, wheat, ploidy, AFLP scoring, autogamy, ***Aegilops geniculata***, ***Aegilops neglecta***, ***Aegilops triuncialis***, ***Aegilops cylindrica***.

Mots-clés transgene, evaluation des risques, flux de gènes entre plantes domestiquées et plantes sauvages, Bassin Méditerranéen, activités humaines, routes commerciales, agriculture, sympatrie, introgression, blé, ploïdie, scoring AFLP, autogamie, ***Aegilops geniculata***, ***Aegilops neglecta***, ***Aegilops triuncialis***, ***Aegilops cylindrica***.

Abstract

The present study answers the need of evaluating the putative transfer of (trans-)genes from crop plants to their wild relatives. We first review the biological features and the consequences of a gene flow between crop plants and their wild counterparts. The ***Triticum*** L. – ***Aegilops*** L. species complex is used as a model. Wheat, with 607 million tons harvested annually is one of the major crop species cultivated nowadays. ***Aegilops*** species, the wild relatives of wheat, include several wheat ancestors and are regularly used by breeders as a source of beneficial traits. The reproductive isolation between both taxa is incomplete. F1 hybrids and first backcrossing generations are known either from natural occurrences or from experimental crossings. The ***Triticum-Aegilops*** complex provides an Eurasian model to investigate crop-to-wild gene flow. In addition, it has worldwide implications since at least three ***Aegilops*** species have a weedy behavior and recently colonized the United-States of America.

The study relies mainly on genomic screening (AFLP), a valuable tool to unravel rare events such as the introgression of wheat genes in the ***Aegilops*** genomes. This marker system provides a large amount of genetic information but requires the assistance of automated analysis programs. We present here our own analysis solution, RawGeno, an R CRAN library that allows a quick and reliable scoring of AFLP results. The study includes data mainly acquired under natural conditions in European agrosystems. It focuses on three ***Aegilops*** species commonly occurring in the Mediterranean basin, ***Ae. geniculata*** Roth, ***Ae. neglecta*** Req. ex Bertol. and ***Ae. triuncialis*** L.

We first investigate the mating system of ***Aegilops*** species. Based on genetic diversity estimates and an outcrossing experiment, we confirm the autogamous status of ***Ae. geniculata*** and highlight a facultative autogamy in ***Ae. neglecta*** and ***Ae. triuncialis***.

Wheat-to-***Aegilops*** gene flow is detected through an indirect approach that compares samples collected along wheat field borders to samples originating in areas isolated from agriculture. The results vary according to ***Aegilops*** species: ***Ae. geniculata***, the most autogamous species included in the study, shows no clear evidence of gene flow from wheat, except for one ***Ae. geniculata*** x ***Triticum*** F1 hybrid discovered in a Spanish population. In contrast, an unexpectedly high amount of wheat genetic markers are observed in ***Ae. neglecta*** and ***Ae. triuncialis*** samples. Interestingly, most of the affected samples originate from populations collected near to crop cultivations. Finally, ***Triticum turgidum*** (the pasta wheat) is the most likely wheat parent of introgressed ***Aegilops***. ***Triticum turgidum*** and the three ***Aegilops*** species investigated in the present study are tetraploid

organisms and share the same chromosome number (2n = 4x = 28). This feature might have enhanced the fertility of F1 hybrids and favoured the introgression of tetraploid wheat genes in the ***Aegilops*** genomes.

The phylogeography of ***Ae. geniculata*** is investigated in the fifth chapter. It reveals biogeographic patterns shaped by ancient and large scale processes such as post-glaciary recolonization and possibly the expansion of agriculture or trade routes. Interestingly, the study outlines several recent migration events that disturbed the genetic diversity of ***Ae. geniculata***. These events were probably triggered by human activities and correspond to millenary trade routes of the Mediterranean basin. Finally, this study reveals recombination events involving migrant and native genotypes, a result that mitigates the autogamy of ***Ae. geniculata***.

The sixth chapter reports attempts to simulate the introgression of a transgene in ***Ae. cylindrica***. This experimental approach succeeds to transfer a transgene into two successive generations (F1 and BC1) and confirm results obtained from previous studies.

TABLE OF CONTENTS

Acknowledgments

En tant que travail de recherche, cette thèse fut un ensemble de bons moments, de mises au point et de questions stimulantes. Ce fut aussi l'occasion de remises en question et de doutes. Cette expérience m'a fait grandir et je tiens à remercier les personnes sans qui l'aventure n'aurait pu avoir lieu :

François Felber et Roberto Guadagnuolo, en m'accordant leur confiance pour mener à bien ce travail, m'ont fait entrer dans l'aventure. Leurs conseils et leur appui lors de la rédaction de cette thèse de doctorat ont étés précieux.

Philippe Küpfer, son humanité et sa bienveillance m'ont mis sur le chemin de la Botanique et de la génétique.

Sven Bürki, pote de lycée, d'uni, puis de bureau, à la fois ami et collègue, avec qui il est difficile de décrocher tant il y a à discuter.

Nadir Alvarez, pour son énergie, son esprit scientifique et pour la formation qu'il m'a donné.

Jacques David et Jean-Claude Dusautoir pour leur précieux soutien Montpelliérain.

Beat Keller et Felix Kessler pour les commentaires et les corrections apportées à cette thèse, en tant que membres du Jury.

Marta Cifuentes et à Nicola Schoenenberger pour leurs discussions éclairées concernant ***Aegilops***.

Pierre Selenet, Isabel Maria Casillas, Juan B. Gallego Fernandez, Abelardo Apparicio, Jesus Rodriguez et Maria Isabel Casillas pour leur aide lors de la planification de collectes en France et en Andalousie.

Elisabeth, Laurent, Sylvian, Edouard, Claire-Aline, Nathalie ainsi que le reste de l'équipe du Jardin botanique de l'Université de Neuchâtel pour leur aide irremplaçable lors des mises en cultures de mes avalanches d'***Aegilops***.

Ernest, Neil, Mélanie, Sophie, Ilhan, Mei Lin, Maxime, Anahi, Jérôme, Astrid, Léonore, Xavier, Marianne, German, Ricardo, Christophe, Aude, Julien, Josiane et Lorenza pour leur appui logistique lors de la mise en place de cette thèse.

Anouk pour sa présence et ses bons conseils au lab.

Sylvain pour les bons délires d'Andalousie.

Un immense merci à Celia pour m'avoir aimé et soutenu au cours de cette aventure.

Mes parents et ma famille, sans qui l'invasion du jardin familial à grands coups de cageots d'***Aegilops*** n'aurait été possible. Leur aide logistique et leur soutien ont toujours étés présents.

Finalement un merci tout particulier à Marc Burgat, pour sa passion de la Nature qu'il a su partager avec des générations d'écoliers. Nulle doute qu'il a suscité des vocations ; je peux en témoigner.

Context

The present PhD thesis was included in the Thematic Group 2 of the NCCR Plant Survival of the University of Neuchâtel, Switzerland. The objective of this thematic group was to assess the consequences of a transgene transfer from crop plants to their wild relative species. Our group used the ***Triticum*** - ***Aegilops*** complex as a model. The works on ***Aegilops*** started at the University of Neuchâtel with the PhD thesis of Roberto Guadagnuolo (2000), followed by the works of Nicola Schoenenberger (2005). The present work mainly presents data acquired in natural conditions by investigating ***Aegilops*** populations collected in agroecosystems of the Mediterranean Basin.

Chapters outline and summary

The present document is organized in six chapters, written as separate works. The two first chapters have already been published during the 2007-2009 period[1,2]. We refer to van Slageren (1994)[3] for the taxonomy of ***Aegilops*** species.

The chapter one is a literature review of the genetic and ecological consequences of a transgene transfer from crop to wild relative species[1]. It develops four main aspects:

- The biotic and abiotic conditions that allow a crop-to-wild gene transfer.
- The case of Switzerland, by reviewing pairs of "crop-wild" species where the transfer of genes may spontaneously occur.
- The importance of gene flow that connects related species, especially in the Poaceae family. The resulting genetic networks are discussed in regard to their possible role in enhancing the potential spread of a crop transgene between several related wild species.
- The ecological and genetic consequences of transgene escape within populations of wild relatives.

The chapter two covers methodological aspects. It describes bioinformatics tools developed in response to the use of new molecular methods. The present PhD relies on Amplified Fragment Length Polymorphisms (hereafter AFLP), a reliable system, which generates a large number of markers. The data are stored on a numeric support and require specific softwares in order to provide a proper valorization of the results. Due to the large number of samples that were investigated, we developed our own automated analysis solution: RawGeno[2], an R CRAN library. We evaluated the accuracy of our program using a model dataset (***Cerastium uniflorum***) and the present chapter describes lessons learned from our calibration trials. In the following chapters, RawGeno ensures the analysis of most of

[1] Felber F, Kozlowski G, Arrigo N, Guadagnuolo R (2007) Genetic and ecological consequences of transgene flow to the wild flora. Adv Biochem Eng Biotechnol 107:173-205.

[2] Arrigo N, Tuszynski JW, Ehrich D, Gerdes T, Alvarez N (2009) Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. BMC Bioinformatics 10.

[3] van Slageren MW (1994) Wild wheats: A monograph of ***Aegilops*** L.and ***Ambylopyrum*** (Jaub and Spach) Eig (Poaceae). Wageningen: Agricultural Univ., Wageningen, The Netherlands and Int. Cent. for Agric. Res. in Dry Areas, Aleppo, Syria.

our molecular analyses. Presently, the library is used by other teams of the University of Neuchâtel and is involved in nine projects.

The chapter three is the starting point of our biological considerations regarding ***Aegilops***. It investigates the mating system of three ***Aegilops*** species: ***Ae. geniculata*** Roth, ***Ae. neglecta*** Req. ex Bertol. and ***Ae. triuncialis*** L. We use population genetics, based on AFLP markers, and an outcrossing experiment to address the question of the commonly assumed autogamy of these species. Our data suggest different mating systems in the three ***Aegilops*** species and ***Ae. geniculata*** is considered to be the most autogamous of the three. ***Ae. geniculata*** appears to be the most structured species with 34% of its genetic variation occurring between the populations. In comparison, ***Ae. neglecta*** and ***Ae. triuncialis*** have only 13% and 20% of genetic variation measured between the populations. Similar results are obtained with genetic diversity measures where ***Ae. geniculata*** appears to be the less diversified species. The outcrossing experiment partly corroborated these results but lacked of statistical power to draw stronger conclusions.

The chapter four investigates the transfer of wheat genes into ***Aegilops*** species. The study is focused on ***Ae. geniculata***, ***Ae. neglecta*** and ***Ae. triuncialis***. For the three species, the sample design compares individuals collected along wheat field borders to samples originating from areas isolated from agriculture, and includes 71 wheat reference lines. The aim was to:

- Detect AFLP bands that are specific to wheat samples.
- Assess and quantify their presence in ***Aegilops*** samples.
- Check whether wheat proximity could increase the occurrence of wheat specific AFLP bands into the ***Aegilops*** samples.

For ***Ae. neglecta*** and ***Ae. triuncialis***, it appears that samples collected along wheat fields have significantly more AFLP bands associated to wheat genotypes. This provides solid evidence of transfer of wheat genes into these two ***Aegilops*** species. In addition, introgressed samples display “normal” morphological phenotypes, suggesting that several backcrossing events had occurred. In contrast, ***Ae. geniculata*** shows no clear evidence of wheat genes introgression; except for a F_1 hybrid, discovered at the edge of a Spanish wheat field. This F_1 hybrid results from a crossing with the bread wheat, ***T. aestivum***. Finally, the most likely wheat parent of the introgressed ***Aegilops*** is ***T. turgidum***, the pasta wheat, a tetraploid plant. Since pasta wheat shares the same ploidy level as the three studied ***Aegilops*** species, one may consider that a better fertility in ***Aegilops*** x ***T. turgidum*** hybrids compared to those of ***Aegilops*** x ***T. aestivum*** hybrids may explain our results.

The chapter five combines AFLP genotyping and sequencing of chloroplast markers (cpDNA) to address the phylogeographic history of ***Aegilops geniculata***, a widespread Mediterranean wild relative of wheat. A total of 438 samples, belonging to 143 populations collected around the Mediterranean basin are included in the present study. Globally, the polymorphisms detected with AFLP and cpDNA markers are congruent and reveal a basal dichotomy in the genetic diversity of ***Ae. geniculata***. The species is organised in two main phylogeograpic clusters; populations collected in Northern

Mediterranean areas (i.e. from Portugal to the Balkans) are genetically distinct from those occurring in Southern Mediterranean locations (i.e. from Bulgaria and Palestine to Morocco). Populations from Northern Mediterranean appear to be less diversified but more genetically structured than those collected in Southern Mediterranean areas: I. We measured significant correlations between the genetic and the geographic distances in populations collected in the Iberian and the Franco-Italian Peninsulas. II. Samples collected in the Iberian and the Franco-Italian Peninsulas belong to two distinct genetic entities. III. Several additional local genetic groups are detected in Spain, France and Italy. In contrast, samples collected in Southern Mediterranean areas do not show any genetic patterns. These results are discussed according to hypotheses that consider either I. a natural expansion or II. an agriculture-mediated colonization of the Mediterranean basin by ***Ae. geniculata***.
Finally, our study reveals recent migration events with samples collected in Northern Mediterranean areas that show genotypes specific to Southern Mediterranean areas. Several of these migrants outcrossed with the local populations, a result that contradicts the commonly assumed strict autogamy of ***Ae. geniculata***.

Finally, chapter six presents several crossing experiments conducted with transgenic wheat and ***Ae. cylindrica***. The natural hybridization between ***Aegilops cylindrica*** Host and the bread wheat (***Triticum aestivum*** L.), as well as the spontaneous occurrence of backcrossed progenies are nowadays largely documented. This process, leading potentially to the introgression of wheat (trans)genes within ***Ae. cylindrica*** represents a major agronomical problem in the United States. Here, we simulate an introgressive series between ***Ae. cylindrica*** and six conventional or transgenic "Bobwhite" bread wheat lines. First generation hybrids were produced using emasculated ***Ae. cylindrica*** florets and the obtained pentaploid hybrids were backcrossed in an experiment simulating natural conditions and using ***Ae. cylindrica*** as pollen donor. Fitness components of the two produced generations were estimated measuring the fertility of spikes and the survival of seedlings. In addition, we investigate the genetic transmission of the transgenes and three molecular markers specific to the "Bobwhite" line using PCR based techniques. The transmission of genetic markers to the backcrossed samples remains difficult to assess due to the low number of available plants. Finally, issues regarding the production of third generation of backcrossed individuals are discussed.

Adv Biochem Engin/Biotechnol (2007) 107: 173–205
DOI 10.1007/10_2007_050

Published online: 15 March 2007

Genetic and Ecological Consequences of Transgene Flow to the Wild Flora

François Felber[1] (✉) · Gregor Kozlowski[2] · Nils Arrigo[1] · Roberto Guadagnuolo[1]

[1]Laboratoire de Botanique évolutive, Institut de Biologie, Université de Neuchâtel, rue Emile-Argand 11, 2009 Neuchâtel, Switzerland
Francois.Felber@unine.ch

[2]Department of Biology and Botanical Garden, University of Fribourg, ch. du Musée 10, 1700 Fribourg, Switzerland

Abstract Gene flow from crops to wild relatives by sexual reproduction is one of the major issues in risk assessment for the cultivation of genetically engineered (GE) plants. The main factors which influence hybridization and introgression, the two processes of gene flow, as well as the accompanying containment measures of the transgene, are reviewed.

The comparison of risks between Switzerland and Europe highlights the importance of regional studies. Differences were assessed for barley, beet and wheat. Moreover, transgene flow through several wild species acting as bridge (bridge species) has been up to now poorly investigated. Indeed, transgene flow may go beyond the closest wild relative, as in nature several wild species complexes hybridize. Its importance is assessed by several examples in Poaceae. Finally, the transgene itself has genetic and ecological consequences that are reviewed. Transgenic hybrids between crops and wild relatives may have lower fitness than the wild relatives, but in several cases, no cost was detected. On the other hand, the transgene provides advantages to the hybrids, in the case of selective value as a Bt transgene in the presence of herbivores. Genetic and ecological consequences of a transgene in a wild species are complex and depend on the type of transgene, its insertion site, the density of plants and ecological factors. More studies are needed for understanding the short and long term consequences of escape of a transgene in the wild.

Keywords Risk assessment · Transgene · Genetically engineered plants · Bridge species · Switzerland

Abbreviations
GE genetically engineered
FOEN Swiss Federal Office for the Environment

1 Introduction

The gene transfer from crops into populations of wild relatives has become an important scientific and public issue since the development and cultivation of genetically engineered (GE) plants in the late 1980s [1]. The concerns related to the cultivation of GE crops, in particular those dealing with the possibility of transgene escape into the wild flora, have generated a multitude of studies on crop-to-wild gene flow [1–9]. While these studies have shown that such gene flow exists for almost all of the most important crops cultivated worldwide, only recently have new studies focused and are focusing on its ecological and genetic consequences [9–11]. Yet, in order to better assess the ecological and agronomic risks associated with the transgene flow to the wild flora, it is fundamental to understand the mechanisms and the consequences of such gene flow [12].

Studies on the existence of crop-to-wild gene flow have already been reviewed several times in the context of the cultivation of GE crops (e.g. [2, 5, 7]). A general overview on the factors influencing gene flow, and containment measures is presented here. Risk of gene flow has a geographical component and we focus in Sect. 2 on the particular case of Switzerland. While gene flow has been mostly investigated from crops to their closest wild relatives, further introgression may occur between wild species. The importance of such "bridge species" is explained in Sect. 3. Finally, the genetic and ecological consequence of transgene flow is evaluated in Sect. 4.

1.1 Factors Influencing Gene Flow

It is widely accepted that hybridization between two taxa depends on several key factors, such as their sympatry, the synchrony of their flowering periods, the existence of a common vector for the gametes, as well as their reproductive compatibility and the viability and fertility of the hybrids [12]. Generally speaking, gene flow between two taxa can thus be viewed as a two step process: (i) a first hybridization event, which leads to the production of first generation hybrids, followed by (ii) the introgression of part of the genome of one species into the other by successive backcrosses.

Hybridization depends mainly on straightforward conditions, such as the need for the plants to grow close to each other and the potential for the exchange of pollen. While the most obvious situation where both conditions are met is represented by cultivated fields where the wild relatives grow in close proximity, it is worth noting that crop plants growing as volunteers within fields of other crops or in other habitats represent additional contact zones between crops and wild relatives. In the case of GE crop, such a situation may lead to transgene escape to the wild flora. A notable example of this latter situation is that of rapeseed, which is extremely common to see in any kind of disturbed habitats even relatively distant from cultivated areas [6].

More generally, the establishment of feral crop populations in the agroecosystems, as well as outside the cultivated areas depends mainly on the crop features, such as seed dispersal by wind, water or animals, absence of dormancy, ripening period, persistence of seeds in the soil [13, 14]. Agricultural practices (harvesting period, crop rotation, till vs. no-till) as well as post-harvesting procedures (transportation), can also greatly influence the emergence of volunteer plants.

Finally, an additional potential source of transgenes is represented by first and subsequent generations of hybrids between GE crops and wild relatives, which can act as "genetic bridges" between the parental species [5, 9].

1.2 Factors Influencing Hybridization

Hybridization is influenced quantitatively by numerous factors, some of them depending on the characteristics of the plants, while others are more related to the environment. Hybridization is frequent in perennial species and especially for outcrossing and clonal plants [2], as the produced hybrids can subsist clonally even in the case of reduced fertility.

Pollen vectors play a major role, at least on the distance at which hybridization can take place. For instance, maize pollen is known to be particularly heavy and intraspecific gene flow at distances greater than 50 meters is un-

likely [15]. In contrast, other wind-pollinated species can show large distance pollen dispersal events. Watrud et al. [16] discovered intraspecific hybrids of *Agrostis stolonifera* 21 km from the pollen source.

While it is obvious that topography influences winds, a flat land favoring the pollen flow over long distances, it is worth mentioning that microtopography seems to have also an impact on the behavior of pollinator insects, by hiding or making more visible potential pollen sources and sinks. However, predictions on the pollen movements seem more complicated in the case of insect pollinated species. Different experimental and modeling studies on the distance at which rapeseed pollen could produce hybrids, generated indeed inconsistent results [17, 18], because these results depend indirectly on the factors influencing the activity of bees [19].

Repeated contacts with crop populations are known to accelerate the introgression process [20]. However, hybridization as well is positively correlated with the frequency and the extent of the contact zones between crops and their wild relatives. Indeed, feral populations or individual volunteer crop plants will not only increase the area of contact, but also increase the potential for the overlap of flowering periods. For instance, while in central Europe fields of rapeseed usually flower simultaneously in May, it is common to observe volunteers flowering from June till late October.

1.3 Factors Influencing Introgression

Most factors influence both hybridization and introgression. While successful introgression is achieved when genes from one taxon are fixed in another one, several hybrid generations and parental individuals can be involved in the process. All of these individuals and generations can coexist and exchange genes simultaneously for many years [5].

Fitness of hybrids is essential to successful introgression. Moreover, independent of the pollen vector, the intensity and symmetry of pollen flow will determine both the direction of hybridization and the speed of introgression in a sink population. Fixation of genes is known to occur more rapidly in small populations, which are also more prone to act as a pollen sink [21, 22].

Both hybridization and introgression are facilitated in genetically close species, such as crop and prickly lettuce (D'Andrea et al., unpublished) or crop and wild sunflower [23], rather than between more distantly related species like rapeseed and wild radish [24]. The actual introgression of crop genes into the genome will depend greatly on the existence of pre- and postzygotic barriers, which strongly depend on factors linked to the evolutionary divergence between the crop and its wild relative, the incompatibilities being generally higher between genetically distant taxa and lower between closely related taxa.

Genetic barriers acting against hybridization between species are considered by several authors as "semi-permeable" [25, 26]. Individual genes or specific genome regions may be transferred during introgression processes, rather than entire genomes [5]. Moreover, genes from one species may not be uniformly transmitted to another via introgressed generations, as selection does not act homogeneously within genomes.

A factor that influences specifically introgression, rather than hybridization, is the observation that beneficial or neutral traits will be preferentially introgressed, compared to detrimental genes. For example, silenced genes can be kept in recipient genomes, until they are eliminated by genetic drift [23]. Additionally, several linked genes may be transferred together, especially if such complexes carry positively selected genes.

The situation is more complex in polyploids where multiple copies of genes make genetic interactions even more complicated. Moreover, related polyploid species often share only part of the genome (e.g *Triticum aestivum* and *Aegilops cylindrica*, *Brassica napus* and *B. campestris*) and introgression from one species to the other is easier for genes located on the homologous chromosomes, than for genes located in the homeologous ones.

1.4 Containment of Transgenes

One clue which arises from the existing studies on crop-to-wild gene flow is that hybridization between most crops and their wild relatives cannot be avoided [4]. Therefore, if the goal is to impede the transfer of transgenes to the wild flora, gene flow has to be stopped at its source. For this purpose, several strategies, each possessing advantages and drawbacks, have been proposed (most are reviewed in [5]), which are mostly linked to the mechanisms and factors influencing introgression presented above.

Since physical barriers, such as isolation by distance or hedge rows bordering fields appeared rapidly to be inefficient, genetic barriers based on the breeding systems of the crops were investigated. One of the first ideas was to decrease or completely block gene flow via pollen, by favoring apomixis. However, many apomictic species preserve low to moderate sexual seed production, and moderate or high levels of pollen [14].

It was thus suggested to induce male-sterility in GE crops. This system was applied to commercialized *Brassica napus* varieties [27]. In this rapeseed variety, the transgenic construct is induced by a *tapetum*-specific promoter, and produces a cytotoxin (*barnase*). Only anthers express the lethal transgene, which leads to the destruction of the mother cells of pollen. However, male sterility does not prevent the formation of hybrids when wild relatives act as paternal parent, like in the case of bolting beets in south Europe [4]. Moreover, these two strategies can only prevent gene flow by pollen, while they have no effect on gene flow by seeds.

It was subsequently suggested to insert transgenes in genomic regions, which have no or reduced mobility. As mentioned previously, genomes are not uniformly transmitted, and some regions are more "mobile" than others [25, 28]. Targeting gene insertion in regions poorly transmitted should decrease the probability of gene escape. However, in order to be efficient, this strategy has to be developed on a case-by-case basis, and introgressive patterns on all possible wild relatives of each crop should be known. A similar idea was proposed for polyploid species, where genomes non-shared by wild relatives could be chosen as insertion sites of transgenes (e.g. [29, 30]). However, recombination events between non-homeologous genomes were observed in wild x crop hybrids involving *Brassica napus* [31], and *Triticum aestivum* [32, 33].

Another proposition was to insert transgenes in the DNA of mitochondria or chloroplast, as organellar DNA is usually maternally transmitted, and should not be carried by pollen grains in Angiosperms [34]. However, paternal inheritance of chloroplasts has already been observed (reviewed by [35]). For instance, transfer of genes from organelles to nucleus occurs at a low frequency in tobacco, as one pollen grain out of 16 000 carries cytoplasmic genome elements in its nucleus [36]. As for the strategies presented above, gene flow via seeds is not prevented.

Therefore, so-called "seed suicide" techniques were proposed (see [12] for a review). In these plants, the transgenic construct induces the production of lethal protein or blocks physiological functions during seed maturation, which makes it impossible for the seeds to germinate, but without disturbing albumen differentiation. However, producing non-germinating seeds would impede farmers from sowing part of their harvest, which is a highly controversial issue from an ethical point of view.

Another recent technique consists of the chemically induced removal of transgene from pollen cells during the gametogenesis. The transgene is flanked by specific sites (*lox*), which allows its removal by a site-specific recombinase (Cre). The recombinase is coded by the transgene and expressed after induction [37]. Recombinase-based techniques present currently two major drawbacks: the controlling system has to be activated by an external signal, that is the application of tetracycline, and basically every single cell involved in the sexual reproduction of the crop should be treated.

Finally, post-hybridization and fitness-based strategies were also suggested to avoid the spread of hybrid derivates in the environment. The idea is to lower the fitness of these plants by linking the transgene with traits which are neutral or beneficial in an agricultural context, but detrimental in the wild. The genes responsible for traits such as dwarfing, loss of dormancy or non-shattering of seeds were proposed as suitable loci to place transgenes [5]. However, there are at least two serious drawbacks in this strategy. First, the current technology does not allow placing of the transgenic construct in

a precise location. Second and probably more important, most of these so-called deleterious traits are recessive loss-of-function alleles related to the domestication of crops [13]. These alleles would thus not be expressed in first generation hybrids with a wild plant, because of the presence of the dominant wild allele in their genome. In further generations, the deleterious allele would only be expressed in homozygous individuals, which would strongly reduce its capability to lower the fitness of these plants. Moreover, if the hybrids are fertile, this strategy would not prevent them acting as a genetic bridge and pollinating the wild parent [11].

Alternatively, this strategy could have a good efficacy when the transgene is coupled within the transgenic construct itself with one or two mutant genes conferring an ecological disadvantage (transgenetic mitigation, [38]), such as dwarfing, as demonstrated in tobacco introgressants [39].

2 Gene Flow between Cultivated Plants and Wild Relatives: the Case of Switzerland

Risks related to transgenic plants are often investigated on a worldwide scale and several reviews have focused on this topic. Nevertheless, a regional perspective is necessary because crops vary among countries, wild species have often a limited geographical range and floras composition changes geographically. Consequently, the distribution of crops and their ability to cross with their wild relatives vary regionally. Moreover, the genetic characteristics of a wild species, as for example its ploidy level, may vary according to their geographical range and can influence largely their ability to hybridize. This is illustrated for example by tetraploid alfalfa, *Medicago sativa* ($2n = 32$ chromosomes). In Switzerland, its wild relative, *Medicago falcata*, is tetraploid and has the same chromosome number ($2n = 32$) except in Unterengadin, where it is diploid ($2n = 16$). Hybrids between the two species, *M. x varia*, are found frequently where both species are tetraploid, but are, on the contrary, very rare in the range of the diploid *M. falcata* [40]. Risks of gene flow are consequently much lower in Unterengadin than in the other areas of Switzerland.

Consequently, the results of one country cannot be necessarily generalized to another country without further investigations. This is particularly true for Switzerland, where topography strongly influences the distribution of wild species and constrains agriculture. Its landscape typically illustrates that risks may vary from one area to the other.

Distribution of wild relatives may also vary in time. For example, global change, including both the global warming and the increase in disturbance as a consequence of human activity, has led to the northern expansion of several Mediterranean species. Similarly, change in agricultural practices may

influence the contact zones between crops and their wild relatives, and consequently influence greatly the risks. Therefore, monitoring over a long term the wild flora and the agricultural areas is necessary in order to evaluate the risks on a regional perspective. Switzerland has voted on November 27, 2005 a moratorium of 5 years on the outdoor cultivation of GE organisms for commercial purposes. Probabilities of large-scale cultivation of transgenic plants are therefore low. Nevertheless, political changes may occur rapidly and therefore, assessment of potential risk of gene flow from crops to wild relative is necessary with a Swiss perspective.

The Swiss Federal Office for the Environment (FOEN) granted a study on risk assessment which focused on the main cultivated plants of Switzerland. For each of them, bibliographical data were collected on the crop and on most of the wild relatives (Table 1). From this, the risks of transfer of transgene to conventional varieties and to the wild or naturalized flora were evaluated (Table 2).

Risks of gene flow were never null between cultivars, as all crops reproduce sexually. Risks were evaluated as null for the wild flora when the crop produces no feral populations and no wild relative exists in Switzerland. It was low to medium in the case of autogamy, or of harvest before flowering (as for lettuce, out of the seed production areas). Risk was considered as high for all allogamous species, those forming spontaneous or subspontaneous populations, or possessing wild relatives that hybridize readily with the crop.

Risks may be examined in different perspectives. For a monitoring program, the priority is to examine commercialized transgenic crops. Prior to the authorization of outdoor cultivation, it is important to evaluate on one hand the risks of contamination of non-transgenic cultivation, and on the other hand those of gene flow to the wild flora.

Table 1 Characteristics collected for the crops and its wild relatives

Common for the crops and its wild relatives	Specific to the cultivated plant	Specific to the wild relatives
Latin name	Extent of cultivation	Ecology
Vernacular names	Feral populations	Hybridization with the crop
Chromosome number	Frequent transformations which have led to a request for a field trial	Hybridization with other wild relatives
Pollen dispersal	Recent transformation	Category of threat according to the Swiss Red List [104]
Breeding system	GE field cultivation	Stability of the distribution
Longevity	Commercialization	
Levels of vegetation		

Table 2 Summary of risks for the main crops cultivated in Switzerland

Name		Crop to crop gene flow*	Gene flow with spontaneous and naturalized flora*		Commer-cialization**	Main transformation
			Suisse	Europe		
Poaceae						
Agrostis stolonifera L.	Creeping bentgrass	++	++	++	pending	Herbicide tolerance, agronomical properties
Avena sativa L.	Oats	+	++	++	FT	Virus Resistance (virus BYDV)
Cynodon dactylon (L.) Persoon	Bermuda grass	++	+	+	FT	Agronomical properties, herbicide tolerance
Festuca arundinacea Schreber s.l., *F. pratensis Hudson s.l.*	Fescue	++	++	++	FT	Product quality, fungal resistance
Hordeum vulgare L.	Barley	+	0	+	FT	Product quality, fungal resistance
Lolium perenne Hudson s.l., *L. multiflorum* Lamarck	Ryegrass	++	++	++	FT	Product quality
Poa pratensis L.	Smooth meadow-grass	++	++	++	FT	Herbicide tolerance
Triticum aestivum L., *Triticum spelta* L.	Wheat	+	+	++	FT	Herbicide tolerance, fungal resistance
Zea mays L.	Maize	++	0	0	Com	Insect resistance, herbicide tolerance

Table 2 (continued)

Name		Crop to crop gene flow*	Gene flow with spontaneous and naturalized flora*		Commercialization**	Main transformation
			Suisse	Europe		
Rosaceae						
Fragaria x ananassa	Strawberry	+	+	+	FT	Fungal resistance, herbicide tolerance
Malus domestica Borkh.	Apple	++	++	++	FT	Product quality (fruit quality), insect and bacterial resistance
Prunus avium L.	Cherry	++	++	++	FT	Modification of metabolism
Prunus domestica L.	European plum	++	++	++	pending	Virus resistance
Pyrus communis L.	Pear	++	++	++	FT	Product quality (fruit maturation)
Rubus idaeus L.	Raspberry	++	++	++	FR	Virus resistance, product quality
Asteraceae						
Cichorium intybus L.	Chicory	++	++	++	Com	Agronomical properties (male sterility), herbicide tolerance
Helianthus annuus L.	Sunflower	++	+	+	FT	Fungal resistance, insect resistance
Lactuca sativa L.	Lettuce	+	+	+	FT	Herbicide tolerance, product quality

Table 2 (continued)

Name		Crop to crop gene flow*	Gene flow with spontaneous and naturalized flora*		Commer-cialization**	Main transformation
			Suisse	Europe		
Fabaceae						
Glycine max L.	Soybean	+	0	0	Com	Herbicide tolerance, product quality
Medicago sativa L.	Alfalfa	++	++	++	Com	Herbicide tolerance, product quality
Pisum sativum L.	Pea	+	+	+	FT	Herbicide tolerance, virus resistance
Solanaceae						
Lycopersicon esculentum Miller	Tomato	+	0	0	Com	Product quality, insect resistance
Nicotiana tabacum L.	Tobacco	++	0	0	Com	Product quality, virus resistance
Solanum tuberosum L.	Potato	+	0	0	Com	Insect resistance (doryphore), product quality
Other families						
Beta vulgaris L.	Beet	+	0	++	Com	Herbicide tolerance, virus resistance
Brassica napus L.	Rapeseed	++	++	++	Com	Product quality (oil quality), herbicide tolerance
Brassica rapa L.	Rape	++	++	++	FT	Insect resistance (lepidopters), herbicide tolerance
Cucumis melo L.	Melon	++	0	0	FT	Virus resistance, product quality (fruit ripening)

Table 2 (continued)

Name		Crop to crop gene flow*	Gene flow with spontaneous and naturalized flora*		Commercialization**	Main transformation
			Suisse	Europe		
Cucumis sativus L.	Cucumber	++	0	0	FT	Virus resistance, agronomic properties (salt tolerance)
Cucurbita pepo L.	Squash	++	0	0	Com	Virus resistance
Daucus carota L.	Carrot	++	++	++	FT	Fungal resistance (*Alternaria* tolerance), product quality
Dianthus caryophyllus L.	Carnation	++	0	0	Com	Product quality (colors modification)
Osteospermum ecklonis (DC) Norl	Cape Daisy	++	0	0	FT	Metabolism modification
Picea abies (L.) Karsten	Norway Spruce	++	++	++	FT	Gene marker
Pinus sylvestris L.	Scots pine	++	++	++	FT	Gene marker, forestry performance
Populus alba x tremula, *Populus* sp.	Poplar	++	+	+	FT	Herbicide tolerance, forestry performance
Vitis vinifera L., *Vitis labrusca* L.	Grape	++	+	+	FT	Fungal resistance, virus resistance

* Evaluation of risks: 0 = no risk, + = low or medium risk, ++ = high risk

** Commercialization: FT = field tests have been carried out; pending = commercialization of transgenic varieties is pending; Com = transgenic varieties are commercialized

2.1
Priority Species for a Monitoring Program

Eleven cultivated species in Switzerland possess commercialized transgenic varieties elsewhere in the world. Six of them present a high risk of gene flow with other cultivars (alfalfa, carnation, chicory, maize, rapeseed and squash). Others represent a lower risk as they are harvested before flowering, such as beet, or because seed do not mature in the regions, such as potato for example. Moreover, crops with an autogamous breeding system such as soybean, tobacco or tomato also present a lower risk of gene flow.

2.2
Potential Risks for the Contamination of Non-Transgenic Crops

Only crops which have commercialized GE varieties are mentioned below. The higher risks originate from the six allogamous species mentioned above. Medium risks are characteristics from either autogamous species with partial allogamy, or those which are not producing fruits in traditional practices (beet, potato). Low risks exist for plants that do not flower, when vegetative parts are collected. Such cultivation necessitates a good management and strict control, in order to avoid any loss of seeds or unintended flowering. Some of the species mentioned above belong to that category, depending on their use.

2.3
Potential Risks for Gene Flow to the Wild or Naturalized Flora

High risk characterizes crops and wild relatives with no or low reproductive barriers, as for oilseed rape and creeping bentgrass. For example, escape of transgenic creeping bentgrass (*Agrostis stolonifera* L.) in non-agronomic areas was observed in the USA [41]. Medium risk occurs if the hybrid is partially fertile and introgression is possible. No commercialized transgenic crops belong to that category: pea, poplar, strawberry, sunflower and wheat. Some cultivated species have no wild relative in Switzerland; this is for example the case for beet, carnation, maize, melon, potato, soybean, tobacco, tomato and squashes. Consequently, they do not represent a genetic threat for the natural flora, even if containment measures are needed to avoid crop to crop gene flow.

2.4
Particularity of the Swiss Flora

Table 2 reveals that, for some species, different risks were assessed between Switzerland and Europe. Barley present no risk for Switzerland, as no an-

cestor grows in this country [42], while in the eastern Mediterranean to Iran and West Central Asia, hybridization occurs readily with *Hordeum spontaneum* [4]. Wheat presents also a lower risk in Switzerland, where only *Ae. cylindrica* forms durable populations, contrasting with the Mediterranean area where several wild relatives of *Aegilops* are frequent. Finally, beet presents no risk of outcrossing with the wild flora because its wild relative *Beta vulgaris* subsp. *maritima* is absent in Switzerland, while it is present close to the Atlantic coast and along the Mediterranean boarder.

3 The Importance of Bridge Species

Historically, the term "bridge species" has been used to designate wild plant species which could act, through artificial or natural hybridization, as a genetic bridge between wild relatives and closely related cultivated plants. Figure 1 shows that there are potentially three possible directions of the gene flow through bridge species: (1) wild-to-crop bridges, (2) crop-to-wild bridges and (3) wild-to-wild bridges.

The wild-to-crop bridges have been used by humans since millennia and are still used by breeders for the introduction of desirable traits from wild relatives into crops [43]. As discussed above, the development of GE crops has brought much more attention to the gene flow the other way around, that is between cultivars/crops and their wild relatives [2, 5, 7]. Surprisingly, the potential further spread of transgenes to other wild relatives via wild-to-wild

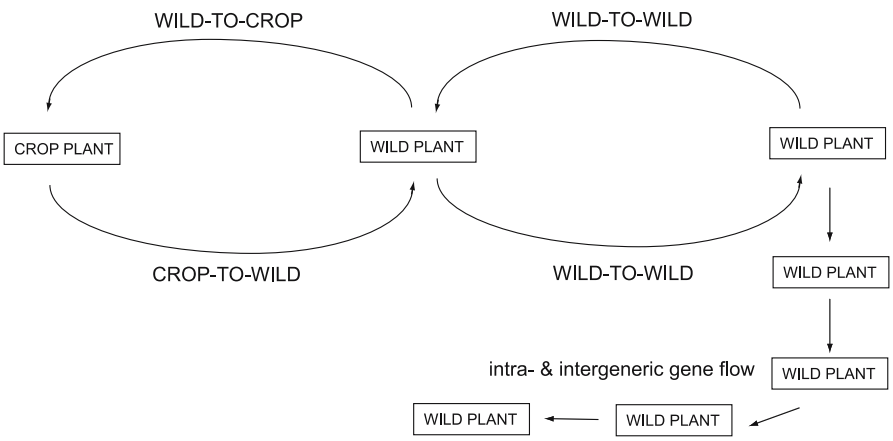

Fig. 1 Bridge species and directions of gene flow in crop-wild hybrid complexes. The potential spread of transgenes into wild populations via wild-to-wild bridges and further introgression has been poorly investigated

bridges and so-called stepping-stones introgression has been rarely if ever studied in details in the GE plants context. Thus, after a short description of the first well studied and described gene flow direction, more attention will be devoted to the still weakly explored subject of wild-to-wild bridges.

3.1 Wild-to-Crop Bridges

The term "bridge species" designates here wild relatives of cultivated plants which are used during artificial and/or natural hybridization procedures for crop improvement to circumvent some experimental or environmental constrains (Fig. 1). The ability to transfer genes between related plant species has been a great benefit in the improvement of cultivars for disease resistance, insect resistance, and/or end-use quality. This has been especially true in allopolyploid crops where there are multiple species that can act as donors. The best documented examples come from studies of gene transfer from wild species to wheat (*Triticum aestivum*). Romero et al. [44] obtained for example the transfer of a cereal cyst nematode resistance gene from *Aegilops triuncalis* (donor) to hexaploid wheat using bridge species *T. turgidum*. Fernandes et al. [43] transferred to wheat stem and leaf rust as well as powdery mildew resistance from *Ae. squarrosa* (donor) through hybridization with *T. durum* (bridge species). Such methods imitate in fact the ancient hybridization events, which happened during evolution and domestication of some crop plants, e.g. the hexaploid wheat. This bridge species method with development of intermediate natural or artificially synthesized amphiploid hybrids is one of the available procedures to facilitate gene flow between wild relatives and crop. It has been used for many decades not only for wheat cultivars [45] but also for many other crops (e.g. *Brassicaceae* [46], *Gossypium* sp. [47], *Cucumis* sp. [48]). However, for numerous plant groups such approaches are very laborious and/or have low or no success (e.g. for some *Solanum* sp. [49]).

3.2 "New Old Issue": Wild-to-Wild Bridges and Stepping-Stones Introgression

The hybridization and introgression between wild plants is a very well known phenomenon. Ellstrand et al. [2] estimated that there are more than 1000 well studied and published examples of spontaneous plant hybridization. Although at generally low frequencies and over long periods of time, genes (and thus also transgenes) can be spontaneously introgressed between different wild species [5]. It is therefore surprising that there are practically no detailed studies and exhaustive reviews on the importance of wild-to-wild hybridizations and wild-to-wild bridges in the context of the transgene flow and GE crops (Fig. 1).

How common are the natural hybridization processes between wild plant taxa? To answer this question we have to remember that there are two possible outcomes of hybridization [1]. The first outcome is the present ongoing introgression. Mallet [50] based on fundamental work on hybrid flora of the British Isles by Stace [51] estimated that at least 25% of all wild plant species are able to hybridize spontaneously and/or are involved in ongoing introgression processes with other wild species. Ellstrand et al. [52] using the same data set concluded that up to 34% of families and 16% of all genera in Great Britain have at least one reported hybrid. Additionally, there are many very well-studied genera with numerous closely related species producing hybrid swarms, such as *Salix* [53], *Quercus* [54] or *Eucalyptus* [55]. Rieseberg [56], based on the calculations of Ellstrand et al. [52], concluded that we could expect a worldwide total of 27 500 hybrid combinations among all Angiosperms. He added however, that it could be strongly underestimated since many regions, especially the tropics, are weakly explored and documented as far as their hybrid flora is concerned.

The second result of hybridization is the ancient and present speciation [56–58]. Indeed, in many families and genera, polyploidization and hybridization were the main mode of speciation and diversification. Ellstrand et al. [2] based on the summarizing works of Grant [59] and Arnold [60] concluded that more than 70% of plant species originated from hybrids.

As a consequence of these two well-documented hybridization outcomes, it has been often stated that the natural interspecific and even intergeneric hybrid formation is ubiquitous and uniform among higher plants [61, 62] or even the rule rather than the exception [63]. However, Ellstrand et al. [52] demonstrated clearly that the spontaneous hybridization is non-randomly distributed among systematic plant groups. By analyzing five biosystematic floras from Europe, North America and the Hawaiian Islands they showed that certain phylogenetic groups are predisposed for hybridization. To the most important hybrid families in practically all analyzed regions belong such crop-plant families as Poaceae, Asteraceae, Rosaceae and Fabaceae. Ellstrand et al. [2] enumerated 13 of the most important food crops grown for human consumption. Among them seven belong to Poaceae (*Eleusine, Hordeum, Oryza, Saccharum, Sorghum, Triticum, Zea*), and three to Fabaceae (*Arachis, Glycine, Phaseolus*). Hybridization seems therefore to be concentrated in a relatively restricted fraction of families and/or genera. Moreover, many members of these highly hybridizing families have been genetically modified and mainly possess numerous wild relatives. Table 3 lists all major plant genera and families of European flora containing crop plants with reported genetic transformation and/or with GE species used for field trials. It shows how many wild relatives of transformed crops could be found in Europe and which of those taxa possess the highest ability for complex hybridization. Here again the family of Poaceae has the most important potential for wild-to-wild bridge formation. Numerous of its mem-

Table 3 Number of wild relatives in European flora of the most important crop plants used for genetic transformation (only genera with GE members commercialized or used for field trials are listed) and estimation of their potential for wild-to-wild bridge formation. Symbols: – hybridization not observed, + hybridization observed but rare, ++ hybridization frequent, +++ hybridization extremely frequent (based on [51, 71], for *Lolium, Festuca* and *Poa* see also [112–160])

Genus	Nb. of wild (naturalized) species in Europe*	Potential for wild-to-wild bridge formation in Europe		
		intrageneric	intergeneric	hybridizing with
***Poaceae*:**				
Agrostis	24 (1)	+++	++	*Polypogon, Calamagrostis*
Avena	10 (2)	++	–	–
Cynodon	1	–	–	–
Festuca	165	+++	++	*Lolium, Vulpia*
Hordeum	8 (1)	++	++	*Agropyron, Elymus*
Lolium	5	+++	++	*Festuca*
Poa	43 (1)	+++	–	–
Triticum	3	+++	+++	*Aegilops, Elymus, Secale*
Zea	0	–	–	–
***Rosaceae*:**				
Fragaria	4 (1)	+	–	–
Malus	6	++	++	*Pyrus, Sorbus*
Prunus	19 (2)	++	–	–
Pyrus	11	++	++	*Malus, Sorbus*
Rubus	c. 75 (c. 3)	+++	–	–
Asteraceae:				
Cichorium	3	–	–	–
Helianthus	3 (7)	+	–	–
Lactuca	15	++	–	–
***Fabaceae*:**				
Glycine	0	–	–	–
Medicago	35 (2)	+	–	–
Pisum	1	–	–	–
***Solanaceae*:**				
Lycopersicon	0	–	–	–
Nicotiana	3(4)	–	–	–
Solanum	3 (9)	+	–	–
Other families:				
Beta	5	+	–	–
Brassica	20	+++	+++	*Raphanus, Sinapis, Diplotaxis, Hirschfeldia, Eruca, Erucastrum*
Cucumis	(1)	–	–	–
Cucurbita	0	–	–	–

Table 3 (continued)

Genus	Nb. of wild (naturalized) species in Europe*	Potential for wild-to-wild bridge formation in Europe		
		intrageneric	intergeneric	hybridizing with
Daucus	10	+	–	–
Picea	2 (c. 8)	–	–	–
Pinus	13 (c. 13)	–	–	–
Populus	4 (c. 6)	+++	–	–
Vitis	1 (c. 9)	+	–	–

* wild: native species including cultivated species capable of formation of weedy subspontaneous populations; naturalized: non-native species naturalized in Europe.

bers form easily both intra- and intergeneric hybrids. The only groups which could be compared with Poaceae are some genera of the family Rosaceae (mainly fruit trees) and the very well-known *Brassica* coenospecies complex.

It is worth mentioning that detailed studies and surveys on natural hybridization in wild taxa are extremely difficult. Abbott [64] and Ellstrand et al. [52] pointed out that the main limitation is the scarcity of modern biosystematic floras containing complete ecological, evolutionary and genetic information needed for such surveys. Additionally, the documentation of hybridization and introgression faces several methodological and theoretical difficulties. There are many methods used in identifying hybrids ranging from relatively simple morphological measurements to complex molecular and phylogenetic analyses. However, the majority of them, if not all, suffer from the fact that there are multiple explanations for the morphological and/or molecular intermediacy of a given hybrid candidate taxon [57]. The morphological similarity for example could be simply a result of convergent evolution. Martinsen et al. [26] concluded that the hybrid detection based on morphological characteristics is additionally constrained by backcrosses, since it is known that a backcrossed hybrid often resembles the parental species. The development of molecular genetic markers has facilitated studies of hybridization and allowed one to detect even very low levels of introgression. Additionally, it is possible with the molecular markers to track both the nuclear and cytoplasmic gene flow. However, the presence in one individual of molecular markers from two different species could be explained not only through recent hybridization but also due to shared ancestral characters (symplesiomorphy [26, 57]). The differentiation between contemporary versus ancient introgression is difficult and has been studied in only a few taxa [26, 65].

3.3 *Poaceae*: Example of a Biologically Predisposed Family for Wild-to-Wild Bridge Formation

Table 3 and detailed comparative studies mentioned above [2, 51, 52] demonstrate clearly the enormous potential of the family *Poaceae* for the wild-wild hybridization processes. The importance of the *Poaceae* as an object of research reflects their ecological and biogeographical success as well as their enormous economic value [66]. They occupy almost every habitat around the world, often being the dominating organisms [67]. The *Poaceae* comprises about 10 000 species and between 600 and 900 genera [68, 69]. In addition to that, the family contains a very high percentage of species and cytotypes of polyploidy origin. More than 80% of grass species have undergone polyploidy which represents the highest percentage in Angiosperms. Such a high level can be explained by successive regressions and extensions of the ranges which would favour secondary contact zones between related taxa, their hybridization and their subsequent polyploidization [70].

According to Wipff [71] one of the most important grass groups being currently used in genetic transformations are the forage grasses as well as grass species used for turf and erosion control. Furthermore, Wipff gives four main reasons why this group is particularly at risk of spreading transgenes: (1) they have undergone relatively little domestication; (2) they have usually numerous wild relatives; (3) they grow often in sympatry with these; (4) they can grow as weeds outside cultivated areas or in other crop cultures. To this grass group belong such common and species-rich European and North American genera as *Lolium, Festuca, Poa* and *Agrostis*. In the United States not less than 187 field tests were carried out between 1993 and 2006 with transgenic *Agrostis stolonifera*, 36 with *Poa pratensis*, 26 with *Festuca arundinacea*, 17 with *Cynodon dactylon* and 6 with *Lolium perenne* [72]. All mentioned species possess numerous wild relatives in Europe (Table 3) and are capable of hybridizing easily with them (e.g. *Festuca* with ca. 165 species, numerous subspecies and swarms of hybrids in Europe).

Figure 2 gives an example of intrageneric wild-to-wild hybrid complexes in *Poa*. Genus *Poa* contains approximately 43 species in Europe (Table 3) and 300 species worldwide. Intergeneric hybridization is extremely common and results in serious classification difficulties [71, 73]. In *Poa pratensis*, which absorbed genomes from many different taxa, it is even impossible to trace its ancestors [71, 74]. It was shown additionally that F1 hybrids between different *Poa* species can be completely fertile [75]. Figure 2 shows that almost 1/3 of all European *Poa* species are able to hybridize. They have mainly sympatric distribution even at a local level and have similar phenology. It is additionally very probable that more detailed studies would reveal much higher levels of intrageneric hybridization between members of the genus *Poa*.

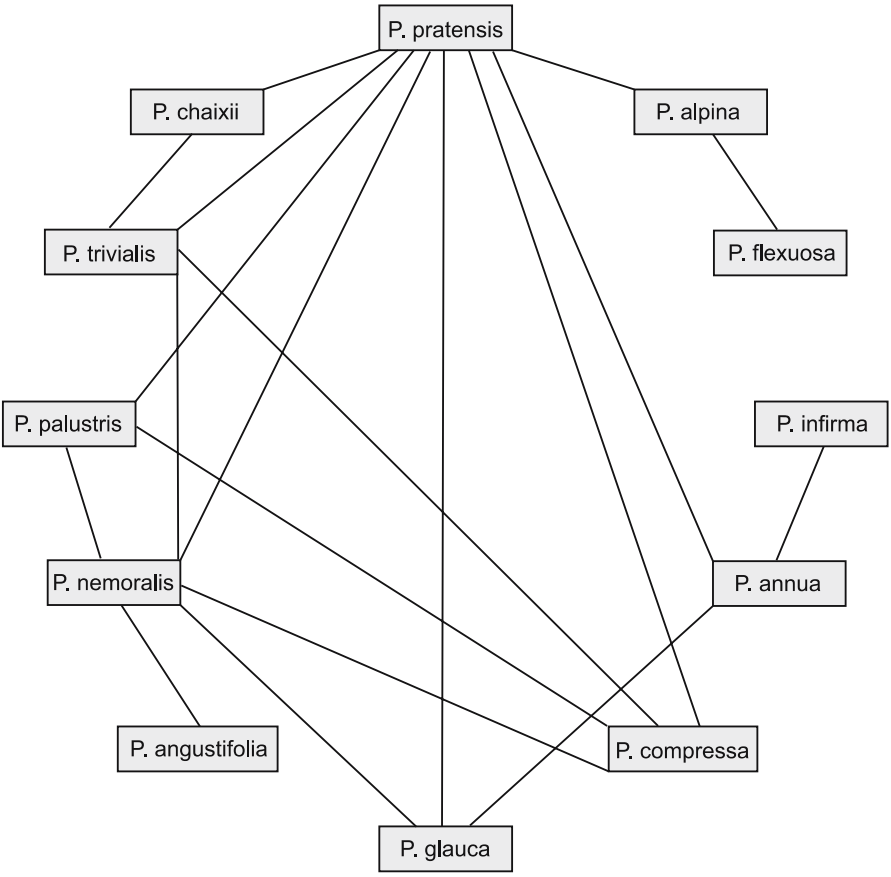

Fig. 2 Wild-to-wild bridges: possible intrageneric hybridization in the grass genus *Poa* [112–160]

Figure 3 gives some further examples of wild-to-wild inter- and intrageneric hybrid complexes in three common European genera of *Poaceae*. The reproductive compatibility and hybrid viability (even at intrageneric level) between *Lolium, Festuca* and *Vulpia* are very well documented (e.g. [51, 76–78]). Intrageneric spontaneous hybrids between *Festuca* and *Lolium* (= x *Festulolium*) are not rare (see also Table 3), they can be fertile and have an ability to backcross with either of the parents [71, 79]. The commonest x *Festulolium* in Europe is the hybrid between *F. pratensis* x *L. perenne* (= x *Festulolium loliaceum*) which can be found in different types of pastures and meadows from Norway to Italy [51, 80, 81]. Figure 3 shows additionally that there are certain species complexes where hybrid combinations are possible in all directions. This is the case for example in the following five species: *Festuca pratensis, F. arundinacea*, *F.*

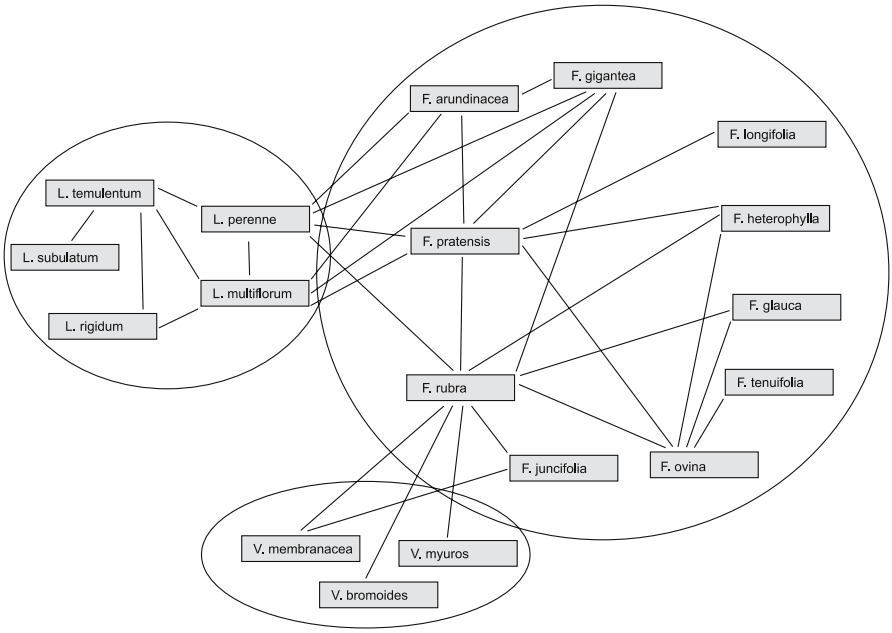

Fig. 3 Wild-to-wild bridges: possible intra- and intergeneric hybridization between selected grass genera: *Lolium*, *Festuca* and *Vulpia* [112–160]

gigantea, *Lolium perenne* and *L. multiflorum*. The close relation of these species could be also demonstrated experimentally [82]. Several authors proposed even to join both genera or to move some *Festuca* species into genus *Lolium* [83, 84]. Additionally, some studies on the chromosome structure of *F. pratensis*, *L. perenne* and *L. multiflorum*, concluded that there are practically no barriers for gene exchange between these species [71, 78].

The majority of species represented in Figs. 2 and 3 fit very well the general characteristic of taxa predisposed for hybridization [12, 52]. They are outcrossing with incomplete reproductive isolation between species, they are mainly perennials with well-developed vegetative spread. Further, they are wind-pollinated and the pollen dispersal up to 21 km has been shown (e.g. for *Agrostis* [16]). Thus, the geographic proximity as well as pollination does not represent any constrains. Additionally, they flower over a very long time period from May till August, thus even at a local scale the phenological overlapping is very common.

The examples described above illustrate clearly that in selected vascular plant families and genera we could potentially expect a stepping-stone spread and exchange of genes with unpredictable effect. Absolute containment of transgenes will be in such taxa practically impossible. Therefore, more experimental and descriptive work has to be done in order to evaluate the

existence and importance of wild-to-wild bridges among a spectrum of taxonomic plant groups as broad as possible.

4
Genetic and Ecological Consequences on Wild Relatives

4.1
Genetic and Ecological Consequences of Outcrossing

Outcrossing in plants may have different impacts, depending on the relatedness of the taxa. When a single species is involved, chromosomes are homologous and pair regularly. On the contrary, when related taxa hybridize, recombinations occur between homeologous chromosomes with the possible consequence of irregular pairing, leading to unbalanced gametes with reduced fertility.

Hybridization between crops and wild relatives is a very ancient phenomenon which has been investigated for a long period from an agronomist point of view, as gene flow from the wild species to the crop might lead to reduced yield and loss of the genetic purity of the cultivated varieties. More recently, while GE plants have been developed, agronomists and ecologists have been concerned by the consequences of transgene escape into non-transgenic crop fields or in wild relatives.

Hybridization has genetic and ecological consequences. Genetic consequences may be defined as the effects of the insertion of the genes in the target species itself and on the expression of genes. On the other hand, ecological consequences are considered here as direct or indirect effects on fitness. We discuss below the two types of consequences separately.

4.2
Consequences of the Transgene

The transgene itself may have genetic consequences for the recipient plants by interacting with other genes and leading to untargeted effects. In order to investigate this aspect, among many others, *Arabidopsis thaliana* has been used as a model species. Metzdorff et al. [85] analyzed, using cDNA microarrays, six independently transformed *A. thaliana* lines characterized by modified flavonoid biosynthesis. Although these transgenic lines possessed different types of integration events, no unintended effects were identified.

Genetic transformation could also affect fitness, and may be in this case associated with a physiological cost. For example, significant reduction of fitness was observed repeatedly associated with resistance to herbicide. Bergelson et al. [86] observed for *A. thaliana* a 34% reduction in seed production for a mutant acetolactase synthetase gene that confers resistance to the herbicide

chlorsulfuron, in comparison to the non-transgenic lines. This cost in fitness was caused by pleiotropic effects due to the presence of the resistance genes itself, while no cost was associated with the expression of kanamycin resistance. Purrington and Bergelson [87] obtained similar results by comparing mutant and transgenic herbicide resistant lines in two different environmental conditions: with or without fertilizer treatments. The cost of resistance appeared in both treatments for the transgenic line, while no cost was associated with the mutant line in the high fertilizer treatment. Other untargeted effects may appear, such as the change in outcrossing rate observed in an outdoor experiment involving transgenic *A. thaliana*, without the proof that it was caused by the transgene itself [88].

Transgenesis may have also unexpected effects on crops. For example, the lignin content of Bt corn was significantly higher than that of non-Bt corn [89]. A change in lignin content might affect the action of herbivores and have ecological consequences.

Moreover, Prescott et al. [90] demonstrated that post-translational modification of a plant protein (α-amylase inhibitor-1 from the common bean (*Phaseolus vulgaris*)) led to the synthesis of a structurally modified form of the protein in pea (*Pisum sativum*). This protein showed altered antigenic properties. While this example concerns human health, it shows that untargeted effects of transgenic plants on protein expressions occur. Consequently, we can infer that similar effects could lead to changes in ecological properties.

4.3 Consequences for Wild Relatives

Introgression involves chromosome segments containing possibly several genes. Therefore, the consequences of introgression will depend on the genes included in the introgressed segment, and on the site of introgression in the recipient species (linkage to other crop genes, pleiotropy). Similarly, introgression from a transgenic crop to a wild relative will also depend on the insertion site of the transgenic line. Genetic consequences are expected to be those of conventional lines, except for the effect of the transgene itself.

4.3.1 Inference from Natural Observations

Natural hybridization is frequent in nature and the fitness of hybrids may be lower, equal or higher than that of their parents [91]. Hybrid inferiority has been recognized as a rule for a long time. More recently, the importance of hybridization for evolution and speciation has emerged (e.g. [2, 59, 60]). Hybrids may be at the origin to new lineages, which may lead to new

species. Speciation might be either progressive or abrupt, when polyploids are formed by chromosome doubling. Because of their different chromosome number, allopolyploids are reproductively isolated, at least partly, from their parents.

According to Burke and Arnold [92], different genetic mechanisms operate behind low and high fitness of the hybrids. Hybrid inferiority would be caused in most cases by negative epistasis, while heterosis would be mostly the consequences of the segregation of additive genetic factors.

4.3.2 Inference from Conventional (Non-Transgenic) Wild x Crop Hybrids

Hybridization followed by repetitive backcrosses lead to introgression, the transfer of a part of the genome of one species to another. Depending on the introgressed genes and on their expression, introgression may lead to new characteristics which could affect the ecological properties of the target species. Experimental data produced a broad spectrum of results on the relative fitness between hybrids and their parents.

The effectiveness of gene flow will depend on the viability and the fertility of hybrids and of subsequent backcrosses. Several studies have involved hybrids between a conventional crop and their wild relatives. For example, Hauser et al. [93, 94] have investigated the fitness of F1 hybrids, as well as F2 hybrids and backcrosses between *Brassica rapa* and oilseed rape (*B. napus*) in experimental crosses. Hybrids were as viable as their parents, produced more pods, but these later contained fewer seeds, with an overall fitness that was intermediary to their parents [93]. The fitness of F2 and backcrosses were on average lower, compared to that of their parents, and varied considerably, including individuals as fit as *B. rapa* [94].

In another study, interspecific F1 hybrids between wild and cultivated radishes (*Raphanus raphanistrum* x *R. sativus*) had a lower fitness than the wild plant. Nevertheless, a field experiment was set up with one half containing F1 wild-crop hybrids and the other half wild. After three years, the dominant white color of the flower of the crop persisted at a frequency ranging from 8% to 22% [95]. A similar study on carrot (*Daucus carota*) showed that hybrids between cultivated and wild carrots were more sensitive to frost than the wild parents, which limited their survival [96].

Weed x crop hybrids between *Sorghum halepense* (Johnsongrass) and *S. bicolor* (cultivated sorghum) did not show any difference in fitness, suggesting that in this case, no barriers to gene flow exist [97]. Contrasting with that, a second generation of hybrids (S1 and BC1) between *Lactuca serriola* (prickly lettuce) and *L. sativa* (cultivated lettuce) germinated and survived better than their wild relative. Seed output of both classes of hybrids was greater with *L. sativa* but no significant difference was found with *L. serriola* [98].

4.3.3
Inference from Transgenic Wild x Crop Hybrids

Table 4 summarizes the results of fitness measurements comparing the hybrids of different generations for five crops, partly derived from Hails and Morley [1]. Non-transgenic hybrids have usually lower, equivalent or intermediary fitness then their parents.

A notable exception is the experiment of Guadagnuolo et al. [11] which demonstrated heterosis of hybrids between glyphosate-tolerant maize (*Zea mays*) and teosinte (*Z. mays* ssp. *mexicana*), when compared to the wild parent. Nevertheless, in the absence of selection pressure with herbicide, no difference was detected between transgenic and non-transgenic hybrids.

Sunflowers have been extensively studied. Burke and Rieseberg [10] investigated transgenic sunflower with an inserted gene of oxalate oxidase (OxOx) conferring enhanced white mold resistance in cultivated sunflower. They backcrossed it with wild sunflower. No cost of the transgene was observed in the absence of the pathogen. When the plants were infected, the transgene decreased the probability of infection, this later having a negative effect on seed output. Moreover, the disease effect varied among locations and no generalization was possible. The authors insisted on the necessity of replicating the experiment over space and time, as well as on the importance of genetic background and of environmental conditions.

Snow et al. [99] demonstrated that, in the field, male sterile wild sunflowers introgressed with a Bt transgene produced more inflorescences than those without the transgene. These advantages were related to a decrease in insect damage. Greenhouse experiments did not reveal any fitness cost of the transgene.

In hybrids with transgenic rapeseed, relative fitness varied considerably, depending on the transgene and on the presence of associated selective pressure by insect herbivores. Transgenic hybrids between wild *Brassica rapa* and rapeseed possessing a Bt transgene performed better than non-transgenic hybrids in the presence of herbivores, while their fitness was lower when herbivores were absent, showing a physiological cost of the transgene [100]. On the contrary, performances were equivalent in hybrids with or without a transgene coding for high laurate content [101, 102]. The same was true for glufosinate tolerance, in the absence of herbicide treatment.

For practical reasons, only one transgenic line was used in most of the cited experiments, instead of using so-called "sister lines", possessing the same transgenic construct but in different insertion sites. It is then delicate to assess the effect of the transgenes themselves on the fitness of hybrids and subsequent backcrosses, because the consequences on fitness may not be due necessarily to the transgene, but could depend on its insertion site. Moreover, environmental conditions and the density of plants may also influence relative fitness [10, 103].

Table 4 Comparisons of fitness of F1 and BCs for some crops; data compiled partly from [1]

		NT (non transgenic), T (transgenic)	Type of transgene	Hybrid generation	Fitness (E=equivalent, H=higher, I=intermediary, L= Lower)	Refs.
Oilseed rape						
Brassica napus	*Brassica rapa*	NT		F1	I	[93]
B. napus	*B. rapa*	NT		F2 and BC1	L	[94]
B. napus	*B. rapa*	T	glufosinate tolerance	F1	female fitness L to H: frequency and density dependent, male fitness L	[103]
B. napus	*B. rapa*	T	Bt	F1	H, in presence of herbivores; L in absence	[100]
B. napus	*B. rapa*	T	Bt	BC2	E (low herbivory)	[105]
B. napus	*B. rapa*	T	high laurate	F1	E	[101]
B. napus	*B. rapa*	T	high laurate	F1	E	[102]
B. napus	*B. rapa*	T	glufosinate tolerance	BC3	E	[106]
B. napus	*B. rapa*	T	glufosinate tolerance	BC1	E	[107]
Sunflowers						
Helianthus annuus	*H. annuus* (wild type)	NT		F1	L	[108]
H. annuus	*H. annuus* (wild type)	T	Bt transgene	BC1	H (natural herbivory)	[99]
H. annuus	*H. annuus* (wild type)	T	white mould resistance	BC3	H to E	[10]

Table 4 (continued)

		NT (non transgenic), T (transgenic)	Type of transgene	Hybrid generation	Fitness (E=equivalent, H=higher, I=intermediary, L= Lower)	Refs.
Sugar beet						
Beta vulgaris ssp. *vulgaris*	*B. vulgaris* ssp. *maritima*	NT		F1, F2	E	[109]
B. vulgaris ssp. *vulgaris*	*B. vulgaris* ssp. *maritima*	T	viral resistance	F1, F2	E	[109]
Squash						
Cucurbita pepo	*Cucurbita pepo* (wild type)	NT		F1, F2 and BC1	L (F1) to E (F2 and BC1)	[110]
C. pepo	*C. pepo* (wild type)	T	viral resistance	F1, BC1 and BC2	L (F1) to H (BC1 and BC2) in case of high disease pressure. L for all in case of low disease pressure	[110]
C. pepo	*C. pepo* (wild type)	T	resistant to two pathogenic viruses	F1	L (survival and seed production)	[111]
Maize						
Zea mays	*Z. mays* ssp. *mexicana*	NT		F1	H then wild species	[11]
Zea mays	*Z. mays* ssp. *mexicana*	T	glyphosate tolerance	F1	H then wild species	[11]

Summarizing, costs of the transgene have been observed in some cases but not in all. On the other hand, it is worth noting that in the mentioned studies, positive effects on fitness of transgenic hybrids has been interpreted as the direct effect of the transgene itself.

5 Conclusion

To date, a considerable amount of data has been gathered on potential and actual gene flow between crops and wild relatives. Such knowledge is extremely useful for the risk assessment associated with GE crops, especially when considering the regional component of the floristic composition. The investigation of the case of Switzerland reveals differences with Europe, which modulate the evaluation of risk for several crops. It reveals also that not only the presence of wild relatives differ geographically, but that their genetic composition may vary and strongly influence gene flow, as was illustrated for alfalfa [40].

Other issues merit further investigation. Indeed, transgenes can be transmitted from crops to the closest wild relatives, but can also migrate further to other species, by successive crosses. A bibliographical survey of Poaceae illustrate that hybridization is widespread in some taxonomical groups.

Genetic and ecological consequences of the transgene in a wild species have also been poorly investigated up to now. The few existing studies show different pictures according to the species and the inserted trait. More investigations are needed to dissociate the importance of the insertion site and of the transgenes themselves. Moreover, the fitness of hybrids may vary according to environmental conditions and these interactions merit evaluation in nature. It is interesting to note that several examples demonstrate that transgene expression give advantages to the wild species in cases of selective pressure, as for the Bt gene in the presence of herbivores. While generalization is difficult, any type of transgene which would influence fitness positively, such as for example the resistance to diseases could confer to the wild species a real ecological advantage.

Finally, given the diversity of the results observed in the various studies on gene flow, and on its consequences, it seems almost impossible to address all the questions experimentally. On the contrary, it is probably necessary to produce enough empirical data, in order to build realistic and reliable predictive models.

Acknowledgements This project was partially funded by the National Centre of Competence in Research (NCCR) Plant Survival, a research program of the Swiss National Science Foundation and by the Swiss Federal Office for the Environment. We especially express gratitude to Dr. Christof Sautter for valuable comments offered on earlier versions of the manuscript.

References

1. Hails RS, Morley K (2005) Trends Ecol Evol 20:245
2. Ellstrand NC, Prentice HC, Hancock JF (1999) Ann Rev Ecol Syst 30:539
3. Halfhill MD, Millwood RJ, Rymer PL, Stewart CN (2002) Environ Biosafety Res 1:19
4. Ellstrand NC (2003) Dangerous Liaisons? When Cultivated Plants Mate with Their Wild Relatives. The John Hopkins University Press, Baltimore
5. Stewart CN, Halfhill MD, Warwick SI (2003) Nat Rev Genet 4:806
6. Wilkinson MJ, Elliott LJ, Allanguillaume J, Shaw MW, Norris C, Welters R, Alexander M, Sweet J, Mason D (2003) Science 302:457
7. den Nijs HCM, Bartsch D, Sweet J (2004) Introgression from Genetically Modified Plants into Wild Relatives. CABI Publishing, Oxon, UK
8. Elliott LJ, Mason DC, Wilkinson MJ, Allainguillaume J, Norris C, Alexander M, Welters R (2004) J Appl Ecol 41:1174
9. Reagon M, Snow AA (2006) Am J Bot 93:127
10. Burke JM, Rieseberg LH (2003) Science 23:1250
11. Guadagnuolo R, Clegg J, Ellstrand NC (2006) Ecol Appl 16:1967
12. Chapman MA, Burke JM (2006) New Phytol 170:429
13. Gepts P (2002) Crop Sci 42:1780
14. Ellstrand NC (2003) Philos Trans R Soc Lond B 358:1163
15. Bannert M (2006) Simulation of transgenic pollen dispersal by use of different grain colour maize. Dissertation no. 16508 Swiss Federal Institute of Technology of Zürich
16. Watrud LS, Lee EH, Fairbrother A, Burdick C, Reichman JR, Bollman M, Storm M, King G, Van de Water PK (2004) Proc Natl Acad Sci USA 101:14533
17. Lavigne C, Klein EK, Vallee P, Pierre J, Godelle B, Renard M (1998) Theor Appl Genet 96:886
18. Damgaard C, Kjellsson G (2005) Agric Ecosyst Environ 108:291
19. Hayter KE, Cresswell JE (2006) J Appl Ecol 43:1196
20. Haygood R, Ives AR, Andow DA (2003) Proc R Soc Lond B Biol Sci 270:1879
21. Ellstrand NC, Elam DR (1993) Annu Rev Ecol Syst 24:217
22. Burgess KS, Morgan M, Deverno L, Husband BC (2005) Mol Ecol 14:3471
23. Whitton J, Wolf DE, Arias DM, Snow AA, Rieseberg LH (1997) Theor Appl Genet 95:33
24. Guéritaine G, Sester M, Eber F, Chevre AM, Darmency H (2002) Mol Ecol 11:1419
25. Rieseberg LH, Whitton J, Gardner K (1999) Genetics 152:713
26. Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Evolution 55:1325
27. Deblock M, Debrouwer D (1993) Planta 189:218
28. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Proc Natl Acad Sci USA 98:11479
29. Zemetra RS, Hansen J, Mallory-Smith CA (1998) Weed Sci 46:313
30. Seefeldt SS, Zemetra R, Young FL, Jones SS (1997) Weed Sci 46:632
31. Halfhill MD, Richards HA, Mabon SA, Stewart CN (2001) Theor Appl Genet 103:659
32. Schoenenberger N, Felber F, Savova-Bianchi D, Guadagnuolo R (2005) Theor Appl Genet 111:1338
33. Schoenenberger N, Guadagnuolo R, Savova-Bianchi D, Küpfer P, Felber F (2006) Genetics 174:2061
34. Daniell H, Datta R, Varma S, Gray S, Lee SB (1998) Nat Biotechnol 16:345
35. Smith SE (1989) Plant Breed Rev 6:361
36. Huang CY, Ayliffe MA, Timmis JN (2003) Nature 422:72
37. Keenan RJ, Stemmer WPC (2002) Nat Biotechnol 20:215

38. Gressel J (1999) Trends Biotechnol 17:361
39. Al-Ahmad H, Galili S, Gressel J (2005) Planta 272:372
40. Savova D, Rufener Al Mazyad P, Felber F (1996) Bot Helv 106:197
41. Reichman JR, Watrud LS, Lee EH, Burdick CA, Bollman MA, Storm MJ, King GA, Mallory-Smith C (2006) Mol Ecol 15:4243
42. Savova-Bianchi D, Keller-Senften J, Felber F (2002) Weed Res 42:325
43. Fernandes MIBM, Zanatta ACA, Prestes AM, Caetano VR, Barcellos AL, Angra DC, Pandolci V (2000) Genet Mol Biol 23:1051
44. Romero MD, Montes MJ, Sin E, Lopez-Braña I, Duce A, Martin-Sanchez JA, Andrés MF, Delibes A (1998) Theor Appl Genet 96:1135
45. Riley R, Kimber G (1966) Annu Rep Plant Breed Inst 1964–1965:6
46. Choudhary BR, Joshi P, Singh A (2000) Theor Appl Genet 101:990
47. Vroh BI, Baudoinm JP, Hau B, Mergeai G (1999) Euphytica 106:243
48. Chen J, Staub J, Qian Ch, Jiang J, Luo X, Zhuang F (2003) Theor Appl Genet 106:688
49. Laferriere LT, Helgeson JP, Allen C (1999) Theor Appl Genet 98:1272
50. Mallet J (2005) Trends Ecol Evol 20:229
51. Stace C (ed) (1975) Hybridization and the flora of the British Isles. Academic Press, London
52. Ellstrand NC, Whitkus R, Rieseberg LH (1996) Proc Natl Acad Sci USA 93:5090
53. Hardig TM, Brunsfeld SJ, Fritz RS, Morgan M, Orians CM (2000) Mol Ecol 9:9
54. Dumolin-Lapegue S, Kremer A, Petit RJ (1999) Evolution 53:1406
55. Griffin AR, Burgess IP, Wolf L (1988) Austr J Bot 36:41
56. Rieseberg LH (1997) Ann Rev Ecol Syst 28:359
57. Rieseberg LH, Baird SJE, Gardner KA (2000) Plant Mol Biol 42:205
58. Hegarty MJ, Hiscock SJ (2005) New Phytol 165:411
59. Grant V (1981) Plant speciation. Columbia University Press, New York
60. Arnold ML (1997) Natural hybridization and evolution. Oxford University Press, New York
61. Raven PH (1976) Syst Bot 1:284
62. Whitham TG, Morrow PA, Potts BM (1991) Science 254:779
63. Stebbins GL (1959) Proc Am Philos Soc 103:231
64. Abbott RJ (1992) Trends Ecol Evol 7:401
65. Marchant AD, Arnold ML, Willinson P (1988) Heredity 61:321
66. Mathews S, Tsai RC, Kellogg EA (2000) Am J Bot 87:96
67. Davies JI, Soreng RJ (1993) Am J Bot 80:1444
68. Dahlgren RMT, Cliford HT, Yeo PF (1985) The families of the monocotyledons. Springer, Berlin Heidelberg New York
69. Clark LG, Zhang W, Wendel JF (1995) Syst Bot 20:436
70. Stebbins GL (1985) Ann Missouri Bot Gard 72:824
71. Wipff JK (2002) Gene flow in turf and forage grasses (*Poaceae*), In: Ecological and Agronomic Consequences of Gene Flow from Transgenic Crops to Wild Relatives, Workshop Proceedings. The Ohio State University, Columbus, Ohio, p 115–133
72. http://www.isb.vt.edu/
73. Stebbins GL (1950) Variation and evolution in plants. Columbia University Press, New York
74. Clausen J (1961) Euphytica 10:87
75. Muntzing A (1940) Hereditas 15:219
76. Melderis A (1955) Proc BSBI 1:390
77. Trist PJO (1971) Watsonia 8:311
78. Jahuar PP (1975) Chromosoma 52:103

79. Terell EE (1966) Bot Rev 32:138
80. Gymer PT, Whittington WJ (1973a) New Phytol 72:411
81. Gymer PT, Whittington WJ (1973b) New Phytol 72:862
82. Terell EE (1979) Taxonomy, morphology, and phylogeny, In: Buckner RC, Bush LP (eds) Tall Fescue, Chapter 3. Agronomy Series No. 20. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Inc., Madison
83. Jaworski A, Sulinowski S, Nowacki E (1975) Genet Pol 16:271
84. Darbyshire SJ (1993) Novon 3:239
85. Metzdorff SB, Kok EJ, Knuthsen P, Pedersen J (2006) Plant Biol 8:662
86. Bergelson J, Purrington CB, Palm CJ, LopezGutierrez JC (1996) Proc R Soc Lond B Biol Sci 263:1659
87. Purrington CB, Bergelson J (1997) Genetics 145:807
88. Bergelson J, Purrington CB, Wichmann G (1998) Nature 395:25
89. Saxena D, Stotzky G (2001) Am J Bot 88:1794
90. Prescott VE, Campbell PM, Moore A, Mattes J, Rothenberg ME, Foster PS, Higgins TJV, Hogan SP (2005) J Agric Food Chem 53:9023
91. Arnold ML, Hodges SA (1995) Trends Ecol Evol 10:67
92. Burke JM, Arnold ML (2001) Ann Rev Genet 35:31
93. Hauser TP, Shaw RG, Østergård H (1998) Heredity 81:429
94. Hauser TP, Jørgensen RB, Østergård H (1998) Heredity 81:436
95. Snow AA, Uthus KL, Culley TM (2001) Ecol Appl 11:934
96. Hauser TP (2002) Conservation Genet 3:75
97. Arriola PE, Ellstrand NC (1997) Ecol Appl 7:512
98. Hooftmann DAP, Oostermeijer GB, Jacobs MMJ, den Nijs HCM (2005) J Appl Ecol 42:1086
99. Snow AA, Pilson D, Rieseberg LH, Paulsen MJ, Pleskac N, Reagon MR, Wolf DE, Selbo SM (2003) Ecol Appl 13:279
100. Vacher C, Weis AE, Hermann D, Kossler T, Young C, Hochberg ME (2004) Theor Appl Genet 109:806
101. Linder CR, Schmitt J (1995) Ecol Appl 5:1056
102. Linder CR, Taha I, Seiler GJ, Snow AA, Rieseberg LH (1998) Theor Appl Genet 96:339
103. Pertl M, Hauser TP, Damgaard C, Jorgensen RB (2002) Heredity 89:212
104. Moser D, Gygax A, Bäumler B, Wyler N, Palese R (2002) Liste Rouge des fougères et plantes à fleurs menacées de Suisse. OFEV, Bern
105. Mason P (2003) Env Biosafety Res 2:263
106. Snow AA, Andersen B, Jorgensen RB (1999) Mol Ecol 8:605
107. Mikkelsen TR, Andersen B, Jorgensen RB (1996) Nature 380:31
108. Snow AA, Moran-Palma P, Rieseberg LH, Wszelaki A, Seiler GJ (1998) Am J Bot 85:794
109. Pohl-Orf M, Morak C, Wehres U, Saeglitz C, Driessen S, Lehnen M, Hesse P, Mücher T, von Soosten S, Schuphan I, Bartsch D (2000) The environmental impact of gene flow from sugar beet to wild beet – an ecological comparison of transgenic and natural virus tolerance genes. In: Fairbairn C, Scoles G, McHughen A (eds) Proc 6th Int Symp Biosafety of Genetically Modified Organisms, Saskatoon, Canada, p 51–55
110. Fuchs M, Chirco EM, Mcferson JR, Gonsalves D (2004) Environ Biosafety Res 3:17
111. Spencer L, Snow A (2001) Heredity 86:694
112. Akerberg E (1942) Hereditas 28:1
113. Akerberg E, Bingefors S (1953) Hereditas 39:1
114. Bangerter EB (1957) Proc BSBI 2:381

115. Benoit PM (1958) Proc BSBI 3:85
116. Benoit PM (1960) Nature Wales 6:59
117. Clapham AR, Tutin TG, Warburg EF (1952) Flora of the British Isles. Cambridge University Press, Cambridge
118. Corkhill L (1945) J Agr 71:465
119. Dale MR, Ahmed MK, Jelenkovic G, Funk CR (1975) Crop Sci 15:797
120. Dandy JE (1958) List of British vascular plants. BSBI, London
121. Druce GC (1919) Rep BEC 5:314
122. Griffiths D (1950) J Agr Sci 40:19
123. Gymer PT (1971) The nature of hybrids between *Lolium perenne* L. and *Festuca pratensis* Huds., PhD Thesis, University of Nottingham
124. Gymer PT, Whittington WJ (1973a) New Phytol 72:411
125. Gymer PT, Whittington WJ (1973b) New Phytol 72:862
126. Hertzsch W (1938) Züchter 10:261
127. Holmberg OR (1922) Hartmans Handbook i Skandinaviens Flora 1:151
128. Hubbard CE (1968) Grasses, 2nd Ed. Penguin Books, Harmondsworth
129. Hylander N (1953) Nordisk Kärlväxtflora 1:356
130. Jahuar PP (1975) Chromosoma 52:103
131. Jenkin TJ (1931) Welsh Pl Breed, Ser H 12:121
132. Jenkin TJ, Thomas PT (1938) J Bot Lond 76:10
133. Jenkin TJ (1954) J Genet 52:282
134. Jenkin TJ (1955a) J Genet 53:442
135. Jenkin TJ (1955b) J Genet 53:467
136. Jenkin TJ (1955c) J Genet 53:94
137. de Langhe J-E, Delvosalle L, Duvigneaud J, Lambinon J, Berghen CV (1973) Nouvelle Flore de la Belgique, du Grand Duché de Luxembourg, du Nord de la France et des Régions voisines. Edition du Patrimoine du Jardin botanique national de Belgique, Bruxelles
138. Lewis EJ (1966) The production and manipulation of new breeding material in *Lolium-Festuca* In: Proc 10th Int Grassland Congr, Helsinki, p 688
139. Linton EF (1907) J Bot Lond 45:296
140. Malik CP, Thomas PT (1966) Chromosoma 18:1
141. Melderis A (1955) Proc BSBI 1:390
142. Melderis A (1957) Proc BSBI 2:243
143. Melderis A (1965) Proc BSBI 6:172
144. Melderis A (1971) Watsonia 8:299
145. Nannfeldt JA (1937) Bot Notiser 1937:1
146. Nilsson F (1935) Hereditas 20:181
147. Nilsson F (1930) Bot Notiser 1930:161
148. Nygren A (1962) Symb Bot Upsal 17:1
149. Patzke E (1970) Biol 51:255
150. Perring FH, Sell PD (1968) Critical supplement to the atlas of the British flora. Nelson, London
151. Petch CP, Swann EL (1968) Flora of Norfolk. Jarrold and Sons Ltd, Norwich
152. Peto FH (1933) J Genet 28:113
153. Stace CA (1975) Hybridization and the Flora of the British Isles. Academic Press, London
154. Stace CA, Cotton R (1974) Watsonia 10:119
155. Terrell EE (1966) Bot Rev 32:138
156. Trist PJO (1971) Watsonia 8:311

157. Tutin TG (1957) Watsonia 4:1
158. Ullmann W (1936) Herbage Rev 4:105
159. Webster MMCC (1973) Watsonia 9:390
160. Willis AJ (1967) Proc BSBI 6:386

Invited by: Professor Sautter

BMC Bioinformatics

Research article

Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring

Nils Arrigo*[1], Jarek W Tuszynski[2], Dorothee Ehrich[3], Tommy Gerdes[4] and Nadir Alvarez[5]

Address: [1]Laboratory of Evolutionary Botany, Institute of Biology, University of Neuchâtel, 11 rue Emile-Argand, CH-2000 Neuchâtel, Switzerland, [2]Science Applications International Corporation (SAIC), 1710 SAIC Drive Suite 3155 McLean, VA 22102, USA, [3]Department of Biology, University of Tromsø, N-9037 Tromsø, Norway, [4]Chromosome Laboratory, Department of Clinical Genetics, Rigshospitalet, Blegdamsvej 9, Copenhagen, Denmark and [5]Laboratory of Evolutionary Entomology, Institute of Biology, University of Neuchâtel, 11 rue Emile-Argand, CH-2000 Neuchâtel, Switzerland

Email: Nils Arrigo* - nils.arrigo@unine.ch; Jarek W Tuszynski - jaroslaw.w.tuszynski@saic.com; Dorothee Ehrich - dorothee.ehrich@ib.uit.no; Tommy Gerdes - tommy.gerdes@rh.regionh.dk; Nadir Alvarez - nadir.alvarez@unine.ch

* Corresponding author

Published: 26 January 2009

BMC Bioinformatics 2009, **10**:33 doi:10.1186/1471-2105-10-33

Received: 18 August 2008
Accepted: 26 January 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/33

Abstract

Background: Since the transfer and application of modern sequencing technologies to the analysis of amplified fragment-length polymorphisms (AFLP), evolutionary biologists have included an increasing number of samples and markers in their studies. Although justified in this context, the use of automated scoring procedures may result in technical biases that weaken the power and reliability of further analyses.

Results: Using a new scoring algorithm, RawGeno, we show that scoring errors – in particular "bin oversplitting" (i.e. when variant sizes of the same AFLP marker are not considered as homologous) and "technical homoplasy" (i.e. when two AFLP markers that differ slightly in size are mistakenly considered as being homologous) – induce a loss of discriminatory power, decrease the robustness of results and, in extreme cases, introduce erroneous information in genetic structure analyses. In the present study, we evaluate several descriptive statistics that can be used to optimize the scoring of the AFLP analysis, and we describe a new statistic, the information content per bin (I_{bin}) that represents a valuable estimator during the optimization process. This statistic can be computed at any stage of the AFLP analysis without requiring the inclusion of replicated samples. Finally, we show that downstream analyses are not equally sensitive to scoring errors. Indeed, although a reasonable amount of flexibility is allowed during the optimization of the scoring procedure without causing considerable changes in the detection of genetic structure patterns, notable discrepancies are observed when estimating genetic diversities from differently scored datasets.

Conclusion: Our algorithm appears to perform as well as a commercial program in automating AFLP scoring, at least in the context of population genetics or phylogeographic studies. To our knowledge, RawGeno is the only freely available public-domain software for fully automated AFLP scoring, from electropherogram files to user-defined working binary matrices. RawGeno was implemented in an R CRAN package (with an user-friendly GUI) and can be found at http://sourceforge.net/projects/rawgeno.

Background

For the past decade, studies on ecology and evolution have relied upon the assessment of genetic diversity in populations and species [1]. Genomic screening approaches for the measurement of diversity are more satisfactory than any other phenotype- or genotype-based techniques in the sense that they reveal a large number of markers [2]. Before the 1990's, restriction fragment length polymorphisms (RFLP [3]), random amplified polymorphic DNA (RAPD [4]) and simple sequence repeats (SSR or microsatellites [5]) were widely used to generate a relatively high number of markers. However, the implementation of amplified fragment length polymorphisms – AFLP [6], a relatively cheap, easy, fast and reliable method [7] – has exponentially increased the number of informative markers, resulting in large datasets. Most of those approaches retrieve genetic information through PCR-based techniques coupled with electrophoretic gels. As a result, markers are identified according to their absolute size that is measured as a function of mobility. The AFLP technique starts by digesting genomic DNA with two restriction enzymes (*EcoR* I and *Mse* I according to the original protocol [6]). This step is followed by the amplification of a subset of the restricted DNA fragments (requiring several intermediate steps, see [7]) through two successive PCR reactions (namely, the preselective and selective PCRs) and the separation of the amplicons by electrophoresis. Amplicons are fluorescently labelled and electrophoresis takes place in a genotyping machine, for instance using the GeneScan technology (i.e. amplicons migrate along a capillary during a span of time proportional to their size). As a result, the reaction conducted on a sample constitutes an "AFLP electropherogram" or "profile", in which each amplicon is recorded as a "peak" that is characterized by its mobility (converted to size and measured in base pairs, "bp") and intensity (measured as relative fluorescence units, "rfu"). The final step of the analysis aims to convert numerous AFLP profiles, that reflect the results of an AFLP reaction conducted on many samples, into a binary matrix where the presence/absence of each amplicon is recorded for each sample. While the GeneScan technology improves the accuracy of the genotyping process, its precision is not absolute and several factors (i.e. biases occurring during restriction, amplification or migration of the fragment in the capillary of the genotyping machine [8]) can affect the recorded size of an amplicon. Consequently, variability of recorded sizes complicates the analysis, as the same amplicon may have similar but non-equal sizes throughout the sampling. Due to these variations, checking and recording the presence/absence of a given amplicon, through the sampling, is generally done manually by the user. This phase is termed "scoring" and is achieved by the following procedures:

I. Defining amplicon size categories (i.e. called "bins") that ideally represent AFLP loci.

II. Recording the presence/absence of an amplicon within each bin and for each sample [9]. As a result, an AFLP locus will be coded as a binary state, where the presence of an amplicon is coded with 1 ("present" allele) while the absence of the amplicon is coded as 0 ("null" allele).

Programs such as Genographer V2.0 [10] (freely available at http://sourceforge.net/projects/genographer [verified on December 30, 2008]) propose a graphical solution by displaying the GeneScan results on a "gel-like" interface and allowing the user to manually define the bins. Although this strategy allows direct control of the scoring, it requires experienced users and remains sensitive to human biases (e.g. see Bonin et al. [11]). Moreover, the procedure becomes problematic for large numbers of samples and adding new samples to the analysis requires a new scoring session. As a consequence, the final results may vary among runs, among users and across time, weakening the reproducibility and reliability of the analysis. These limitations justify the use of partially automated (e.g. Peakmatcher V6.1 [12], freely available at http://perennialgrains.org/wiki/index.php?title=User:Dehaan [verified on December 30, 2008]) or fully automated scoring procedures (e.g. GeneMapper V3.7, Applied Biosystems, Foster City, CA), where the user does not directly score the dataset, but parameterizes a scoring algorithm. At least three main issues must be considered when scoring a dataset (either manually or using automated procedures).

I. Size homoplasy [13], which can arise in two ways: a) when two non-identical amplicons are considered to be homologous because they display identical mobility or b) when two amplicons are scored as absent at the same locus for differing reasons (e.g. amplicon size polymorphism or mutation in the restriction site).

II. Occurrence of bin definition errors resulting from a non-optimal scoring parameterization. This bias can lead to two contrasting errors: either a) "oversplitting", in which bins are too thin and may split variant locations of the same amplicon into smaller and erroneous sub-bins, or b) including an exaggerated range of amplicon sizes within the same bin, thus introducing an artificial similarity between unrelated samples. Although they differ in their causes, we assume that this second bias has comparable consequences for the dataset quality as those caused by size homoplasy. We will therefore term this bias "technical homoplasy" in order to distinguish it from size homoplasy.

III. Difficulties in detecting amplicons due to variable quality of AFLP reactions. Low quality runs can lead to the introduction of a noisy signal (i.e. "false-negatives" or "false-positives") within the dataset. This bias can be limited by using optimized and standardized laboratory AFLP protocols [8] and by running blank samples in order to

determine the background noise associated with the genotyping machine.

In addition, recent studies propose the evaluation of the quality of both bins and alleles in order to increase the final dataset quality. Several of these procedures are applied once the scoring is complete [14]. However, it is also possible to proceed before the scoring phase, while analysing the AFLP profiles, for instance by tuning the peak detection parameters [15].

The scope of the present study covers methodology of automated AFLP scoring in the framework of population genetics and phylogeography. We propose a new automated solution for scoring AFLP electropherograms: RawGeno, a program implemented as a package in the widely used R CRAN freeware. We investigate the effects of sub-optimal settings on our algorithm, by focusing on two upstream processes of the AFLP electropherograms analysis: the bin definition and the recording of alleles. For this purpose, we tuned scoring parameters and produced five datasets differing in the average width of bins. This strategy, applied to the model-species *Cerastium uniflorum* (Caryophyllaceae), produced five datasets with increasing technical homoplasy. In parallel, we produced five analogous datasets with the commercial software GeneMapper. Finally, we scored the dataset manually by using the freeware Genographer. Using these eleven datasets, we first evaluate several descriptive statistics that can be used to optimize the scoring of the AFLP data. We then investigate the effects of the AFLP scoring settings, as well as the choice of the scoring method, on downstream analyses such as data mining statistics (ordination techniques), inter-individual and inter-population distance, Maximum Likelihood clustering (using PSMix [16], an algorithm aimed to investigate patterns of genetic structure) and population diversity indices.

Methods

Technical features of RawGeno

The analysis begins by detecting and calculating the size of peaks within the AFLP profiles. This preliminary analysis is conducted either with GeneScan V3.1.2 (ABI) or with the freeware PeakScanner V 1.0 (ABI, http://www.appliedbiosystems.com/peakscanner [verified on December 30, 2008]) to produce an exhaustive list of detected amplicons generated by the AFLP reaction. This list records the size, fluorescence and sample origin of each amplicon which are used as the input data for RawGeno. It should be noted that our program can potentially be modified to handle results from other genotyping machines.

The bin definition algorithm [17] is the core of our library (see Figure 1. for a thorough explanation). In brief, it begins by defining the locations and limits of the bins. The bins are identified on the basis of the range of amplicon sizes from all samples, regardless of their fluorescence intensity (amplicons are preliminarily filtered according to their intensity during the detection and sizing analysis; e.g. 50 rfu as in our present study). This strategy aims to minimize the bin widths while maximizing the intervals between bins. Amplicons are sorted according to their sizes and size intervals are computed between each consecutive amplicon. Transitions between successive bins are identified by considering the inter-bin size intervals to be larger than the intra-bin ones. Two additional rules are involved in bin definition. First, bins must fit within a user-specified width range (i.e. the "maximum bin width" is parameterized). Second, a bin can include only one amplicon per sample. The application of this second rule is modulated by the "minimum bin width" parameter, in order to allow the occurrence of technical homoplasy. As a result, "thin" bins are accepted only if they effectively co-occur in at least one sample. Similarly, the definition of "wide" bins that would include more than one amplicon per sample within a single bin is prevented as often as possible but remains possible by manipulating the "minimum bin width" parameter. Once bins are defined, the algorithm notes the presence or absence of an amplicon in each bin for each sample and builds a binary matrix. Finally, amplicon features (i.e. fluorescence and size) are used to improve the binary matrix quality (see Additional file 1).

Empirical data and experimental design

In order to investigate the effects of the scoring procedure on genetic analyses, a dataset from an extensive study on intra- and inter-specific plant biodiversity (IntraBiodiv Consortium [18]) was chosen as a model. This dataset includes samples covering the whole geographic range of *Cerastium uniflorum* (Caryophyllaceae), a perennial, diploid (2n = 36) plant distributed throughout the European Alps in subnival habitats. A total of 209 individuals (including 40 individuals that were replicated in the DNA extraction step) from 46 populations (four individuals per population on average) were analysed with three selective AFLP primer pairs. Details on the sampling scheme, the primer pairs used and the scoring methods are provided in Gugerli et al. [18] and Bonin et al. [19]. Raw data were obtained after running AFLP reactions on an ABI 3100 sequencing machine (ABI) and analysing electropherograms using GeneScan V3.1.2 (ABI). The program was set up with default detection parameters; only peaks ranging between 50 and 500 bp with a minimum fluorescence intensity of 50 rfu (i.e. the minimum reportable peak height according to Gilder et al. [20]) were included in the analysis. In the context of the Intrabiodiv Consortium, the scoring was performed manually using Genographer V1.6 [10], through a classical user-defined protocol [18], in which non-reproducible bins were discarded before being recorded (as proposed in Bonin et al. [11]). As a consequence, this manually scored dataset did not include

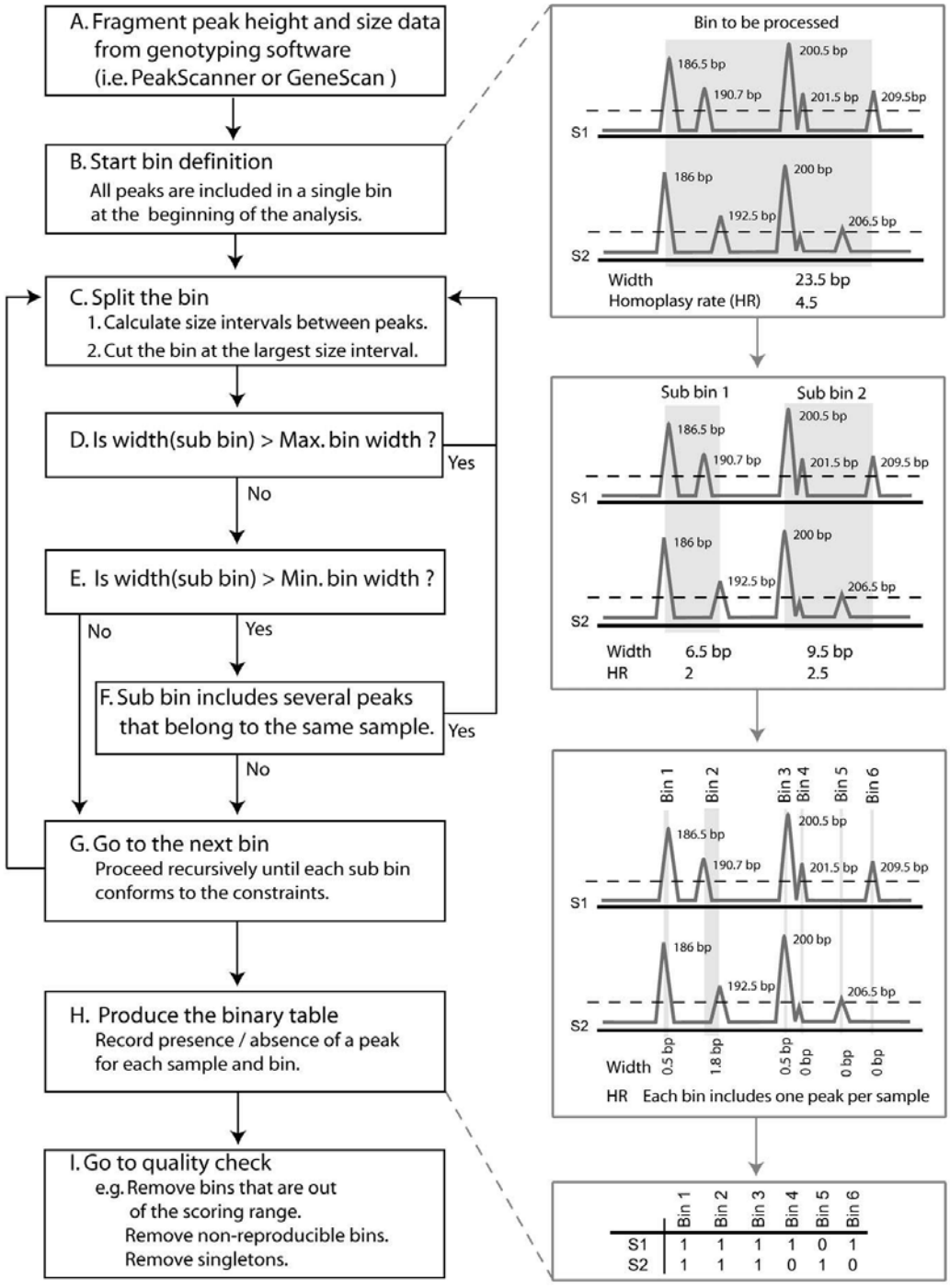

Figure 1

Flowchart illustrating the bin definition algorithm of RawGeno. On the right side: an example showing two samples (S1 and S2) where a total of nine AFLP peaks were preliminarily detected and sized. The bin width (i.e. the difference in size between the longest and the shortest amplicons included in the considered bin) and the technical homoplasy (i.e. the mean number of peaks belonging to the same sample that are included in a same bin) are indicated for each bin.

information regarding rates of reproducibility of alleles and individuals. The resulting scored matrix (referred to as "manual" below) was coded as binary states with lines and columns recording presence or absence of amplicons in samples and bins respectively. Finally, ten datasets were produced by tuning the scoring parameters of RawGeno and GeneMapper. We tuned the "minimum bin width" parameter of RawGeno and the "bin width" parameter of GeneMapper in order to force the algorithms to assign amplicons displaying close but differing sizes to the same bin. As a consequence, the technical homoplasy of the datasets increased proportionally while increasing the bin width.

Automated scoring procedures

The parameters of RawGeno were determined as follows (refer to Figure 1. for a scheme of the scoring algorithm and its parameters): the "maximum bin width" was left unconstrained and set as being the maximum observed amplicon size in the dataset (default value from the original algorithm). As a consequence, RawGeno avoided systematically the oversplitting bias (see above). In contrast, the "minimum bin width" was constrained during the analysis and set to 0.2 bp, 1 bp, 2 bp, 5 bp and 10 bp. This strategy produced datasets with an increasing amount of technical homoplasy since RawGeno was forced to use wider bins. It bears repeating that this setting strategy was used to intentionally increase the technical homoplasy in the produced datasets and not to produce optimally scored datasets (refer to the Additional file 1 for recommendations to conduct a proper scoring with RawGeno). The parameters in GeneMapper were determined using the standard detection settings and a polynomial degree of three was used for peak recognition in the electropherograms. The scoring step was achieved by tuning the "bin width" parameter with values ranging from 0.2 bp to 10 bp (identically to RawGeno settings). The precise effect of the "bin width" parameter on the GeneMapper scoring algorithm could however not be predicted *a priori* and was deduced from the results we obtained. Replicated samples (i.e. ~20% of the sampling) were included in all ten datasets, which allowed the calculation of error rates (see below) and subsequently, the removal of non-reproducible bins (as done during the manual scoring) in the final dataset. Monomorphic bins and singletons were also removed. The whole set of downstream analyses (see below) were carried out on these cleaned datasets. The ten resulting matrices were coded as binary states in the same way as for the manual dataset (see above). For RawGeno, datasets were labelled as follows: RG_0.2, RG_1, RG_2, RG_5 and RG_10. GeneMapper datasets were labelled as follows: GM_0.2, GM_1, GM_2, GM_5 and GM_10.

Descriptive statistics

Several indices were computed. First, the final number of bins (nbin) was recorded for each dataset. Second, the mean homoplasy rate was computed within the RawGeno datasets. The homoplasy rate (HR) was defined as the number of amplicons belonging to the same individual that are assigned within the same bin. This statistic was calculated for each sample/bin and averaged for the whole dataset. The frequency of the "present" allele was computed for each bin and frequencies were plotted against the bin sizes (in bp). Finally, the level of correlation was calculated between each of the ten automatically scored datasets and the manually scored one (i.e. performing Pearson's correlations; hereafter referred to as "R2 Manual"). This was achieved by: I. Calculating Jaccard similarity indices [21] between samples, within each dataset. This calculation is defined as being asymmetric as it only accounts for presences in individual genotypes while absences are not considered. II. Calculating Pearson's correlation between the resulting similarity matrices, using the similarity matrix obtained with the manually scored dataset as reference.

Ordination techniques

The datasets produced from the automated analysis were compared to the manually scored one by using a partial constrained correspondence analysis [21]. We used the "vegan" package [22] implemented in the R CRAN environment and applied the "cca" function (using a "scaling 1" procedure, in order to optimally represent the samples coordinates). This analysis was used to produce residuals containing information that is specific to the automatically scored dataset, according to the following model: $V_a - V_m = R_a$, where V_a is the variance of the automatically scored dataset, V_m is the variance of the manually scored dataset and R_a is referenced to as the residuals that are specific to the automatically scored dataset. We further measured the ability of these residuals to discriminate populations, in order to assess whether the automatically scored datasets effectively contained more information than the manually scored one. This calculation was achieved by applying a Mantel test between the residuals matrix and a contrast matrix comprising the population origin of each sample. This test required the computation of Euclidean distance on the contrast and residual matrices, for which we used the "mantel" function of the "vegan" R CRAN package (1000 permutations).

Error rates and optimality criterion

For each dataset, two estimators of the error rate were computed by considering the information comprised in the replicated samples.

I. The mismatch error rate [11] defined as $E_b = M_{repl}/nbin$, where M_{repl} is the total number of mismatches between a sample and its replicate and nbin is the total number of bin. This statistic was computed for each sample-replicate pair.

II. The Bayesian error rates ε1.0 and ε0.1 that represent, respectively, the probability of mis-scoring the presence or the absence of AFLP fragments. We used MasterBayes (an R CRAN package [23]) and the AFLPScore R CRAN script collection [14] to compute 1000 estimates of these statistics.

These two estimators required the inclusion of replicated samples and did not allow addressing the quality of datasets where non reproducible bins had been removed (e.g. in the manually scored dataset). As a consequence, we propose a new "optimality criterion" based on the information content per bin (I_{bin}). This statistic was calculated for each sample of the dataset and was defined as $I_{bin} = M_{sampling}/nbin$ where $M_{sampling}$ is the average number of mismatches between the considered sample and the other samples of the dataset and nbin is the total number of bins in the dataset. This criterion can be computed at any stage of the scoring process and does not require the inclusion of replicated samples. Here, we applied it after the removal of non-reproducible bins.

Biogeographic structure

We investigated the spatial genetic structure in our datasets at two levels of complexity: first between individuals by computing the Jaccard similarity index [21], second between populations by using an estimator of the FST and by performing Maximum Likelihood clustering (assuming Hardy-Weinberg equilibrium). FSTs were computed with the program AFLP-Surv [24] (allelic frequencies were estimated with a Bayesian method using a non-uniform prior distribution and assuming Hardy-Weinberg equilibrium). Jaccard similarity indices (see above) and FST values obtained with the various datasets were compared using scatterplots and linear regressions. Maximum Likelihood clustering was performed using PSMix [16], a package implemented under the R CRAN environment. It uses Maximum Likelihood methods in order to assign individuals into a predefined number of groups with an associated probability. The algorithm assumes Hardy-Weinberg equilibrium within groups and linkage equilibrium between loci. The datasets were coded as follows: each individual genotype was duplicated, in order to simulate a diploid co-dominant dataset. The absences (0) of the original genotype were coded as absences in the duplicated genotype (i.e., "0" is coded as "0-0") whereas presences (1) occurring in the original genotype were coded as missing data in the duplicated genome (i.e., "1" is coded as "1-?"). This coding scheme is adapted from Bonin et al. [19]. The default settings of PSMix were applied (except for itMax [i.e. the maximum number of iterations] that was set to 100000). The number of investigated groups (K) ranged from two to nine, with ten replicated runs per K. For each K value, only the run showing the highest likelihood value was selected for further analysis. The resulting assignment probabilities were compared using pie-charts mapped on geographical maps.

Diversity indices

Three indices, revealing the level of diversity in each population, were computed: I. the estimated Heterozygosity (Hj), by using AFLP-Surv [24] (we used the same parameters as above). II. The percentage of polymorphic loci (PLP), using AFLPdat, an R CRAN script collection [25]. III. The presence/absence rarity index (i.e. "rarity 2" according to the IntraBioDiv Consortium's [18] explanations at http://www.intrabiodiv.eu/IMG/doc/Diversity_Uniqueness_EN_v03.doc [verified on December 30, 2008]). The values estimated with the various datasets were compared using scatterplots and linear regressions.

Results and discussion

Applying the scoring

The manual scoring required approximately 100 hours of "human work" and produced 102 reproducible bins. Automated scoring procedures, in contrast, produced the scorings in about 1 hour (depending on the kind of dataset and the computer power). Moreover, the automatic scoring procedures preserved a large number of bins as we obtained 502, 456, 316, 177 and 116 bins for the RawGeno datasets (RG_0.2 to RG_10) and 4126, 1338, 742, 340 and 183 bins for the GeneMapper datasets (GM_0.2 to GM_10). As expected, the use of larger bin widths decreased the number of bins and introduced technical homoplasy. Interestingly, technical homoplasy occurred more rapidly, with the increase of the bin width, in small fragment sizes (i.e. sizes < 200 bp, Figure 2A) than in larger fragments. This result is explained as follows. Amplicon sizes are generally asymmetrically distributed because short amplicons are often over-represented compared to larger ones (and especially in genomes with a low GC content when *EcoR* I and *Mse* I are used as restriction enzymes) [13,26]. As a consequence, these small amplicons were shown to be especially prone to reflect size homoplasy [13] and in addition, our results showed that they were also likely to accumulate more technical homoplasy than the rest of the AFLP profile.

Using manual scoring as a reference point

We first investigated the quality of datasets by referring to the manually scored dataset. Technical homoplasy, generated by forcing the scoring algorithm to create large bins (i.e. RG_5, RG_10, GM_2, GM_5 and GM_10), clearly impacted the dataset quality, as shown by a decreasing correlation to the manual dataset with increasing bin width (see Figure 2B). This result matched our expectations and outlined the effects of technical homoplasy. Interestingly, we observed an optimal bin width effect in the GeneMapper datasets (Figure 2B), where datasets

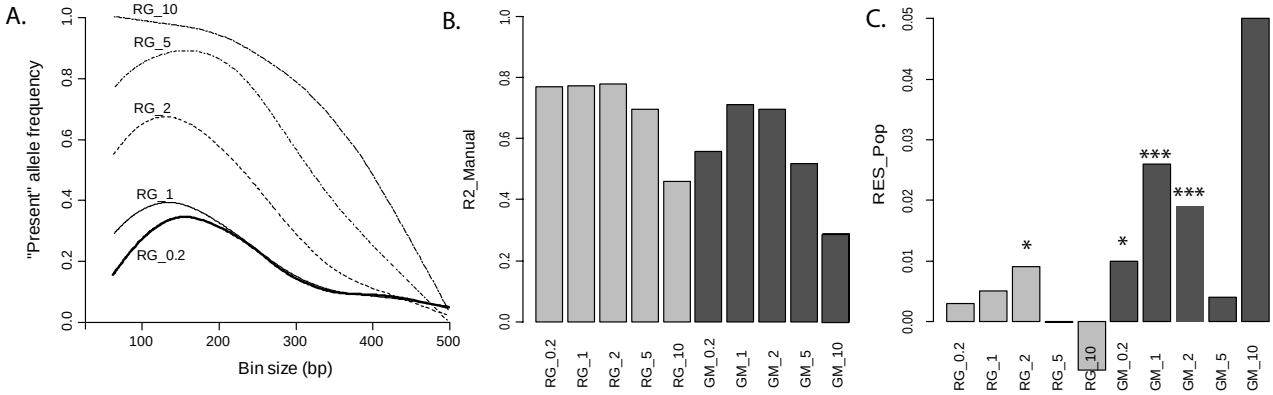

Figure 2

General statistics. In light grey, RawGeno datasets labelled with "RG" as prefix; in dark grey, GeneMapper datasets labelled with "GM" as prefix. A. Comparison of the frequency of the "present" allele (i.e. the occurrence of an amplicon within a bin) according to the bin size, in datasets scored with an increasing bin width using RawGeno. B. Pearson's correlation calculated between the automatically scored dataset and the manually scored one (R2_Manual). C. Partial constrained analysis results. This data mining technique allows the removal of the variation explained by the manually scored dataset in the automatically scored one. As a result, the residuals contain information that is specific to the automatically scored dataset. RES_Pop is a measure of the discrimination power of these residuals in identifying populations. We computed and tested Pearson's correlations between the residuals and a contrast matrix indicating the population origin of samples (Mantel test, 1000 permutations). Significance levels: * ($p < 0.05$), ** ($p < 0.01$) and *** ($p < 0.001$).

scored with small bin widths showed less correlation to the manually scored one than the datasets scored with a greater bin width (i.e. GM_0.2 and GM_1 showed correlations of 0.558 and 0.713, respectively, with the manually scored dataset). Conversely, in RawGeno this phenomenon was avoided since correlation values remained relatively constant between datasets that were scored with small to medium bin widths (i.e. RG_0.2, RG_1 and RG_2 showed correlations of 0.771, 0.774 and 0.779, respectively, with the manually scored dataset). Divergences between RawGeno and GeneMapper can probably be explained by differences in algorithm settings. While RawGeno allowed the user to set both the "minimum" and the "maximum bin width", we suspect that the "bin width" parameter of GeneMapper might have effects similar to those of the "maximum bin width" parameter of RawGeno. Decreasing the "maximum bin width" parameter can exaggerate the splitting of bins and, as a result, appropriately sized bins might be split into smaller and erroneous sub-bins. This bias is expected to produce datasets with a high number of low-quality bins ("oversplitting of bins"). Such a hypothesis is in accordance with the very high number of bins produced in GM_0.2 and GM_1 for instance. Holland et al. [15] showed that, by using GeneMapper, choosing a bin width below 0.4 bp was misleading since it resulted in oversplitting. We could however not obtain an absolute confirmation of this hypothesis since GeneMapper algorithms are strongly black-boxed.

Divergences related to varying bin width parameters were also detected by the partial constrained correspondence analysis (Figure 2C). This procedure inspected the residuals of the automatically scored datasets (after having removed the variation explained by the manually scored dataset). It appeared that the residuals contained relevant information in several datasets (e.g. the residuals of RG_2, GM_0.2, GM_1 and GM_2 significantly discriminated populations). This was particularly true for GeneMapper datasets where the scoring algorithm outlined several additional biogeographic patterns (see Additional file 2). These patterns, however, were seldom interpretable since they segregated single populations from the rest of the samples and were identified neither by the manual scoring nor by RawGeno (see below). We could however consider that the very high number of bins associated with these GM datasets (GM_0.2, GM_1 and GM_2) might have included such additional private alleles. Finally, both oversplitting and technical homoplasy decreased the information content of residuals. This result was reinforced by an optimum bin width effect where only RG_2, from the RawGeno datasets and GM_1 and GM_2 from the GeneMapper datasets showed the highest residual information content, while technically biased datasets presented lower values.

Error rates and optimality criteria

The previous statistics used the manually scored dataset as a reference. This strategy, however, might be misleading

for several reasons. First, this reference point is not available when analysing a new dataset and second, manual scoring may introduce subjective biases (for instance resulting from different filtering strategies) in the dataset. It is therefore clear that the evaluation of dataset quality requires a much more absolute criterion, such as error rates. However, the properties of the different estimators must be evaluated taking into account technical homoplasy and bin oversplitting. For instance, the mismatch error rate (E_b), defined by the number of mismatches between a sample and its replicate, divided by the number of bins ($E_b = M_{repl}/nbin$), failed to unambiguously detect the oversplitting, while technical homoplasy was slightly detected (Figure 3A). On one hand, oversplitting could not be detected by an increase of the mismatch error rate since it dramatically increased the number of bins. Technical homoplasy, on the other hand, was hardly detected because it artificially decreased the number of mismatches between a sample and its replicate, leading to an underestimation of E_b. Similar results were described by Holland et al. [15] where the E_b rate (i.e. referred to as the Euclidean Error rate by the authors) was shown to be unable to discriminate datasets scored with varying bin sizes. As a result, we advise that the mismatch error rate should not be used to optimize the bin definition and the scoring of AFLPs. This criterion provided, however, valuable information when the number of bins remained similar among datasets. This situation was encountered, for instance, during further quality checking steps, such as fluorescence or bin reproducibility filtering. Interestingly, the two Bayesian error rates (Figure 3B and 3C) detected both the oversplitting (with ε1.0) and the technical homoplasy (with ε0.1). These indices thus provided a valuable solution for optimizing the whole range of scoring parameters, including downstream filtering procedures as shown by Whitlock et al. [14]. These statistics, however, required a large computational time and a quicker alternative may be desirable. We therefore propose the use of a new statistic, i.e. the information content per bin (Figure 3D), as a quality criterion to optimize the first steps of AFLP scoring (see above for explanations regarding this new criterion). In our study, this statistic detected both the oversplitting and technical homoplasy. Oversplitting increased the number of bins faster than the number of mismatches. In contrast, the technical homoplasy decreased the accuracy of the dataset at a faster rate than the number of bins decreased. Maximizing the I_{bin} represents an interesting trade-off between the accuracy of the dataset and the number of bins that are used to record the AFLP information. We propose applying this criterion to optimize the bin definition, while the other quality criteria (e.g. Whitlock et al. [14]) can be used during downstream scoring steps. According to the I_{bin} criterion, we assumed that RG_2 and GM_2 datasets had a reasonably good quality. This result was confirmed for RG_2 by both the correlation to the manually scored dataset and the partial constrained correspondence analysis (see Figure 2), while according to these same statistics, GM_1 might perform better than GM_2. It bears repeating, however, that datasets scored with bin widths up to 2 bp probably included some level of technical homoplasy. The presence of stutter-bands (i.e. PCR artefacts) might contribute to this. The effect of this phenomenon on scoring could either produce rare alleles that led to the definition of non-informative bins or cause erroneous results when the range of the stutter bands extended into another bin (e.g. oversplitting bins or mis-scoring of absences or presences). Moreover, stutter-bands generally had a decreased fluorescence and their various peaks might not be reproducible. Therefore, scoring parameters that considered a cluster of stutter-bands as a single allele probably handled the situation in a more appropriate way.

Analysis sensitivity

Biogeographic structure analyses were moderately affected by the scoring system (Figure 4). For instance, both inter-individual (Jaccard similarity index) and inter-population (FST) distances provided comparable results from one scoring method to the other (although RG datasets matched the manually scored ones better than did GM datasets). Additionally, these measures were also moderately affected by the scoring parameters in cases when technical homoplasy and oversplitting were avoided (for instance, convergent results were obtained with the manual scoring, RG_0.2, RG_1 and RG_2 or GM_1 and GM_2 respectively [data not shown]).

The PSMix clustering analyses provided similar patterns (Figure 5). The spatial genetic structure of *C. uniflorum* seemed to be composed of five main phylogeographic regions (the second-derivative of the Maximum Likelihood values showed a small decrease when investigating more than five groups [data not shown]; this result was also accompanied by a decrease in the clustering accuracy). These regions were identified by the clustering analysis in the following order (see Figure 5 and Additional file 2): Meridional (K = 2), Oriental (K = 3), Cryptic (K = 4), and finally Occidental and Central regions (K = 5). Interestingly, the automatically scored datasets showed much higher assignment probabilities to clusters than did the manual dataset (see Additional file 2). RawGeno and manual scoring provided congruent genetic structures while GeneMapper identified specific patterns that were difficult to interpret (as mentioned above). These patterns also disrupted the clustering analyses and similar values of K may not provide converging results when comparing the three scoring methods. The main phylogeographic structures were observed when inspecting the entire set of clustering runs (Additional file 2, with K ranging from two to nine), although results were slightly influenced by the

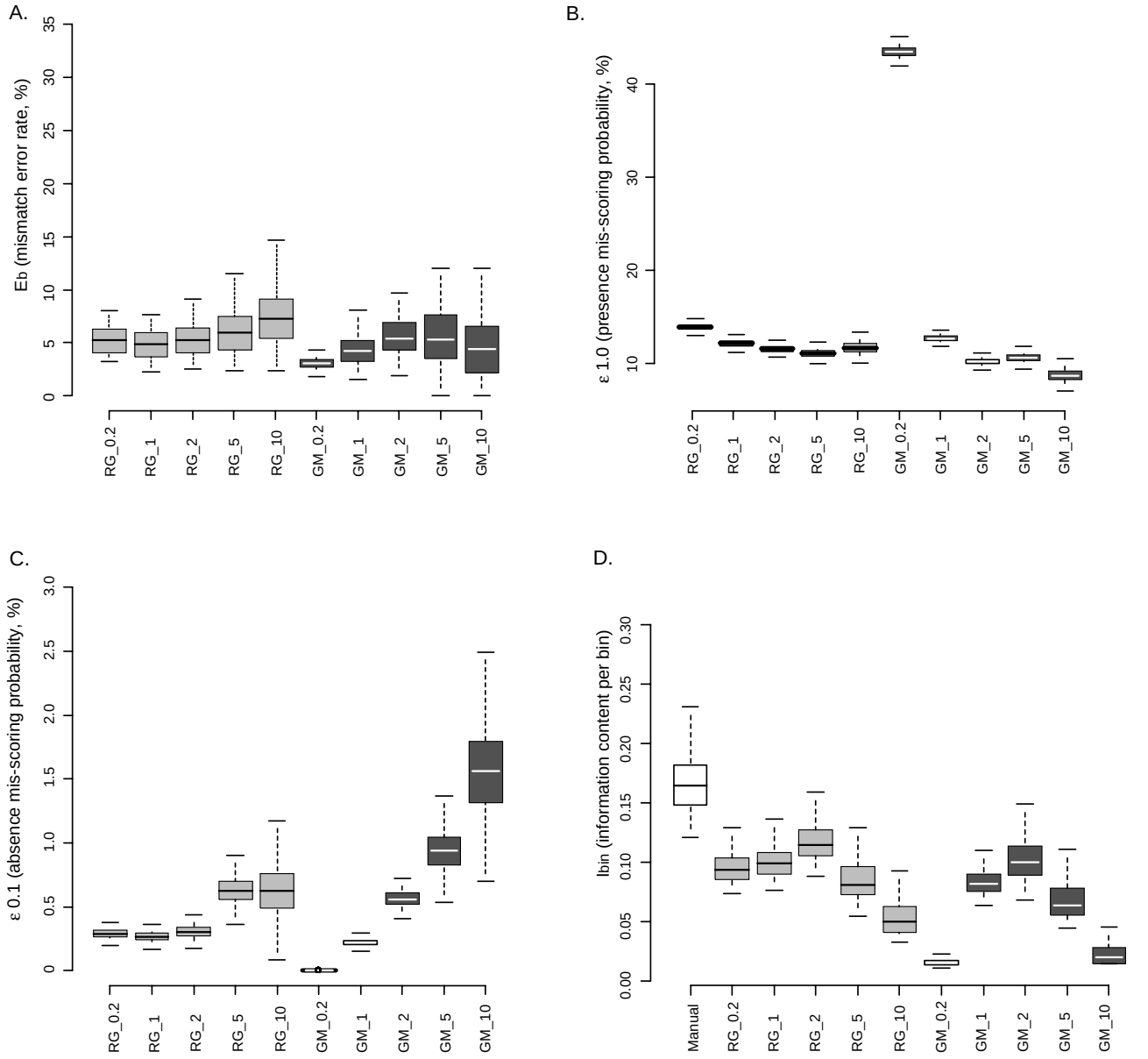

Figure 3

Error rates and information content. In light grey, RawGeno datasets labelled with "RG" as prefix; in dark grey, GeneMapper datasets, labelled with "GM" as prefix. Only the quartiles are displayed on the box-plots. A. Mismatch error rate (E_b). B. and C. Bayesian error rates representing the probability (%) of mis-scoring allele presence (ε1.0) or absence (ε0.1). D.Information content per bin (I_{bin}), calculated as the mean inter-individual distance divided by the number of bins in the dataset. This statistic could also be computed for the manually scored dataset since it does not require replicated samples.

scoring algorithm. The scoring parameters were also important. The consequences of technical homoplasy ranged from a loss of discrimination power (Figure 5 and Additional file 2) for the datasets with moderate homoplasy (e.g. RG_5), to erroneous results for the most biased ones (RG_10 which was largely non-informative). In this situation, coherent results could be observed only up to four groups (i.e. K = 4, Additional file 2) where the most evident phylogeographic regions were detected (Meridional and Occidental clusters) while detailed groupings (such as the Cryptic, Oriental and Central clusters) were missing or identified at higher values of K (see Additional file 2). Interestingly, and despite the relatively high number of bins (e.g. 177 bins for RG_5 vs. 102 for the manually scored dataset), accuracy of clustering was not observed beyond five groups in datasets displaying mod-

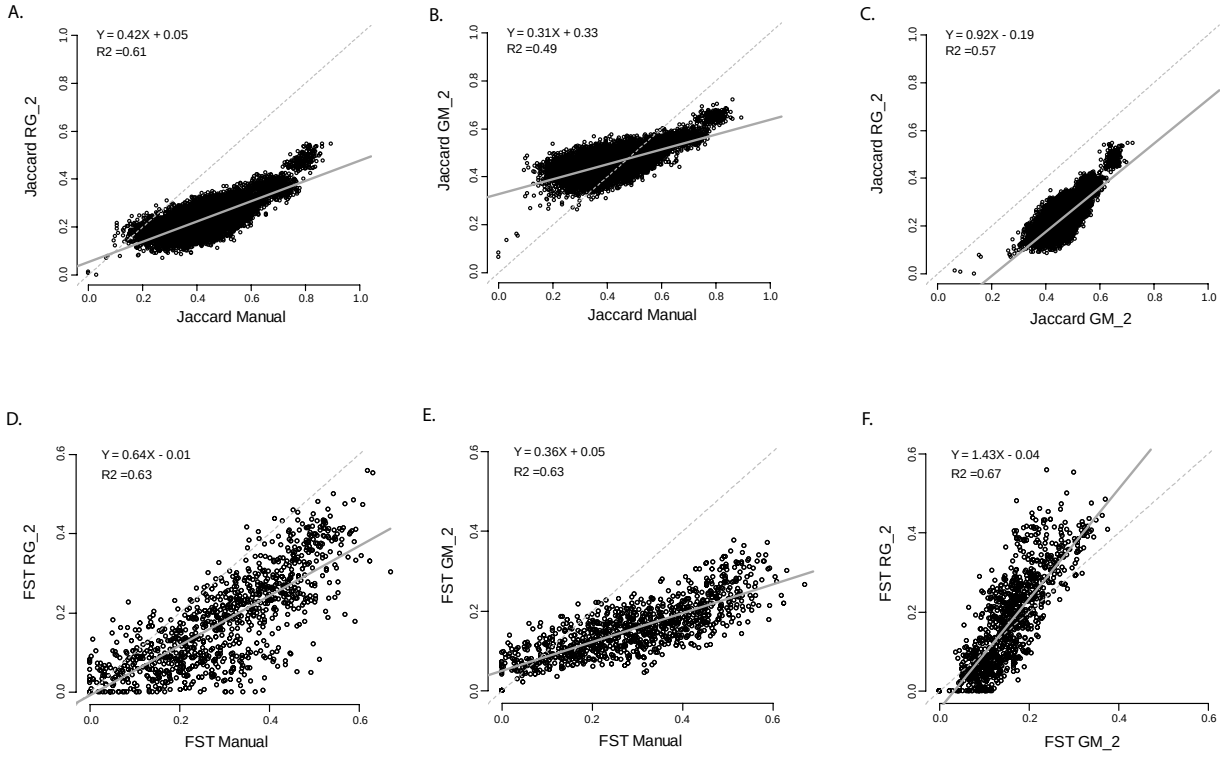

Figure 4

Comparison of genetic structure estimators. The statistics evaluate the consistency of two structure analyses, computed with the manual dataset and two automatically scored datasets assumed to be accurate according to the I_{bin} criterion (i.e. "RG_2" and "GM_2"). The estimations obtained with the three datasets are compared with pairwise scatterplots (RG versus Manual, GM versus Manual and RG versus GM) and linear regressions (displayed as a solid line). The R^2, the slope and the intercept of the regression are indicated. A. to C. Inter-individual similarities (Jaccard similarity index). D. to F. Inter-population distances (FST).

erate homoplasy, indicating that the additional bins of these datasets probably contained noisy information (Additional file 2). This result also agreed with the I_{bin} values that showed a decrease in the information content of these datasets.

In contrast to genetic structure patterns, population diversity indices were extremely sensitive to the scoring features since the estimated Heterozygosity, the percentage of polymorphic loci and the rarity index showed important discrepancies among the datasets (Figure 6 and Additional file 3). Interestingly, the percentage of polymorphic loci and the estimated Heterozygosity were the most robust measures that we tested (Figure 6A to 6F). Detection of rare loci was more successful when the datasets were finely scored, but this statistic was particularly sensitive to the technical homoplasy (Figure 6G to 6H and Additional file 3C). We explain these results as follows. First, the sampling scheme of our model species might not be robust enough (with an average of four individuals per population) and in this context, estimates of genetic diversity might be strongly influenced by technical biases. Second, scoring biases such as the oversplitting of bins or the technical homoplasy affected the definition of bins and the recording of presence/absence of alleles. As a consequence, statistics that relied directly on the polymorphic status of presence/absence of alleles, (such as diversity measures at the within population level) were likely to be particularly sensitive to the dataset noise. As a result, scoring biases may have reinforced the problems caused by size homoplasy in estimating the genetic diversity [26]. This point raised the question of how to obtain reliable diversity estimates by using AFLP datasets and the use of other indices, such as the proportion of different genotypes in a population (not tested here because of the small sample sizes), might provide more robust results. In contrast, synthetic statistics such as distance-based methods or clustering analyses probably buffered the technical noise with the rich signal of AFLP and finally provided coherent results. Similar observations were made in other studies [15,19,26] where distance-based analyses proved to be robust regardless of the choice of the scoring algo-

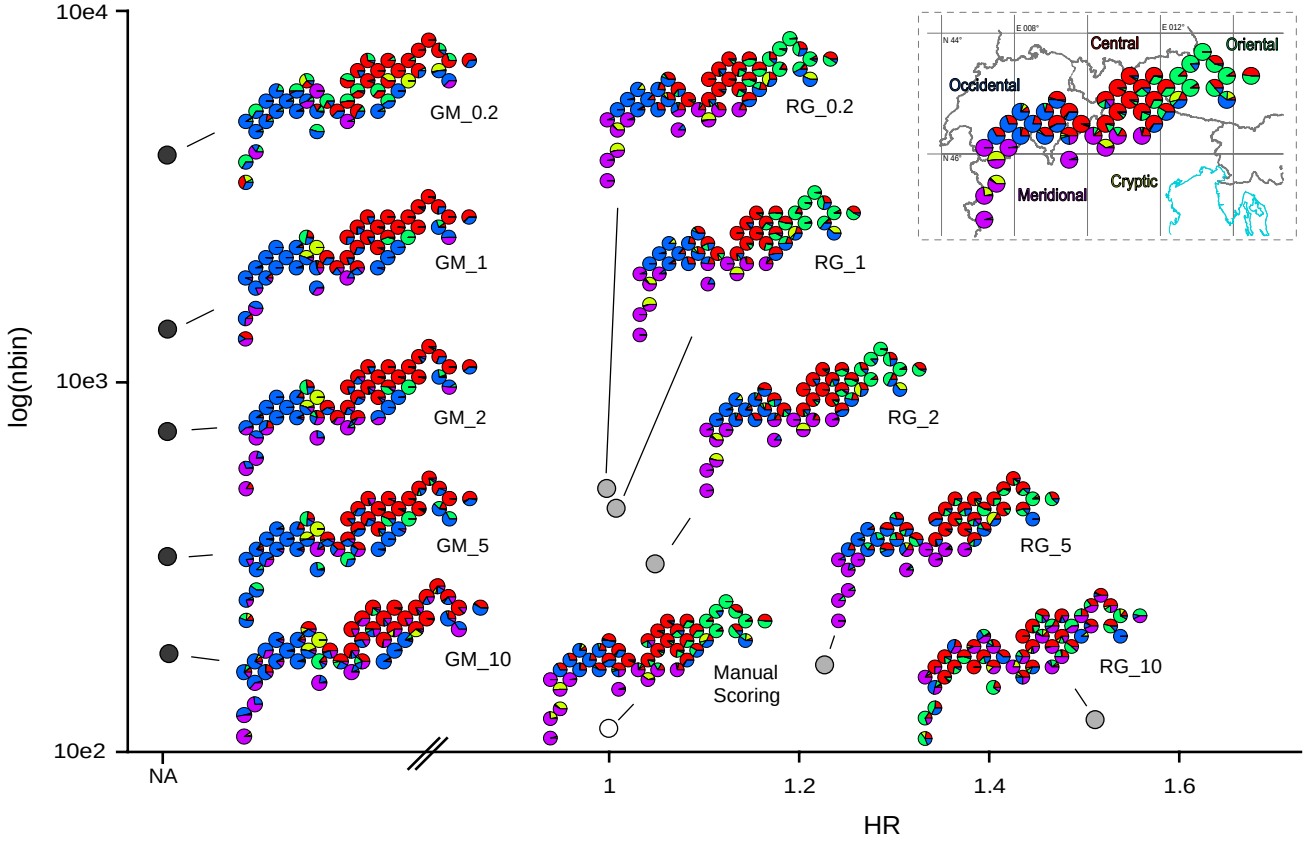

Figure 5

Effect of the scoring parameters on the PSMix clustering analysis. Scatterplot of the number of bins (nbin, log-scaled y-axis), according to the mean homoplasy rate (HR, x-axis). The mean homoplasy rate is defined as the average number of peaks belonging to the same individual that are affiliated within the same bin. The eleven datasets are displayed as dots (in dark grey, GeneMapper datasets labelled with "GM" as prefix; in light grey, RawGeno datasets labelled with "RG" as prefix and in white, manual dataset) and as maps on the scatterplot. The maps represent *C. uniflorum* populations with pie-charts representing assignment probabilities obtained with PSMix (K = 5 groups). On the upper-right corner: geographical location of the sampling and names of the identified phylogenetic groups (violet – Meridional cluster, blue – Occidental cluster, red – Central cluster, green – Oriental cluster and yellow – Cryptic cluster). The mean homoplasy rate could not be calculated for the GeneMapper datasets.

rithm and its settings. We also report convergent observations for gene flow detection (N. Arrigo and S. Lappe, unpublished data) where F1 hybrids and introgressed individuals were detected accurately, regardless of moderate levels of technical homoplasy. Finally, several steps that might help to eliminate unsatisfactory bins or peaks (such as filtering bins according to their average fluorescence [14]) were not applied in the present study. Better results may be obtained by optimizing this specific aspect of the analysis. Such filtering methods, as well as new tuning strategies, will be progressively implemented in RawGeno. In any case, however, we stress that users must carefully choose the parameters of the scoring algorithms to reflect the aim of the analysis.

Conclusion

Scoring a dataset manually is a time consuming process that is complicated when investigating a large number of samples. Using an automated system can significantly increase the reproducibility of the dataset as all the electropherograms are scored uniformly and as genotyping errors are limited to technical factors (e.g. PCR or migration variations). Our study showed that, with high quality AFLP and GeneScan raw data, automated procedures can be particularly efficient in producing ready-to-use datasets, at least in the context of population genetic or phylogeographic studies (i.e., RawGeno was not tested in a genomics framework [e.g., for gene mapping] and further investigations are needed before validating its extension

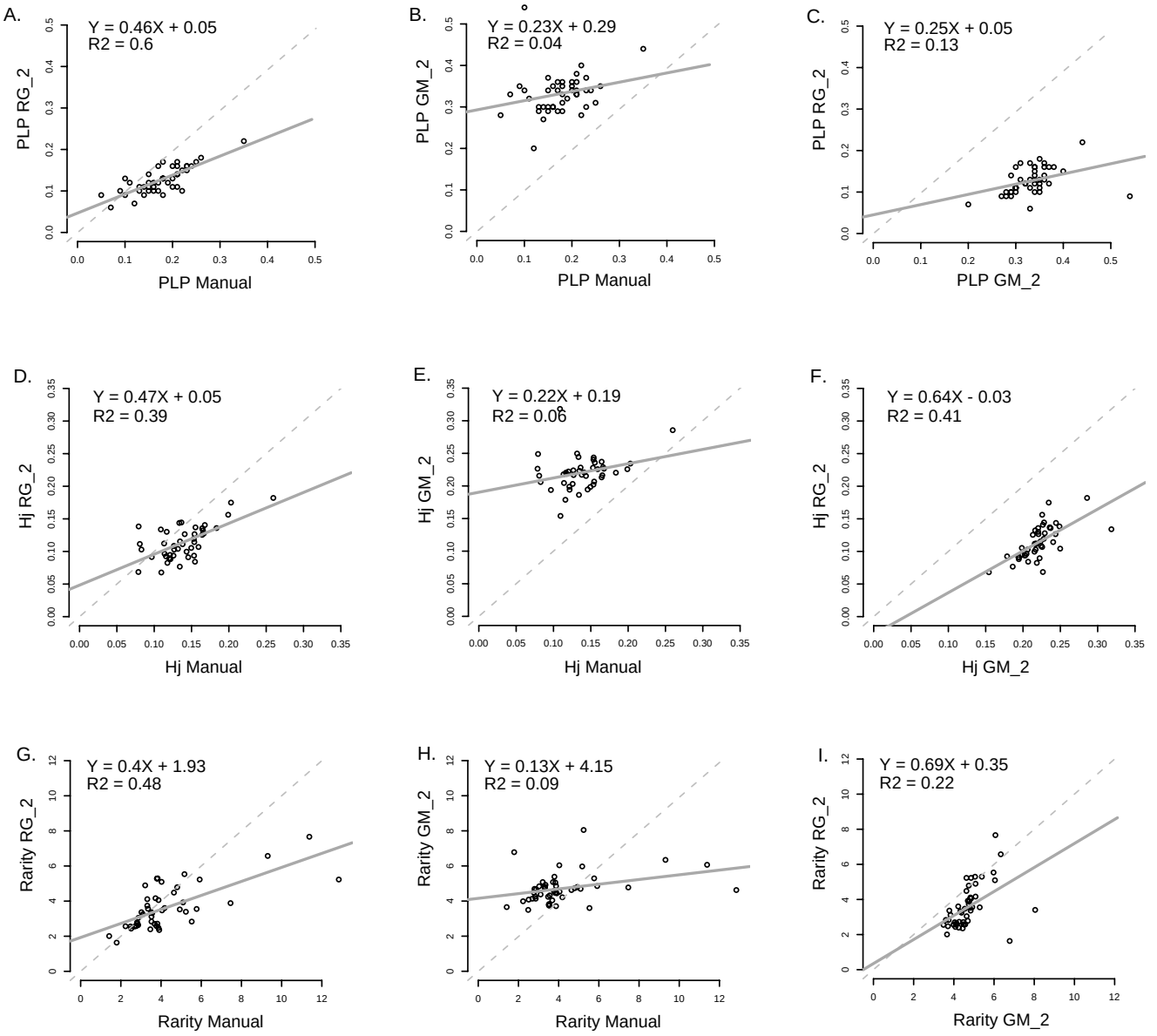

Figure 6

Genetic diversity statistics. The statistics evaluate the consistency of three diversity estimators calculated with the manual dataset and two automatically scored datasets assumed to be accurate according to the I_{bin} criterion (i.e. "RG_2" and "GM_2"). The estimations obtained with the three datasets are compared with scatterplots (RG versus Manual, GM versus Manual and RG versus GM) and linear regressions (displayed as a solid line). The R^2, the slope and the intercept of the regression are indicated. A. to C. Percentage of polymorphic loci (PLP). D. to F. Estimated Heterozygosity (Hj). G. to I. Rarity index (Rarity).

to this field). However, the automated scoring of AFLPs is a multiple-step process and a trade-off based on several quality criteria may be desirable since it might provide more relevant information than a single statistic. Optimizing the parameters used in the scoring algorithm therefore represents one of the most important steps of the whole analysis. Using RawGeno, we were able to evaluate the impact of technical homoplasy and bin oversplitting on optimality criteria and genetic structure patterns by intentionally biasing our starting datasets. Interestingly, our results demonstrated that a high number of redundant and informative bins might overcome techni-

cal homoplasy due to scoring errors, at least when investigating biogeographic structures. While allowing for some plasticity during the optimization of the scoring procedure, this result also reinforces the use of the AFLP methodology for its ability to produce highly informative datasets. By contrast, the estimation of genetic diversity may be considered with caution since scoring biases are likely to reinforce problems caused by size homoplasy.

Finally, RawGeno provided results at least as accurate as those obtained by scoring the dataset manually (even when considering bin widths as wide as 2 bp, representing an error range much higher than the technical error rate of the genotyping machines) or by using a commercial software such as GeneMapper. To our knowledge, RawGeno is the only freely available program proposing a fully automated scoring solution, from electropherogram files to user-defined working binary matrices. Benefiting from the open source R platform, RawGeno can be potentially enhanced and used by any user.

Authors' contributions

NA1 carried out the main design of the study, organized the package, programmed and debugged the R code, and drafted the manuscript. JWT, DE and TG participated in the design of the study and programmed parts of the code. NA2 participated in the main design and coordination of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Detailed technical features of RawGeno. *This document provides a description of scoring and bin filtering solutions proposed by RawGeno. In addition, it includes recommendations to achieve a proper scoring.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-33-S1.doc]

Additional file 2

Spatial genetic structures for all datasets. *The graphics are displayed on an array where the lines represent the different datasets (Manual – manual scoring, RG – RawGeno datasets and GM – GeneMapper datasets) and the columns display the different analyses. The three first columns contain the scatterplots of three genetic diversity indices (Rarity – rarity index, Hj – estimated Heterozygosity and PLP – percentage of polymorphic loci). The values obtained by using the automatically scored datasets (displayed on the y-axis) are compared to those obtained with the manually scored dataset (displayed on the x-axis). The red line represents a linear regression between values obtained by both datasets (the Pearson's correlation indices of these regressions are displayed in the Figure 6). The next columns contain individual clustering results (with the number of* a priori *groups ranging between K = 2 and K = 9).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-33-S2.pdf]

Additional file 3

Effect of the scoring parameters on diversity estimators. *Scatterplot of the number of bins (y-axis, log-scaled), according to the mean homoplasy rate (x-axis). The mean homoplasy rate (HR) is defined as the average number of peaks belonging to the same individual that are affiliated within the same bin. The eleven datasets are displayed as dots (in dark grey, GeneMapper datasets labelled with "GM" as prefix; in light grey, RawGeno datasets labelled with "RG" as prefix and in white, manual dataset) and as maps on the scatterplot. The maps represent* C. uniflorum *populations with circles. The radius of circles is a function of the measured diversity. A. Percentage of polymorphic loci (PLP). B. Estimated Heterozygosity (Hj). C. Rarity index (Rarity).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-33-S3.pdf]

Acknowledgements

The authors would like to thank the IntraBiodiv Consortium for providing raw AFLP datasets and particularly Conny Thiel-Egenter for producing the GeneMapper outputs. They also warmly thank Sarah Kenyon, Russell Naisbit and three anonymous referees for providing valuable comments on earlier versions of the manuscript. RawGeno includes software developed by SAIC for the National Cancer Institute and by Tommy Gerdes for the Rigshospitalet. Roberto Guadagnuolo, François Felber, Celia Bueno, Sven Buerki, Michael Stipe, Thom Yorke and Anouk Beguin provided helpful ideas and discussions; Fabien Fivaz and Anthonny Lehmann helped to code the Graphical User Interface. This study was funded by the National Centre of Competence in Research (NCCR) Plant Survival and by the Swiss National Science Foundation (project No. 3100A0-116778/1).

References

1. Avise JC: *Molecular markers, natural history, and evolution* Sunderland: Sinauer; 2004.
2. Karp A, Seberg O, Buiatti M: **Molecular techniques in the assessment of botanical diversity.** *Ann Bot* 1996, **78:**143-149.
3. Botstein D, White RL, Skolnick M, Davis RW: **Construction of a genetic-linkage map in man using Restriction Fragment Length Polymorphisms.** *Am J Hum Genet* 1980, **32:**314-331.
4. Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV: **DNA polymorphisms amplified by arbitrary primers are useful as genetic-markers.** *Nucleic Acids Res* 1990, **18:**6531-6535.
5. Tautz D: **Hypervariability of simple sequences as a general source for polymorphic DNA markers.** *Nucleic Acids Res* 1989, **17:**6463-6471.
6. Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M: **AFLP – A new technique for DNA-fingerprinting.** *Nucleic Acids Res* 1995, **23:**4407-4414.
7. Vos P, Kuiper M: **AFLP analysis.** In *DNA markers: Protocols, Applications and Overviews* Edited by: Caetano-Anolles G, Gresshoff PM. New-York: J. Wiley & Sons; 1997:115-131.
8. Pompanon F, Bonin A, Bellemain E, Taberlet P: **Genotyping errors: causes, consequences and solutions.** *Nat Rev Genet* 2005, **6:**847-859.
9. Meudt HM, Clarke AC: **Almost forgotten or latest practice? AFLP applications, analyses and advances.** *Trends Plant Sci* 2007, **12:**106-117.
10. Benham J, Jeung JU, Jasieniuk M, Kanazin V, Blake T: **Genographer: a graphical tool for automated AFLP and microsatellite analysis.** *J Agric Genomics* 1999, **4:**3.
11. Bonin A, Bellemain E, Bronken Edeisen P, Pompanon F, Brochmann C, Taberlet P: **How to track and assess genotyping errors in population genetics studies.** *Mol Ecol* 2004, **13:**3261-3273.

12. DeHaan LR, Antonides R, Belina K, Ehlke NJ: **Peakmatcher: software for semi-automated fluorescence-based AFLP.** *Crop Science* 2002, **42:**1361-1364.
13. Vekemans X, Beauwens T, Lemaire M, Roldan-Ruiz I: **Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size.** *Mol Ecol* 2002, **11:**139-151.
14. Whitlock R, Hipperson H, Mannarelli M, Butlin RK, Burke T: **An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error.** *Mol Ecol Res* 2008, **8:**725-735.
15. Holland BR, Clarke AC, Meudt HM: **Optimizing automated AFLP scoring parameters to improve phylogenetic resolution.** *Syst Biol* 2008, **57:**347-366.
16. Wu BL, Liu NJ, Zhao HY: **PSMIX: an R package for population structure inference via maximum likelihood method.** *BMC Bioinformatics* 2006, **7:**317-325.
17. **The caMassClass 1.0. Library (Tuszynski 2003)** [http://cran.r-project.org/web/packages/]
18. Gugerli F, Englisch T, Niklfeld H, Tribsch A, Mirek Z, Ronikier M, Zimmermann NE, Holderegger R, Taberlet P: **Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation – a project synopsis.** *Perspect Plant Ecol* 2008, **10:**259-281.
19. Bonin A, Ehrich D, Manel S: **Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists.** *Mol Ecol* 2007, **16:**3737-3758.
20. Gilder JR, Ford S, Doom TE, Raymer ML, Krane DE: **Systematic differences in electropherogram peak heights reported by different versions of the GeneScan® Software.** *J Forensic Sci* 2004, **49:**92-95.
21. Legendre P, Legendre L: *Numerical ecology* Amsterdam: Elsevier; 2000.
22. **Vegan: community ecology package (Oksanen et al. 2008)** [http://cc.oulu.fi/~jarioksa/softhelp/vegan.html]
23. Hadfield JD, Richardson DS, Burke T: **Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework.** *Mol Ecol* 2006, **15:**3715-3730.
24. **AFLP-SURV version 1.0. (Vekemans 2002)** [http://www.ulb.ac.be/sciences/ecoevol/aflp-surv.html]
25. Ehrich D: **AFLPdat: a collection of R functions for convenient handling of AFLP data.** *Mol Ecol Notes* 2006, **6:**603-604.
26. Caballero A, Quesada H, Rolan-Alvarez E: **Impact of Amplified Fragment Length Polymorphism size homoplasy on the estimation of population genetic diversity and detection of selective loci.** *Genetics* 2008, **179:**539-544.

Chapter 3

Investigating the mating system of three Mediterranean *Aegilops* L. species

Arrigo N[1], Guadagnuolo R[1], Kaya I[2] and Felber F[1]

[1]Laboratory of Evolutionary Botany, Institute of Biology, University of Neuchâtel, 11 rue Emile-Argand, 2000 Neuchâtel, Switzerland

[2]Yüzüncü Yıl University, Agriculture Faculty, Plant Protection Department, 65080 Van – Turkey

Abstract

Gene transfer between crop plants and their wild relatives necessarily starts with first generation hybrids. Mating system plays a key role since gene flow is directly related to outcrossing. The present study investigates the mating system of three Mediterranean wild relatives of wheat: *Aegilops geniculata* Roth, *Ae. neglecta* Req. ex Bertol. and *Ae. triuncialis* L.. We use population genetic statistics, based on Amplified Fragment Length Polymorphisms and outcrossing experiments to assess the commonly assumed autogamy of *Aegilops* L. species. Population genetics suggests different mating system for the three investigated species. *Aegilops geniculata* appears to be the most autogamous species while *Ae. neglecta* and *Ae. triuncialis* show evidence of facultative allogamy. Similarly, the outcrossing experiment, by comparing fertilities in spikes forced to autopollination or allowed to allopollinate, shows that the pollen origin plays no role in *Ae. geniculata*. In contrast, a slight increase in fertility is observed in *Ae. triuncialis* and *Ae. neglecta* when allopollination is allowed. Our results corroborate that autogamy is not strict and cannot be expected to act as an absolute barrier against gene transfer involving wheat and *Aegilops* species.

Introduction

The genus *Aegilops* L. belongs to the Poaceae family and includes 22 species (van Slageren 1994). With a diploid origin located in the transcaucasian region, the genus includes several allopolyploid species that colonized the Mediterranean basin (both southern and northern coasts) and reached Central Asia. *Aegilops* species commonly occur in disturbed environments such as pastures, tree-cultivations (oaks, pistachio or olive tree yards), roadsides, industrial wastelands or edges of cultivations. Some species have large colonizing abilities and can occasionally become invasive by dominating the local vegetation. As a result, large *Aegilops* populations occur into pastures or cultivated fields (van Slageren 1994). Finally, *Aegilops* populations were reported as adventives in most parts of northwestern and northern Europe (van Slageren 1994). Seven species were introduced in the United States of America where at least three of them are locally dominant (*Ae. geniculata* Roth) or even considered as serious weeds (*Ae. cylindrica* Host and *Ae. triuncialis* L.) (van Slageren 1994; Morrison et al. 2002; Meimberg et al. 2006).

Aegilops species grow in habitats characterized by hot and dry summers and rainy winters. Their life strategy relies on a vegetative growth during winter, a reproductive period achieved prior to summer and a predominantly autogamous mating system. These traits represent a selective advantage in dry or continental habitats (van Slageren 1994). Nevertheless, the pollination is anemophilous and the importance of autogamy varies according to *Aegilops* species. For instance, mating systems range from strictly selfing species where florets remain closed when the pollen is released (a strategy termed "cleistogamy") to almost allogamous species where large amounts of pollen are released out of the spikes. Allogamy is assumed to be an ancestral state in the evolution of *Aegilops* species while autogamy is more frequent in polyploid than in diploid species (Hammer 1980).

The genus *Aegilops* is closely related to wheat (*Triticum* L.) as several *Aegilops* species have been involved in the domestication and the evolution of the modern wheat. Some species are direct ancestors as *Ae. speltoides* Tausch and *Ae. tauschii* Coss. which provided respectively the B and D genomes of bread wheat (Waines and Barnhart 1992). Moreover other species as *Ae. umbellulata* Zhuk., *Ae. comosa* Sm. In Sibth. & Sm., *Ae. longissima* Schweinf. & Muschl., *Ae. ventricosa* Tausch, *Ae. peregrina* (Hack. In J. Fraser) Maire & Weiller, *Ae. neglecta* Req. ex Bertol., *Ae. geniculata* and *Ae. triuncialis* were used in wheat improvement programs (Zaharieva et al. 2001b; Schneider et al. 2008). Their common genetic background with the *Triticum* genus, the ecological niche and the wide geographical range of most of the *Aegilops* species offer multiple chances for hybridization events with both *durum* and bread wheat.

In the current context of modern agriculture, crop to wild gene flow is a major concern, especially because of the cultivation and commercialization of genetically modified crops (GM crops). Wheat is of great commercial interest: with 607 million tons harvested annually, it is currently the most important cultivated crop (FAOSTAT 2007). The recent progresses in biotechnologies allowed the production of experimental GM wheat varieties that expressed an increased yield, agronomical traits of interest (softer endosperm, increase of the glutenin content), tolerances to herbicides (glyphosate, glyphosinate), fungi (*Fusarium* or smut), pests (production of toxins such as snow-drop lectin) or drought. Up to now, if many of these varieties have been evaluated under field tests

(http://www.isb.vt.edu/, accessed the 13th of March 2009), none are commercialized. However, while the possibility of (trans-) gene transfer to some *Aegilops* species by sexual reproduction is now assessed (Schoenenberger et al. 2006; Loureiro et al. 2007), very little is known about the mechanisms of such gene flow. The mating system is of major importance in *Aegilops* species, as it directly affects the magnitude of gene flow by pollen. We investigate here the breeding system of three *Aegilops* species (*Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis)* that are frequent in the occidental Mediterranean area. We infer it on one side from the population genetics of natural provenances and on the other side with an experiment which compares the fertility of autogamy vs autogamy plus allogamy mating. We discuss the characteristics of the breeding system of the wild plants in the context of crop-to-wild gene flow.

Material and Methods

Plant material

The sampling included 40 populations (19 – *Ae. geniculata*, 8 – *Ae. neglecta* and 13 – *Ae. triuncialis*) that were collected as seeds during field surveys conducted in Spain, France, Italy, Croatia and Turkey. Seeds were grown at the Botanical Garden of the city and University of Neuchâtel and provided the plant material required by our experiments (Table 1).

Table 1 Plant material. The country, the geographical coordinates (N and E, decimal degrees) and the number of individuals included in the AFLP and outcrossing experiments are indicated for each population.

Population	Country	N	E	AFLP	Outcr.
Ae. geniculata					
C0723A	Croatia	43.59	16.73	14	24
C0723C	Croatia	43.47	16.69	12	
E0609B	Spain	40.01	-2.96	20	
E0610B	Spain	39.98	-2.09	11	
E0610C	Spain	39.85	-2.02	20	24
E0610E	Spain	39.22	-1.94	9	
E0611B	Spain	38.15	-3.02	11	
E0615A	Spain	38.26	-5.70	11	
E0616D	Spain	39.97	-4.71	12	
E0617A	Spain	40.75	-4.66	14	
E0618A	Spain	41.22	-3.88	20	
F070406D	France	43.71	3.68	16	
F070606D	France	43.68	3.91	12	
F1207A	France	44.05	6.06	9	24
F1307A	France	43.39	5.13	9	
I0623A	Italy	43.54	11.42	10	24
I0623D	Italy	43.18	11.37	5	
I0624A	Italy	43.15	11.18	17	
I0624C	Italy	43.07	10.91	7	
				239	**96**
Ae. neglecta					
C0723C	Croatia	43.47	16.69	7	
E0613A	Spain	36.77	-5.27	11	
E0615C	Spain	39.17	-6.25	14	24
E0616A	Spain	39.31	-5.72	10	
E0616B	Spain	39.46	-5.17	7	
I0623B	Italy	43.59	11.57	7	
I0624B	Italy	43.18	11.14	15	24
Orb1	Italy	42.47	11.34	12	24
				83	**72**
Ae. triuncialis					
C0723C	Croatia	43.47	16.69	9	
C0724A	Croatia	43.51	16.09	10	
E0610B	Spain	39.98	-2.09	5	
E0610E	Spain	39.22	-1.94	8	
E0611B	Spain	38.15	-3.02	7	
E0612A	Spain	37.54	-4.34	7	24
E0612B	Spain	37.46	-4.32	9	24
E0613A	Spain	36.77	-5.27	6	
E0615A	Spain	38.26	-5.70	8	
E0615C	Spain	39.17	-6.25	9	
E0618A	Spain	41.22	-3.88	7	
I0626B	Italy	44.69	9.99	10	
Turk6	Turkey	38.43	43.26	10	24
				105	**72**

Population genetics

Molecular analyses

A total of 427 samples were included in the AFLP analysis and represented an average of 11 samples per population. DNA extractions were carried out on leaves previously dried and stored in silica-gel, using a CTAB protocol (Chen and Ronald 1999). DNA concentrations were standardized (10 ng/ l) after a quality and quantity check on 1% agarose gels.

The AFLP protocol (Gugerli et al. 2008) used two selective primer pairs (MCTG-EACT and MCAT-EAGT). Reactions were conducted in a PTC-100 thermocycler (MJ Research). We used 96-wells PCR plates, where samples were randomly distributed. A large proportion of AFLP reactions were replicated in order to check the reproducibility of the analysis (i.e. representing 14% of the whole sampling; replicates were randomly chosen from the original samples pool). PCR products were recorded using the GeneScan technology (EACT and EAGT were fluorescently labelled with FAM) with a capillary sequencer (ABI 3730XL, service provided by Macrogen Inc. Seoul, Korea). The scoring was performed using RawGeno (Arrigo et al. 2009), an automated scoring algorithm that was parameterised as follows: scoring range = 100 – 400 bp, minimum intensity = 100 rfu, minimum bin width = 0 bp, maximum bin width = 2 bp and closely sized bins (5%) were removed.

Diversity analysis

We investigated the intraspecific patterns of genetic diversity using AMOVA. The analysis was computed on each *Aegilops* species separately using Arlequin V3.11 (Excoffier et al. 2007). We excluded a single *Ae. triuncialis* population from these calculations since it was collected out of the main sampling area (i.e. Turk6, collected in Turkey while the rest of *Ae. triuncialis* samples were collected in Europe) and would have potentially biased the interpopulation variation measures.

At the intrapopulation level, we estimated the percentage of polymorphic loci and the expected heterozygosity with AFLPsurv (Vekemans et al. 2002). The analysis was conducted separately with each *Aegilops* species. It assumed Hardy-Weinberg equilibrium (preliminary tests using our dataset indicated that AFLPsurv was robust to deviations from Hardy-Weinberg equilibrium) and used a Bayesian method with non-uniform prior distribution to estimate allele frequencies. In order to provide unbiased estimates due to an unequal sampling effort between populations, the percentage of polymorphic loci and the expected heterozygosity were bootstrapped by resampling 1000 times a constant number of individuals in each population (5 individuals per population).

Outcrossing experiment

Experimental design

In order to compare the effect of pollen origin on the seed production, an outcrossing experiment was planned with a subset of 10 populations (4 – *Ae. geniculata*, 3 – *Ae. neglecta* and 3 – *Ae. triuncialis*, Table 1). The experiment compared the fertility of spikes that were either forced to autopollinate or allowed to allopollinate. An average of 14 experimental blocks were used in each population. Each experimental block included two individuals, with three spikes each. On one hand, two spikes belonging to the same individual were isolated within a paper bag in order to produce autopollinated seeds. Spikes were manipulated prior to flowering. On the other hand, two spikes belonging to the two individuals were joined within another paper bag in order to allow allopollination (autopollination remained possible since spikes were not castrated). Flowering overlaps were guaranteed by pairing spikes that had a similar phenological development. Spikes were collected at the end of the experiment and fertility was calculated as a percentage, Fertility = 100 * nseeds / nflowers where nseeds and nflowers were the number of produced seeds and the number of flowers in the spike.

Statistical tests

We compared the fertilities obtained in both spike categories at the populations and the species levels. The statistical tests were computed either by considering each population separately (intrapopulation level) or by including all the samples from a same species in the comparison (intraspecific level). Our dataset was organised in experimental blocks and the independency of the results between both spike categories could not be directly assumed. We applied the following model using CANOCO (Braak 1990): Fertility = * mating system - * experimental block, where the “mating system” parameter indicated the spike category (autopollinated or allo-/autopollinated spike) while the “experimental block” was declared as a covariable in order to remove its effects from the model. We then computed a Monte Carlo permutation test in order to assess the significance of the “mating system” variable. The permutations were performed within the “experimental blocks” and between the different “mating systems” (999 permutations).

Results

Genetic diversity

The AFLP analysis produced a total of 814 bands (442 and 372 bands for EACT-MCTG and EAGT-MCAT respectively) where 572, 456 and 562 bands were observed respectively in *Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis*. An average of 170 bands were scored in each sample and the measured error rate was 5.8% of non-matching bands per sample-replicate pair.

The intraspecific patterns of genetic diversity, as revealed by the AMOVA analysis (Table 2), showed that most of the genetic diversity in *Aegilops* species occurred within the populations with respectively 66%, 86% and 80% for *Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis*. *Aegilops geniculata* showed the highest differentiation between populations with 34% of the genetic diversity that occurred between the populations.

Table 2 AMOVA analysis based on AFLPs. Computations were run independently for each *Aegilops* species. Significance tests (1023 permutations), d.f. degrees of freedom, *** = P<0.0001.

	d.f.	Sum of squares	Variance components	Percentage of variation	FST
Ae. geniculata					
among populations	18	4432.71	17.08	34.21	0.34***
within populations	220	7223.16	32.83	65.79	
Ae. neglecta					
among populations	7	670.19	5.81	13.81	0.14***
within populations	75	2717.31	36.23	86.19	
Ae. triuncialis					
among populations	11	1520.04	11.62	20.00	0.2***
within populations	83	3858.5	46.89	80.00	

At the intrapopulation level, *Ae. geniculata* appeared to be the less diverse and showed low values for the percentage of polymorphic loci and the expected heterozygosities (Figure 1). In contrast, *Ae. neglecta* and *Ae. triuncialis* populations were globally more diversified. The highest diversities were observed for some populations of *Ae. neglecta* (E0616B, I0623B and C0723C) with 0.3 as expected heterozygosities. These high values were due to introgression as many private AFLP bands that had a wheat origin were present in these populations (Arrigo et al. in prep.). The expected heterozygosity of other populations ranged between 0.18 and 0.25. Convergent results were obtained with the proportion of polymorphic loci.

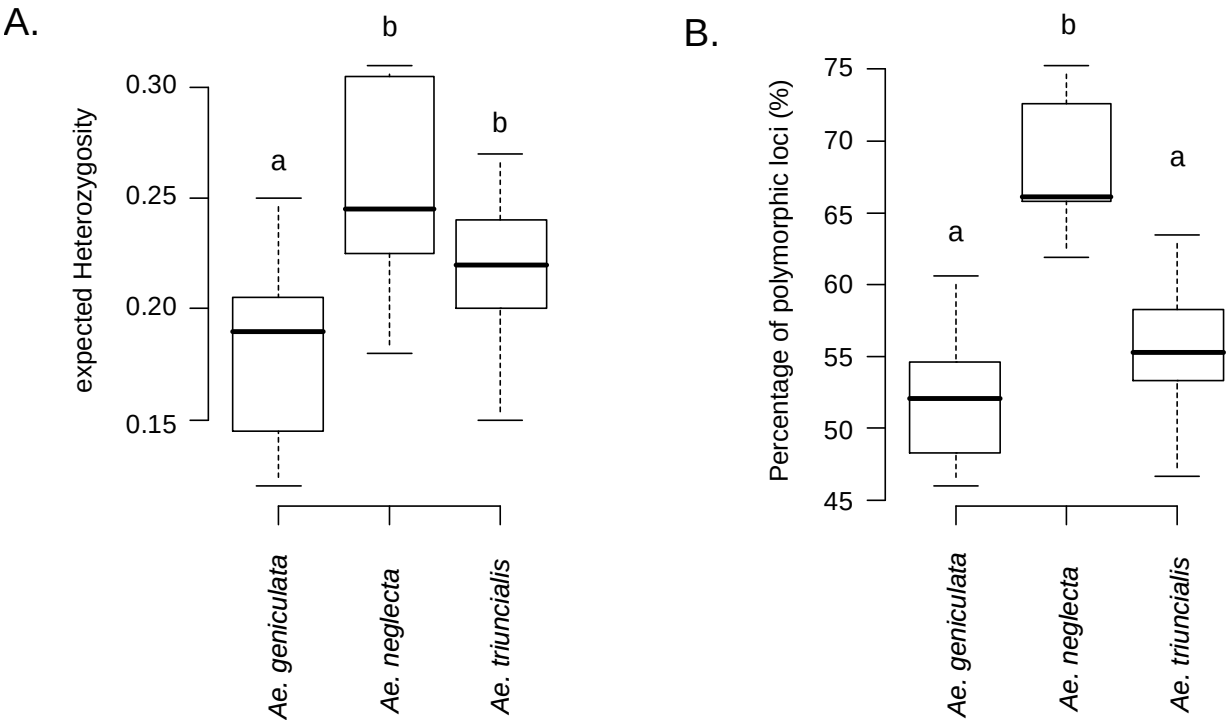

Figure 1 Diversity measures. Boxplots comparing expected heterozygosity (A.) and the percentage of polymorphic loci (B.) between populations of *Ae. geniculata* (19 populations), *Ae. neglecta* (8 populations) and *Ae. triuncialis* (13 populations). Groups that are significantly different are indicated with different letters (5% P-value level, Tukey honest differences test applied on a linear model).

Outcrossing experiment

On average, 45 experimental blocks, representing 82 selfed versus 42 allopollinated spikes could be used for each *Aegilops* species (Table 3). The fertility of spikes ranged between 5.8% and 75% but the values varied between species. Respective average fertilities were indeed 40%, 34% and 17% for *Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis*.

The experiment did not reveal a clear effect of the pollen origin on the fertility of spikes when considering the intrapopulation level. A single population (*Ae. triuncialis*: E0612A) showed a significant fertility difference between selfed spikes (19.58% of fertility) and spikes where allopollination was allowed (33.77% of fertility). The remaining populations did not show significant difference between the two spikes categories. At the intraspecific level, *Ae. geniculata* showed no difference in fertility between the two spike categories while *Ae. neglecta* and *Ae. triuncialis* had slightly increased fertilities when allopollination was allowed (this difference was significant in *Ae. triuncialis*).

Table 3 Outcrossing experiment results. Sampling effort: the number of experimental blocks (Blocks), the number of autopollinated spikes (Na) and the number of spikes allowed to allopollinate (Naa) are indicated for each population. Mean fertility indicates: the mean fertility measured in autopollinated spikes (Auto) and spikes allowed to allopollinate (Allo+Auto) and the fertility difference between both spike categories (Δ). Finally, the F-ratio and P-value are the results of the statistical tests (significance levels, 1000 permutations: * = P<0.1 and *** = P<0.05).

	Sampling effort			Mean fertility (%)			Statistical test		
	Blocks	Na	Naa	Auto	Allo+Auto	Δ	F-ratio	P-value	
Ae. geniculata									
C0723A	14	26	14	40.82	46.25	5.43	2.31	0.129	
E0610C	21	40	18	30.27	27.57	-2.70	1.38	0.263	
F1207A	10	18	10	44.58	38.75	-5.83	1.80	0.180	
I0623A	14	28	14	52.52	51.62	-0.89	0.05	0.815	
Total	59	112	56	42.05	41.05	-1.00	0.01	0.921	
Ae. neglecta									
E0615C	9	12	6	18.51	26.41	7.90	1.78	0.199	
I0624B	14	28	13	30.22	30.58	0.36	0.01	0.936	
Orb1	24	46	24	38.26	41.13	2.87	0.64	0.440	
Total	47	86	43	28.99	32.70	3.71	2.97	0.094	*
Ae. triuncialis									
E0612A	10	18	8	19.58	33.77	14.20	8.55	0.012	***
E0612B	10	16	10	16.42	15.79	-0.63	0.115	0.959	
Turk6	10	14	10	11.04	11.50	0.45	0.12	0.738	
Total	30	48	28	15.68	20.35	4.67	3.97	0.047	***

Discussion

Population genetics

The high number of bands produced by the AFLP analysis was explained by the high genetic diversity that was explored (i.e. the three species were sampled throughout a large geographical area). The mismatch error rate was similar to values obtained in other studies using automated scoring procedures (Bonin et al. 2004; Whitlock et al. 2008). *Aegilops neglecta* produced less AFLP bands than the two other species. This result could be explained as only eight European populations were analysed while a higher number of populations were investigated in *Ae. geniculata* and *Ae. triuncialis*.

For a wide range of plant species, the mating system played a critical role for the patterns of genetic variation between and within populations. Previous studies indicated that *Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis* were globally autogamous species (Hammer 1980; van Slageren 1994; Zaharieva et al. 2001a). In autogamous species, most of the genetic diversity was expected to occur between populations (Bussel 1999; Tang et al. 2003), a phenomenon that was not really supported by our AMOVA analysis since most of the diversity occurred within the populations (Table 2). This deviation might be explained by the technical noise associated to the automated scoring of AFLPs. The AMOVA analysis provided however consistent results when comparing the genetic diversity patterns of the different *Aegilops* species. The proportion of genetic diversity between populations was the highest in *Ae. geniculata* while populations of *Ae. neglecta* and *Ae. triuncialis* were less differentiated (Table 2). This result was in accordance with previous studies that outlined slight differences in the expected mating system of *Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis*. For instance, *Ae. geniculata* was considered as a relatively strict autogamous species since it revealed a poor pollen production, small stigma and important morphological variations between populations (Hammer 1980; Zaharieva et al. 2003; Bandou et al. 2009). In contrast, the higher pollen production, the larger stigma and the lower interpopulation differentiations observed in *Ae. neglecta* and *Ae. triuncialis* suggested at least some levels of allogamy (Hammer 1980; Prieto-Baena et al. 2003; Zaharieva et al. 2003). Our measures of intrapopulation diversity confirmed these results as the expected heterozygosity and the percentage of polymorphic loci (Figure 1) outlined the lowest values in *Ae. geniculata* populations. In contrast, *Ae. neglecta* and *Ae. triuncialis* were more diversified and showed estimates similar to *Ae. speltoides*, an allogamous species (Dvorak et al. 1998; Singh et al. 2006). This putative allogamy was confirmed at least in *Ae. neglecta* where several populations revealed especially high values of diversity (E0616B, I0623B and C0723C) due to gene flow originating from wheat (Arrigo et al. in prep). Finally, the results obtained with *Ae. geniculata* were in accordance with population genetic theory (Bataillon et al. 2003; Lowe et al. 2004). Indeed, an autogamous mating system increases inbreeding, decreases the frequency of heterozygotes and directly affects genotypes diversity and allelic diversity (Charlesworth 2003). In this context, the genotypes diversity depends mainly on historical factors such as founder effects, immigration or mutation events that bring new genotypes in the populations.

Outcrossing experiment

Globally, the outcrossing experiment (Table 3) did not show differences in fertility between spikes submitted to autogamy or to facultative allogamy. At the intrapopulation level, only one *Ae. triuncialis* population showed a higher fertility when allopollination was possible, compared to forced selfing. At the intraspecific level, *Ae. neglecta* and *Ae. triuncialis* showed a slight increase in fertility when allopollination was allowed. These differences suggested that allogamy may increase fertility in *Ae. neglecta* and *Ae. triuncialis*. Our experiment however lacked statistical power to draw stronger conclusions since a limited number of populations were included in the analysis. Nevertheless, the autogamy of *Aegilops* species relies mainly on cleistogamy (Hammer 1980). This mechanism may occasionally fail as widely opened spikes were observed even in putatively strict autogamous species. As an example, the Figure 2 provides pictures of spikes observed in natural conditions that opened their floral pieces and extruded fertile anthers. As a result, between 30% and 80% of the produced pollen can be shed out of spikes (Zaharieva and Monneveux 2006). Hammer et al. (1980) also reported this phenomenon in stressed spikes that failed to develop anthers. As a consequence, spontaneous hybridisation events occurring between *Aegilops* species (Feldman 1965b, 1965a) or between *Triticum* and *Aegilops* species were regularly reported and showed that autogamy in *Aegilops* species was not absolute (Benavente et al. 2001; Guadagnuolo et al. 2001; David et al. 2004; Schoenenberger et al. 2005; Loureiro et al. 2006; Schoenenberger et al. 2006; Loureiro et al. 2007, 2008). In addition, these studies outlined the quasi-absence of pre- or postzygotic barriers in the breeding system of *Aegilops* species since hybrid seeds were viable.

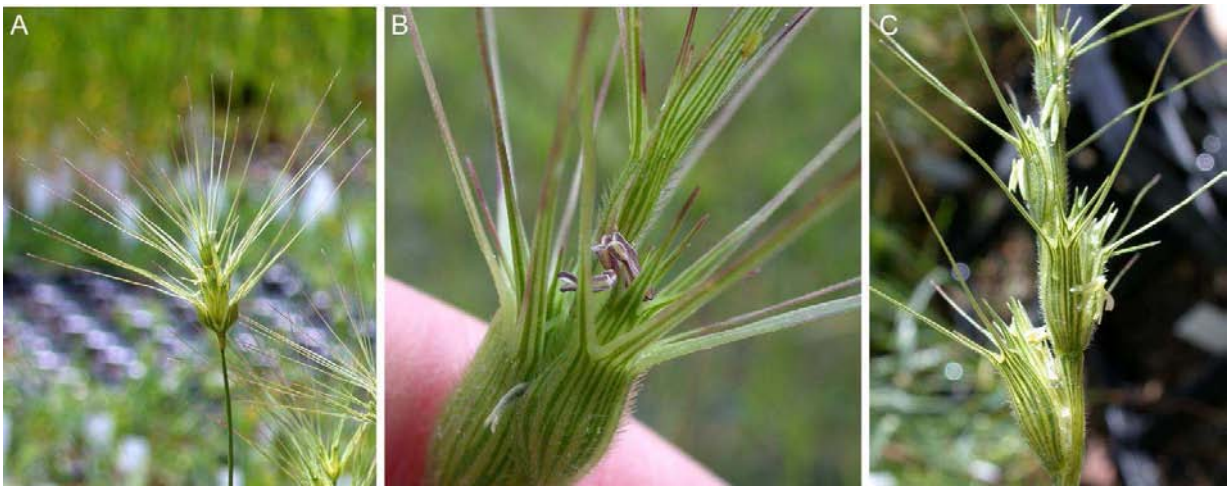

Figure 2 Fertile spikes of *Aegilops* species (A. *Ae. geniculata*, B. *Ae. neglecta* and C. *Ae. triuncialis*). These images were taken in natural conditions; they show spikes that were occasionally widely open and extruding fertile stamens. This phenomenon could potentially lead to allopollination events.

Conclusion

Population genetics analyses revealed unexpected levels of within population genetic diversity which suggest facultative allogamy for *Ae. neglecta* and *Ae. triuncialis.* Moreover, the outcrossing experiments demonstrate for these species a slight increase of fertility when allopollination was allowed, suggesting also that both species are not strictly autogamous. On the contrary, our study confirmed the higher autogamy status of *Ae. geniculata* inferred by both populations genetics and experimental outcrossing experiments.

We suggest that cleistogamy, one of the main pre-zygotic isolation mechanisms, is not absolute as widely opened spikes or hybridization events could be observed in the three species. We could assess that even if cleistogamy could decrease hybrid formation; it cannot prevent risk of transgene flow from GM wheat to *Aegilops* species. Since other studies revealed low pre- or postzygotic barriers to gene flow, we suggest that only the poor fertility of hybrids or fitness mitigation strategies could provide valuable solutions to prevent the spread of a wheat transgene in *Aegilops* populations.

Acknowledgments

This project was funded by the National Centre of Competence in Research (NCCR) Plant Survival, a research programme of the Swiss National Science Foundation. We are grateful to Renée and Olivier Arrigo and the team of the Botanical Garden of Neuchâtel for their help in plants cultivation. Nicola Schoenenberger, Anouk Beguin and Sophie Pasche provided usefull comments and technical assistance.

References

Arrigo N, Tuszynski JW, Ehrich D, Gerdes T, Alvarez N (2009) Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. BMC Bioinformatics 10.

Bandou H, Rodriguez-Quijano M, Carrillo JM, Branlard G, Zaharieva M et al. (2009) Morphological and genetic variation in *Aegilops geniculata* from Algeria. Plant Syst Evol 277: 85-97.

Bataillon T, Bonnin I, David J, Lavigne C, Le Corre V et al. (2003) Dynamique de la diversité génétique dans les populations naturelles et artificielles de plantes autogames. Les actes du BRG 4: 187-199.

Benavente E, Alix K, Dusautoir JC, Orellana J, David J (2001) Early evolution of the chromosomal structure of *Triticum turgidum–Aegilops ovata* amphiploids carrying and lacking the *Ph1*gene. Theoretical And Applied Genetics 103: 1123-1128.

Bonin A, Bellemain E, Bronken Edeisen P, Pompanon F, Brochmann C et al. (2004) How to track and assess genotyping errors in population genetics studies. Molecular Ecology 13: 3261-3273.

Braak T (1990) CANOCO - a Fortran program for canonical community ordination by correspondence analysis, principal components analysis and redundancy analysis. Wageningen, Netherlands: Agricultural Mathematics Group.

Bussel JD (1999) The distribution of random amplified polymorphic DNA (RAPD) diversity amongst populations of *Isotama petraea* (Lobeliaceae). Molecular Ecology 8: 775-789.

Charlesworth D (2003) Effects of inbreeding on the genetic diversity of populations. Phil Trans R Soc Lond B 358: 1051-1070.

Chen DH, Ronald PC (1999) A rapid DNA minipreparation method suitable for AFLP and other PCR applications. Plant Mol Biol Rep 17(1): 53-57.

David JL, Benavente E, Bres-Patry C, Dusautoir JC, Echaide M (2004) Are neopolyploids a likely route for a transgene walk to the wild? The *Aegilops ovata* x *Triticum turgidum durum* case. Biol J Linnean Soc 82(4): 503-510.

Dvorak J, Luo MC, Yang ZL (1998) Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. Genetics 148: 423-434.

Excoffier L, Laval G, Schneider S (2007) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47-50.

Feldman M (1965a) Fertility of interspecific F1 hybrids and hybrid derivatives involving tetraploid species of *Aegilops* section *Pleionathera*. Evolution 19(4): 556-562.

Feldman M (1965b) Further evidence for natural hybridization between tetraploid species of *Aegilops* section *Pleionathera*. Evolution 19(2): 162-174.

Guadagnuolo R, Savova-Bianchi D, Felber F (2001) Gene flow from wheat (*Triticum aestivum* L.) to jointed goatgrass (*Aegilops cylindrica* Host.), as revealed by RAPD and microsatellite markers. Theoretical And Applied Genetics 103(1): 1-8.

Gugerli F, Englisch T, Niklfeld H, Tribsch A, Mirek Z et al. (2008) Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation - a project synopsis. Perspect Plant Ecol 10(4): 259-281.

Hammer K (1980) Vorarbeiten zur Monographischen Darstellung von Wildpflanzensortimenten: *Aegilops* L. Kulturpflanze 28: 33-180.

Loureiro I, Escorial MC, Garcia-Baudin JM, Chueca MC (2006) Evidence of natural hybridization between *Aegilops geniculata* and wheat under field conditions in central Spain. Environ Biosafety Res 5: 105-109.

Loureiro I, Escorial MC, Garcia-Baudin JM, Chueca MC (2007) Hybridization between wheat (*Triticum aestivum*) and the wild species *Aegilops geniculata* and *A. biuncialis* under experimental field conditions. Agric Ecosyst Environ 120(2-4): 384-390.

Loureiro I, Escorial MC, Garcia-Baudin JM, Chueca MC (2008) Hybridization, fertility and herbicide resistance of hybrids between wheat and *Ae. biuncialis*. Agron Sustain Dev.

Lowe A, Harris S, Ashton P (2004) Ecological genetics: Blackwell Publishing. 344 p.

Meimberg H, Hammond JI, Jorgensen CM, Park TW, Gerlach JD et al. (2006) Molecular evidence for an extreme genetic bottleneck during introduction of an invading grass to California. Biological Invasions 8(6): 1355-1366.

Morrison LA, Cremieux L, Mallory-Smith CA (2002) Infestations of jointed goatgrass (*Aegilops cylindrica*) and its hybrids with wheat in Oregon wheat fields. Weed Science 50: 737-747.

Prieto-Baena JC, Hidalgo PJ, Dominguez E, Galán C (2003) Pollen production in the *Poaceae* family. Grana 42: 153-160.

Schneider A, Molnar I, Molnár-Láng MM (2008) Utilisation of *Aegilops* (goatgrass) species to widen the genetic diversity of cultivated wheat. Euphytica 163(1): 1-19.

Schoenenberger N, Felber F, Savova-Bianchi D, Guadagnuolo R (2005) Introgression of wheat DNA markers from A, B and D genomes in early generation progeny of *Aegilops cylindrica* Host x *Triticum aestivum* L. hybrids. Theoretical And Applied Genetics 111(7): 1338-1346.

Schoenenberger N, Guadagnuolo R, Savova-Bianchi D, Kupfer P, Felber F (2006) Molecular analysis, cytogenetics and fertility of introgression lines from transgenic wheat to *Aegilops cylindrica* host. Genetics 174(4): 2061-2070.

Singh H, Chabane K, Valkoun J, Blake NK (2006) Optimum sample size for estimating gene diversity in wild wheat using AFLP markers. Genet Resour Crop Evol 53: 23-33.

Tang T, Zhong Y, Jian S, Shi S (2003) Genetic diversity of *Hibiscus tiliaceus* (Malvaceae) in China assessed using AFLP markers. Annals of Botany 92: 409-414.

van Slageren MW (1994) Wild wheats: A monograph of *Aegilops* L.and *Ambylopyrum* (Jaub and Spach) Eig (*Poaceae*). Wageningen: Agricultural Univ., Wageningen, The Netherlands and Int. Cent. for Agric. Res. in Dry Areas, Aleppo, Syria.

Vekemans X, Beauwens T, Lemaire M, Roldan-Ruiz I (2002) Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Molecular Ecology 11: 139-151.

Waines JG, Barnhart D (1992) Biosystematic research in *Aegilops* and *Triticum*. Hereditas 116(3): 207-212.

Whitlock R, Hipperson H, Mannarelli M, Butlin R, Burke T (2008) An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error. Mol Ecol Resour 8: 725-735.

Zaharieva M, Monneveux P (2006) Spontaneous hybridization between bread wheat (*Triticum aestivum* L.) and its wild relatives in Europe. Crop Science 46(2): 512-527.

Zaharieva M, Santoni S, David J (2001a) Use of RFLP markers to study genetic diversity and to build a core-collection of the wild wheat relative *Ae. geniculata* Roth (= *Ae. ovata* L.). Genetics Selection Evolution 33: S269-S288.

Zaharieva M, Gaulin E, Havaux M, Acevedo E, Monneveux P (2001b) Drought and heat responses in the wild wheat relative *Aegilops geniculata* Roth: Potential interest for wheat improvement. Crop Science 41(4): 1321-1329.

Zaharieva M, Dimov A, Stankova P, David J, Monneveux P (2003) Morphological diversity and potential interest for wheat improvement of three *Aegilops* L. species from Bulgaria. Genetic Resources and Crop Evolution 50: 507–517.

Chapter 4

Evidence of gene flow in the *Aegilops* L. – *Triticum* L. species complex: the case of *Ae. geniculata* Roth, *Ae. neglecta* Req. ex Bertol. and *Ae. triuncialis* L.

Arrigo N[1], Guadagnuolo R[1], Lappe S[1], Pasche S[1] and Felber F[1]

[1]Laboratory of Evolutionary Botany, Institute of Biology, University of Neuchâtel, 11 rue Emile-Argand, 2000 Neuchâtel, Switzerland

Abstract

The domestication of wheat (*Triticum* L.), one of the major crop presently cultivated, results from recurrent allo-polyploidization events involving several *Aegilops* L. species. The reproductive isolation between species of both genera is incomplete. *Aegilops* species were traditionally used by wheat breeders as a source of beneficial traits for selection of new varieties. Conversely, F1 hybrids between wheat cultivars and *Aegilops* species were regularly reported in agroecosystems. Despite these observations, long term consequences of gene transfer of wheat in the *Aegilops* genetic pool have been poorly documented. This topic is however of great concern in risk assessment studies related to the commercial release of transgenic wheat.

The present study investigates three European tetraploid *Aegilops* species (*Ae. geniculata* Roth, *Ae. neglecta* Req. ex Bertol. and *Ae. triuncialis* L.) in natural Mediterranean populations growing in Spain, France, Italy and Croatia. The sample design compared samples collected along wheat field borders to samples harvested in areas isolated from agriculture. AFLP markers were used to assess the presence of wheat genetic markers in *Aegilops* populations.

The results varied according to *Aegilops* species: *Ae. geniculata*, the most autogamous species included in our study, showed no clear evidence of gene flow from wheat, except for one *Ae. geniculata* x *Triticum* F1 hybrid discovered in a Spanish population. In contrast, an unexpectedly high amount of wheat genetic markers were observed in *Ae. neglecta* and *Ae. triuncialis* samples. Interestingly, most of the introgressed samples were from populations collected near crop fields. In contrast, samples collected in areas isolated from agriculture showed almost no evidence of gene flow from wheat. These results provided reasonable evidence that long-term introgression of wheat markers has occurred at least in *Ae. neglecta* and *Ae. triuncialis*. Finally, *Triticum turgidum* L. (the tetraploid pasta wheat) was the most likely wheat parent of introgressed *Aegilops*. In contrast, *Triticum aestivum* L. (the hexaploid bread wheat) appeared only to be the parent of the *Ae. geniculata* x *Triticum* F1 hybrid.

Introduction

The recent progresses in biotechnologies allowed the development of genetically modified organisms. As for many other crop species, new wheat varieties were genetically modified to confer for example, increased yield or other agronomical traits of interest (softer endosperm, increase of the glutenin content), tolerances to herbicides (glyphosate, glyphosinate), fungi (*Fusarium* or smut), pests (production of toxins such as snow-drop lectin) or drought. Up to now, if many of these varieties have been evaluated under field tests (http://www.isb.vt.edu/, accessed the 13th of March 2009), none are commercialized. Indeed, wheat is of great commercial interest as it is the most important crop species with 607 million tons harvested annually (FAOSTAT 2007). The main species, the "bread wheat" *Triticum aestivum* L., is a cosmopolitan crop grown under temperate climates and represents 95% of the annual yield. The remaining five percents are essentially represented by the "durum wheat", *Triticum turgidum* L., a crop grown mainly under semi-arid climates in Western Asia, Southern Europe, North America and the Canada and used for the pasta and semolina production (USDA, http://www.fas.usda.gov/pecad/highlights/2005/07/durum2005/, accessed the 13th of March 2009). Both species are allopolyploids; *T. turgidum* is a tetraploid species (2n = 4x = 28) with AABB as genomic formula while *T. aestivum* is hexaploid (2n = 4x = 42) with AABBDD as genomic formula (Waines and Barnhart 1992). The genomes are organized as sets of seven chromosomes that originate from one *Triticum* and two diploid *Aegilops* species: *Triticum urartu* Tumanian ex Gandilyan (A genome), *Ae. speltoides* Tausch (B genome), and *Ae. tauschii* Coss. (D genome).

Aegilops L. includes wild relatives of the wheat and represents a gene pool that is regularly used by breeders in wheat improvement programs (Schneider et al. 2008). Several other genomes occur in *Aegilops* (C, M, S, U, D and N genomes); they originate from parental diploid species and are shared between derived allopolyploid species. For instance, *Ae. geniculata* Roth (MMUU), *Ae. triuncialis* L. (UUCC) and *Ae. neglecta* Req. ex Bertol. (UUMM or UUMMNN), the three most common *Aegilops* species in southern Europe (van Slageren 1994), are tetraploid or hexaploid (Waines and Barnhart 1992). *Aegilops* is a Mediterranean-Western asiatic genus that grows in habitats characterized by hot and dry summers and rainy winters. The life strategy of *Aegilops* species relies on an vegetative growth during winter and a reproductive period achieved prior to summer. They occur in ruderal and segetal habitats such as for instance, edges of cultivations, roadsides, pastures, orchards or olives groves. Some polyploid species spread on large areas and are even considered as invasive in the USA. For instance, *Ae.* cylindrica Host is a major wheat fields weed (Morrison et al. 2002) while *Ae. geniculata* and *Ae. triuncialis* are mainly invading pastures and disturbed ecosystems of California (van Slageren 1994; Watanabe and Kawahara 1999; Meimberg et al. 2006).

The reproductive isolation between *Aegilops* and *Triticum* is not complete. Both genus grow sympatrically in Europe and their flowering periods overlap (Zaharieva and Monneveux 2006). As a result, the natural occurrence of viable F1 hybrids between several *Aegilops* species and the *Triticum* genus is regularly reported (Guadagnuolo et al. 2001; Morrison et al. 2002; David et al. 2004; Karagöz 2006; Loureiro et al. 2006, 2007a, 2008). The hybridization rates can be significant and range between 0.20% and 0.45% (Loureiro et al. 2006). F1 hybrids are generally male sterile while having low percentages of female fertility. A recent review reported that female fertilities of *Aegilops x Triticum*

F1 hybrids ranged between 1 and 4% (Zaharieva and Monneveux 2006). As a result, backcrossed generations can be obtained experimentally or under field conditions (Schoenenberger et al. 2005; Schoenenberger et al. 2006; Loureiro et al. 2007a, 2008). BC1 individuals generally have a partially restored male fertility that allows selfing (Schoenenberger et al. 2006). Female fertility is also increased with values as high as 15.1% and 37.4%, reported for the BC1 and BC2 generations in *Ae. cylindrica* backcrosses (Zaharieva and Monneveux 2006). Recombination events involving homoeologous chromosomes were observed during meioses of F1 hybrids (Wang et al. 2000; David et al. 2004; Cifuentes et al. 2006). The mechanism was shown to be governed genetically by some gene complexes (*Ph* genes) located on the wheat 5B chromosome (Sears 1976) that prevents homoeologous pairings. Although the first steps of a wheat gene transfer into the *Aegilops* genome are documented, only few studies (Weissmann et al. 2005) report an effective introgression involving wheat and *Aegilops* species under natural conditions.

The present study assesses the putative transfer of wheat genes within three common Mediterranean *Aegilops* species: *Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis*, using Amplified Fragment Length Polymorphism, a reliable method that was successfully used to identify *Aegilops* species and detect hybridization and introgression (O'Hanlon et al. 1999; Monte et al. 2001; Pester et al. 2003; Bonin et al. 2007). The study relies on the detection of wheat AFLP bands (hereafter termed “WB”) into Mediterranean *Aegilops* populations. The sample design compares populations growing along cultivation borders to populations collected in agriculture isolated areas, in order to allow the identification of WBs originating from introgression from those jointly inherited from a common ancestor between wheat and *Aegilops* species.

Material and methods

Sample design and plant material

The present study compared:

I. Populations growing close to wheat fields. These samples were collected between 0 and 50 meters away from a wheat field and were designated as “Close to wheat” samples.

II. Populations isolated from wheat cultivations. These samples were collected at more than 50 meters from a wheat field and were designated as “Distant from wheat” samples.

The comparisons of “Close to wheat” and “Distant from wheat” samples involved three *Aegilops* species (Table 1): *Ae. geniculata* (250 samples, 20 populations), *Ae. neglecta* (90 samples, 10 populations) and *Ae. triuncialis* (133 samples, 15 populations). Populations where several species grew in sympatry were also considered since we assumed a reproductive isolation between the three *Aegilops* species included in the present study. This assumption was supported by the works of Chee et al. (1995). Samples were collected as seeds during field surveys conducted in Spain, France, Italy, Croatia and Turkey. They were grown at the Botanical garden of the city and University of Neuchâtel to provide a large collection of plants conserved as herbaria sheets and silica-gel dried leaves. As a result, the dataset included 473 samples and was organized in three subsets (*Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis* subsets) where “Close to wheat” and “Distant from wheat” samples were compared. Each *Aegilops* subset was completed with 71 wheat reference samples that were either collected during field collections or provided by the Agroscope Changins-Wädenswil. Wheat samples represented a significant proportion of the wheat genome diversity (with 28 lines of tetraploid *T. turgidum* and 43 lines of hexaploid wheat, *T. aestivum*) and were used as reference to detect a putative transfer of wheat genes into the *Aegilops* pool.

Table 1. Description of plant populations : country, geographical coordinates (decimal degrees, NDD, EDD), distance to wheat cultivations (D.wheat), number samples of *Ae. geniculata* (*Ae. gen*.), *Ae. neglecta* (*Ae. negl*.), *Ae. triuncialis* (*Ae. triun*.), *T. turgidum* (*T. turg*.) and *T. aestivum* (*T. aest*.) included in the AFLP analysis are indicated.

Population	Country	NDD	EDD	D.wheat	*Ae. gen.*	*Ae. negl.*	*Ae. triun.*	*T. turg.*	*T. aest.*
Croatia									
C0724A	Croatia	43.51	16.09	Close to wheat (10m)			10		
C0723C	Croatia	43.47	16.69	Close to wheat (30m)		8	9		
France									
F070606D	France	43.68	3.91	Close to wheat (1m)	17				
F1207A	France	44.05	6.06	Close to wheat (50m)	11				
F1307A	France	43.39	5.13	Distant from wheat (600m)	13				
F070406D	France	43.71	3.68	Distant from wheat (1.3Km)	16				
Italy									
I0623B	Italy	43.59	11.57	Close to wheat (1m)		9			
I0624B	Italy	43.18	11.14	Close to wheat (1m)		15			
I0624C	Italy	43.07	10.91	Close to wheat (1m)	11	1			
I0626B	Italy	44.69	9.99	Distant from wheat (100m)			12		
I0624A	Italy	43.15	11.18	Distant from wheat (120m)	17				
I0623D	Italy	43.18	11.37	Distant from wheat (400m)	4	1			
I0623A	Italy	43.54	11.42	Distant from wheat (1.2Km)	12				
Orb1	Italy	42.47	11.34	Distant from wheat (>20Km)		12			
Spain									
E0609B	Spain	40.01	-2.96	Close to wheat (1m)	23		4		
E0610B	Spain	39.98	-2.09	Close to wheat (1m)	15		6		
E0615A	Spain	38.26	-5.70	Close to wheat (1m)	5		13		
E0616A	Spain	39.31	-5.72	Close to wheat (1m)		11			
E0616B	Spain	39.46	-5.17	Close to wheat (1m)		7			
E0618A	Spain	41.22	-3.88	Close to wheat (1m)	22		9		
E0610E	Spain	39.22	-1.94	Close to wheat (10m)	5		9		
E0612A	Spain	37.54	-4.34	Close to wheat (10m)			7		
E0616D	Spain	39.97	-4.71	Close to wheat (20m)	11		7		
E0617A	Spain	40.75	-4.66	Close to wheat (20m)	14				
E0615C	Spain	39.17	-6.25	Distant from wheat (200m)		14	7		
E0613A	Spain	36.77	-5.27	Distant from wheat (800m)	3	12	8		
E0610C	Spain	39.85	-2.02	Distant from wheat (2.4Km)	25				
E0612B	Spain	37.46	-4.32	Distant from wheat (2.8Km)			14		
E0613B	Spain	36.76	-5.39	Distant from wheat (10.4Km)	7				
E0611B	Spain	38.15	-3.02	Distant from wheat (5Km)	14		7		
E0611D	Spain	37.57	-4.29	Distant from wheat (5Km)	5				
Turkey									
Turk6	Turkey	38.43	43.26	Distant from wheat (>20Km)			11		
Reference wheat samples									
CHGSEspturg	Spain							5	
CHGSEspdurum	Spain							4	
CHGSEspaest	Spain								7
CHGSEspspelta	Spain								2
CHGSFradurum	France							3	
CHGSFraturg	France							2	
Tdur INRA	France							12	
CHGSFraaest	France								13
CHGSItaturg	Italy							2	
CHGSItaaest	Italy								18
CHGSItaspelta	Italy								1
tspI0623B	Italy								1
tspI0626A	Italy								1

Molecular analyses and AFLP scoring

DNA was extracted from 10 mg of silica-dried leaves (see below) using a CTAB protocol (Chen and Ronald 1999). DNA concentrations were standardized at 10 ng/ l. We followed the AFLP protocol developed by Gugerli et al. (2008). Reactions were conducted with a PTC-100 (MJ Research) thermocycler, using 96-wells plates where samples were randomly distributed. We used two selective primer pairs (EACT-MCTG and EAGT-MCAT). Primers and adaptors sequences are indicated in Table 2. PCR products were recorded using the GeneScan technology (primers flanking *Eco-RI* restriction site were fluorescently labelled with 5' 6-carboxy-fluorescein [FAM]) with a capillary sequencer (ABI 3730XL, service provided by Macrogen Inc. Seoul, Korea). Raw electropherograms were analysed with PeakScanner V1.0 (ABI, peak detection parameters: default parameters except a light smoothing and 100 rfu as fluorescence threshold) in order to detect and calculate the size of AFLP bands. The scoring (consisting in recording the presence / absence of AFLP bands for each sample) was performed on the complete set of samples (i.e. 473 *Aegilops* sp. plus the 71 wheat samples) using RawGeno, an automated scoring R CRAN package (Arrigo et al. 2009). The library was settled as follows: scoring range = 100 – 400 bp, minimum bin width = 0 bp, maximum bin width = 2 bp and closely sized bins (5%) were removed. Finally, the scored matrices of both primer pairs were concatenated in a single binary matrix where individuals and bands were stored as lines and columns respectively.

We replicated several AFLP reactions: five samples and one blank were distributed on each 96-wells plate to check the between plate variability and the purity of templates. Additionally, 11 samples (chosen randomly from each plate and representing 12 percents of the final dataset) were replicated on another plate in order to calculate the error rate (Bonin et al. 2004) and subsequently eliminate non-reproducible bands. We discarded bands that were not reproducible at a 95% threshold (representing 17% and 15% of the total number of detected bands for EACT-MCTG and EAGT-MCAT respectively).

Table 2. Primer sequences. The AFLP analysis needed the use of AE and AM adaptators during the ligation step (AE is built by annealing AET with AEB while AMT and AMB are used to build AM). The preselective and selective primers are indicated. Selective primers flanking the *Eco-RI* restriction site were labelled fluorescently using 5' 6-carboxy-fluorescein (FAM).

Analysis	Primer name	Sequence (5'-3')
AFLP Ligation	Adaptator AET	CTCGTAGACTGCGTACC
AFLP Ligation	Adaptator AEB	AATTGGTACGCAGTCTAC
AFLP Ligation	Adaptator AMT	GACGATGAGTCCTGAG
AFLP Ligation	Adaptator AMB	TACTCAGGACTCAT
AFLP Pre-Selective PCR	Primer EcoRI-A	GACTGCGTACCAATTCA
AFLP Pre-Selective PCR	Primer MseI-C	GATGAGTCCTGAGTAAC
AFLP Selective PCR 1	Primer EcoRI-AGT	5'FAM-GACTGCGTACCAATTCAGT
AFLP Selective PCR 1	Primer MseI-CAT	GATGAGTCCTGAGTAACAT
AFLP Selective PCR 2	Primer EcoRI-ACT	5'FAM-GACTGCGTACCAATTCACT
AFLP Selective PCR 2	Primer MseI-CTG	GATGAGTCCTGAGTAACTG

Data analyses

Principal coordinates analysis

The complete dataset (i.e. the 473 *Aegilops* sp. plus 71 wheat samples) was first investigated using a principal coordinates analysis using the Jaccard distance. R CRAN was used to compute the analysis using the "vegan" R package.

STRUCTURE analyses

Bayesian Inference methods (STRUCTURE 2.2, (Pritchard et al. 2000), version adapted for dominant datasets) were used to analyze *Aegilops* samples according to wheat reference samples. The algorithm works without *a priori* information and assigns samples within K groups (K, the number of groups, was the only constraint given to STRUCTURE) that were defined in a way to optimize the Hardy-Weinberg equilibrium within the K groups. We used an admixture model to detect a mixed ancestry in *Aegilops* samples (we used the default settings). Since wheat and *Aegilops* species shared a common ancestry, we assumed that allelic frequencies were correlated between the K groups. The Markov chain was set up with 100000 steps of burn-in and 400000 steps for data acquisition.

The analysis was performed separately with each *Aegilops* subset (where "Close to wheat" and "Distant from wheat" samples were analysed conjointly). Within each *Aegilops* subset, STRUCTURE was used to assign samples to K = 3 groups. This value was chosen since three groups could be expected (i.e. the *Aegilops* sp., the *T. turgidum* and the *T. aestivum* groups). Each STRUCTURE analysis was replicated three times and only runs that obtained the highest maximum-likelihood values were kept for further analyses. As expected, STRUCTURE systematically assigned *Aegilops* sp., *T. turgidum* and *T. aestivum* samples to three different genetic groups. As a result we obtained, for each subset of samples, a table characterizing each sample with three assignment probabilities: $P_{Aegilops}$, P_{4x} and P_{6x} that were respectively the probability of belonging to the "*Aegilops* sp.", the "*T. turgidum*" or the "*T. aestivum*" genetic groups. As $P_{Aegilops} + P_{4x} + P_{6x} = 1$, we calculated, $P_{wheat} = P_{4x} + P_{6x} = 1 - P_{Aegilops}$, the probability of belonging to the "wheat" genetic pool (sensu lato).

Wheat assignment probabilities (P_{wheat}) of both "Close to wheat" and "Distant from wheat" samples were compared with histograms within each *Aegilops* subset and distributions of wheat assignment probabilities were compared between populations (Additional file 1). Finally, the complete set of assignment probabilities were displayed using triangle plots in order to assess whether *T. turgidum* or *T. aestivum* was the most likely source of wheat genes in *Aegilops* samples.

Validation of STRUCTURE results

We validated STRUCTURE results using two additional analyses that were applied independently on each *Aegilops* subset:

I. Hybrid index. We used H_{index} (Buerkle 2005), a robust algorithm that uses ML methods and assumes Hardy-Weinberg equilibrium within the defined parental pools. The calculation needed the "a priori" definition of two "parental" pools to calculate a hybridization index (hereafter "H_{index}") on "descent" samples. The H_{index} ranged between zero and one and indicated the probability that a sample belong to one of the two parental pools. In the present analysis, the parental pools were the *Triticum* pool (that included *T. turgidum* and *T. aestivum* samples) and the *Aegilops* sp. pool. The descent pool, on which we calculated the H_{index}, was the *Aegilops* sp. pool. This strategy was applied to avoid the "a priori" definition of potentially introgressed *Aegilops* sp. samples.

II. Detection of wheat specific AFLP bands (hereafter "WB analysis"). The analysis was inspired from the works of Doebley (1990) and assumed "a priori" that wheat and *Aegilops* sp. samples were distinct genetic pools. It started with computing the frequency of AFLP bands in the *Triticum* (*T. aestivum* and *T. turgidum* samples were pooled) and *Aegilops* sp. pools. On this basis, AFLP bands that were frequent in the *Triticum* pool (i.e. AFLP bands present in more than 50% of the *Triticum* samples) while rare in the *Aegilops* pool (i.e. absent in more than the 90% of *Aegilops* samples) were selected. These bands, assumed to be wheat specific, were designated as "wheat specific bands" (i.e. hereafter "WB") and their occurrence was counted in *Aegilops* samples.

Finally, wheat assignment probabilities obtained with STRUCTURE (i.e. P_{wheat}) were compared to the results of Hindex and the WB analysis through scatter plots and Pearson's correlation tests. Only *Aegilops* sp. samples were included in these comparisons.

Results

AFLP dataset accuracy

The AFLP analysis provided an accurate dataset with a total of 1067 bands (570 and 497 bands for EACT-MCTG and EAGT-MCAT respectively) and an average of 170 bands per sample. The measured error rate was 5.8% of non-matching bands per sample-replicate pair.

The dataset included many private alleles and allowed an accurate identification of species. As a result, *Ae. geniculata*, *Ae. neglecta*, *Ae. triuncialis*, *T. turgidum* and *T. aestivum* were clearly identified as distinct genetic entities (Figure 1). Both species of *Triticum* were not clearly separated on the Figure 1, but clearly displayed distinct genotypes when removing *Aegilops* samples from the principal coordinates analysis (data not shown). This discriminatory power was confirmed by the high percentages of variance explained by the first axes of the principal coordinates analysis. For instance, the three first axes captured 51%, with respectively 23%, 15% and 13% of the total variance.

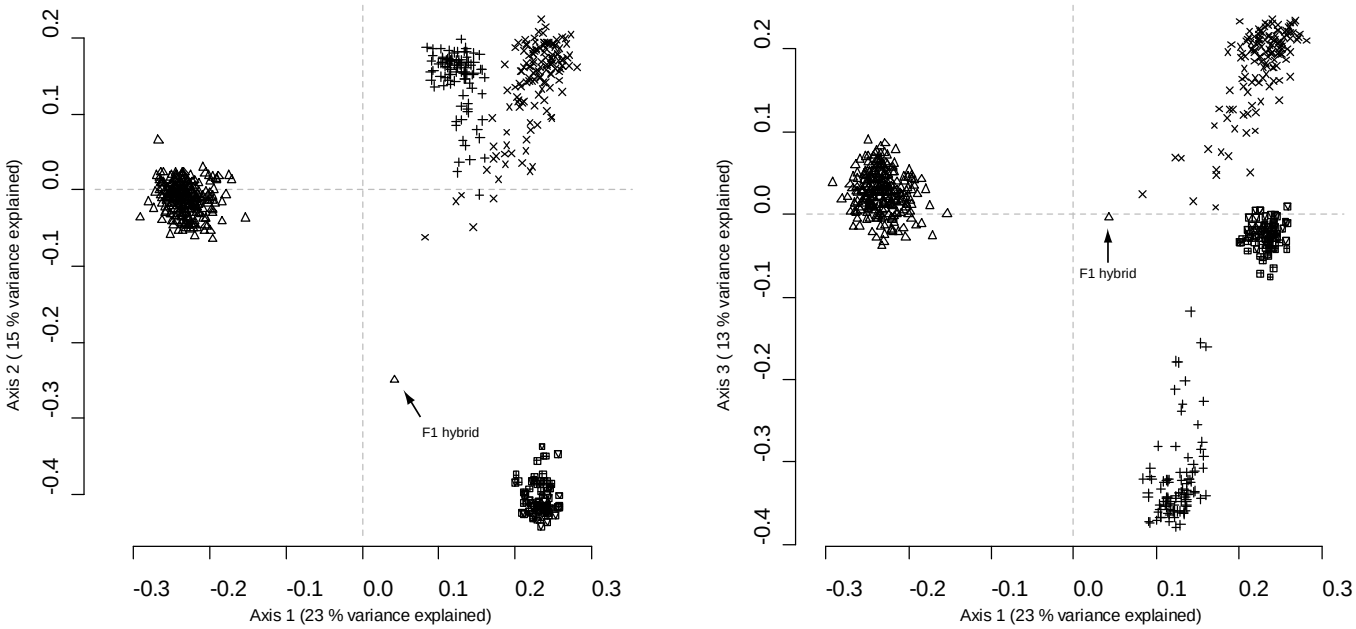

Figure 1. Principal coordinates analysis of the complete AFLP dataset. The three first axes and their contribution in explaining the global variance are displayed. Samples are labeled according to their species, *Ae. geniculata* (∆); *Ae. neglecta* (+), *Ae. triuncialis* (x), *T. aestivum* ("+" square) and *T. turgidum* ("v" square). Left panel: axis one and axis two, right panel: axis one and axis three. The *Ae. geniculata* x *T. aestivum* F1 hybrid is highlighted.

Validation of STRUCTURE results

Wheat assignment probabilities followed a "poisson-like" distribution as $P_{wheat} \leq 0.1$, $P_{wheat} \leq 0.2$ and $P_{wheat} \leq 0.35$ included respectively the 90%, 95% and 99% of the samples. These results were validated with H_{index} and the WB analysis (Figure 2). Hybrid indexes were highly correlated to STRUCTURE results and showed Pearson's correlations of 0.69***, 0.97*** and 0.89*** respectively for the *Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis* subsets. Similar results were obtained with the WB analysis as correlations of 0.38***, 0.83*** and 0.81*** were obtained with the *Ae. geniculata*, *Ae. neglecta* and *Ae. triuncialis* subsets.

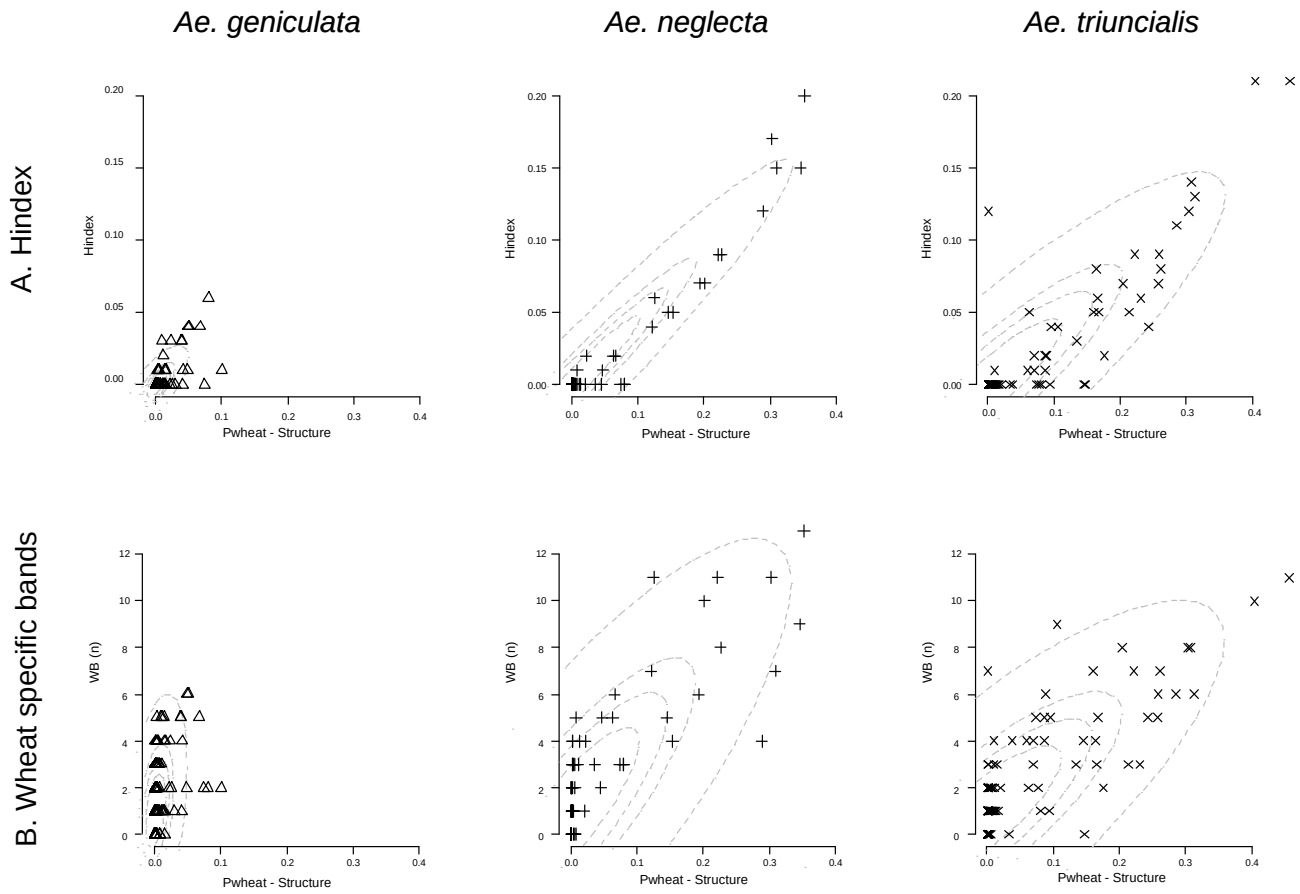

Figure 2. Validation of STRUCTURE results. The three *Aegilops* species were analyzed independently and comparisons are displayed separately. Scatterplots are displayed in an array where lines represent the different statistics that are compared while columns represent the three *Aegilops* species; *Ae. geniculata* (Δ); *Ae. neglecta* (+), *Ae. triuncialis* (x). Wheat assignment probabilities calculated with STRUCTURE (P_{wheat}) are compared with either A. the Hybrid index (H_{index}) or B. the number of wheat specific bands (WB) per sample. Ellipses were added to represent the distribution of samples (50%, 75%, 90% and 99.9% of the samples). The *Ae. geniculata* x *T. aestivum* F1 hybrid was not included in the analysis.

Indications of gene flow

The survey of European populations revealed one spontaneous hybridization event between *Ae. geniculata* and *T. aestivum*. The F1 hybrid, discovered in the progenies of a Spanish population that grew along a wheat field border (Castille e Leon, near Valdesimonte). This sample showed an intermediate phenotype between its parental species. Its hybrid status was successfully confirmed by the AFLP analysis that revealed an intermediate genotype between *Ae. geniculata* and *T. aestivum* (Figure 1). The *Ae. geniculata* subset provided no other clear patterns as the majority of samples obtained low wheat assignment probabilities (the highest probability was observed in a sample distant from 600 meters of a wheat field, with $P_{wheat} = 0.10$). In contrast, a significant number of samples obtained wheat assignment probabilities exceeding 0.1 in *Ae. neglecta* (13 samples, 15% of the subset) and *Ae. triuncialis* (23 samples, 17% of the subset). These samples appeared on the principal coordinates analysis (Figure 1) as a gradient of intermediate genotypes between *Aegilops* sp. and the wheat pool. These measures were confirmed by H_{index} and the WB analysis that revealed an increased number of wheat specific AFLP bands in the intermediate genotypes (Figure 2). Interestingly, these samples occurred mainly in populations growing close to wheat cultivations (Figure 3). Indeed, 90 % of the samples that obtained a $P_{wheat} > 0.1$ were observed within the "Close to wheat" samples and were collected within a 50 meters range from a wheat field. The remaining 10% of introgressed samples were collected 800 meters away from a wheat field, in a single Spanish population located in a mixed landscape of pastures and crop cultivations (E0613A, Andalousia). At the population level, the occurrence of samples with $P_{wheat} > 0.1$ ranged between one and seven samples per population and represented on average 15% of the population samplings (Additional file 1).

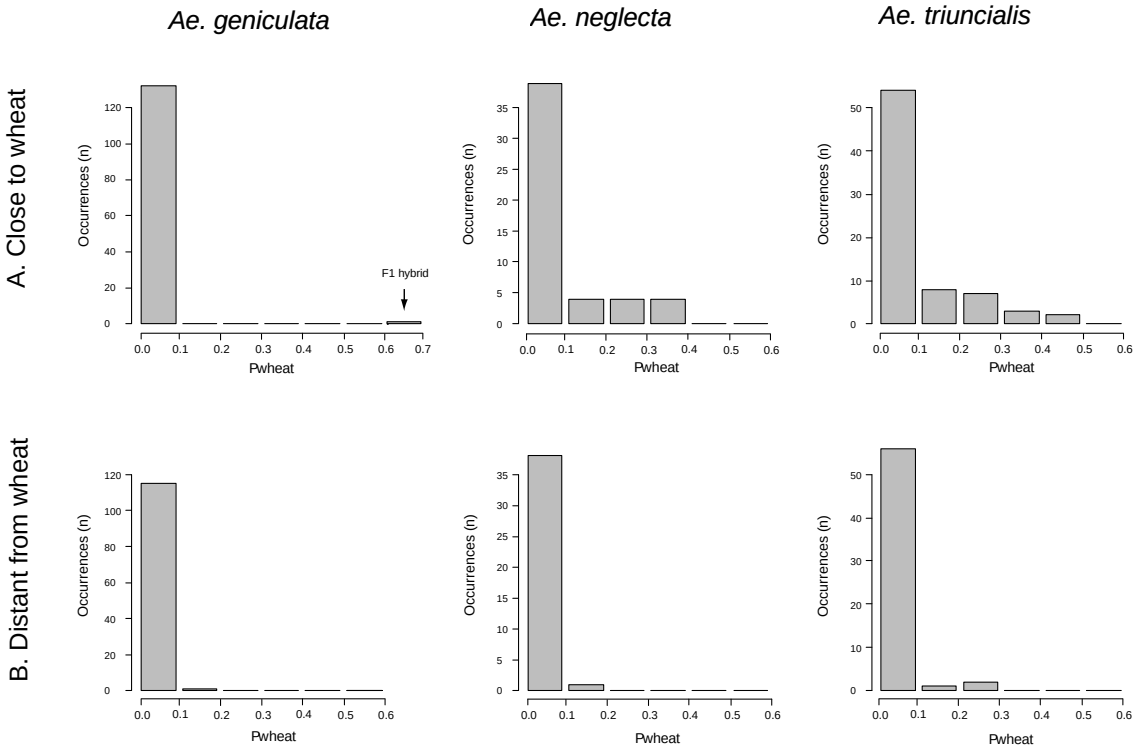

Figure 3. Histograms of absolute frequencies of wheat assignment probabilities (P_{wheat}), for the three *Aegilops* species, collected either close (A.) or distant (B.) from wheat cultivations. The *Ae. geniculata* x *T. aestivum* F1 hybrid is highlighted.

Finally, a more detailed investigation of STRUCTURE results revealed that most of the putatively introgressed samples possessed a P_{4x} greater than P_{6x} (Figure 3.). In other words, *T. turgidum* appeared to be the most likely wheat parent of the introgressed samples.

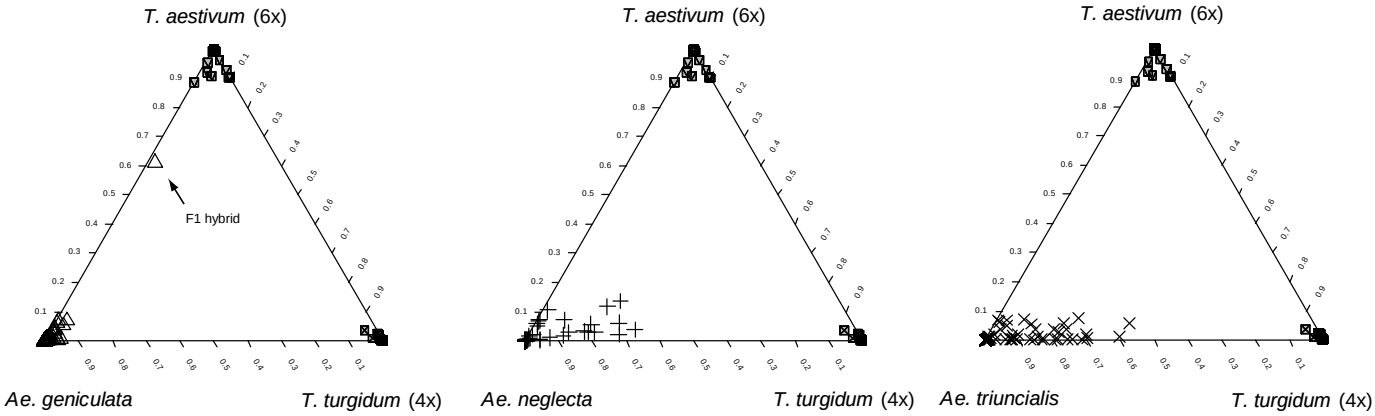

Figure 4. Triangle plot. The three *Aegilops* species were analyzed independently and comparisons are displayed separately. Each sample is displayed according to the three assignment probabilities that were calculated using STRUCTURE. $P_{Aegilops}$, P_{4x} and P_{6x} are assignment probabilities to the "*Aegilops* sp.", the "*T. turgidum*" or the "*T. aestivum*" genetic groups. Samples are labeled according to their species, *Ae. geniculata* (Δ); *Ae. neglecta* (+), *Ae. triuncialis* (x), *T. aestivum* ("+" square) and *T. turgidum* ("v" square). The *Ae. geniculata* x *T. aestivum* F1 hybrid is highlighted.

Discussion

The AFLP analysis provided valuable results. First, the error rate was comparable to values obtained by other AFLP studies that also used automated scoring procedures (Bonin et al. 2004; Whitlock et al. 2008). Second, we obtained accurate assignment probabilities. The high number of private AFLP bands allowed a clear identification of the different species, as shown by the principal coordinates analysis (Figure 1) and STRUCTURE that systematically discriminated *Aegilops* species from wheat samples. These results were confirmed by two independent approaches, Hindex and the WB analysis (Figure 2), that ruled out the hypothesis of a calculation bias in our wheat assignment probabilities. In addition, the F1 hybrid showed an AFLP genotype supporting it expected genomic composition. This sample was probably pentaploid since it derived from the *Ae. geniculata* x *T. aestivum* crossing; where the parents contributed 14 (U^gM^g) and 21 (ABD) chromosomes respectively. As a result, *Ae. geniculata* probably provided 40% of the F1 genome while *T. aestivum* provided the remaining 60%, a proportion supported by the calculated wheat assignment probability for this F1 hybrid (P_{wheat} = 0.63). A similar additive phenomenon in hybrid genotypes was observed with RAPD analyses of F1 hybrids produced with *Ae. cylindrica* and *T. aestivum* (Guadagnuolo et al. 2001).

Interestingly, the majority of putatively introgressed samples occurred within a 50 meters perimeter from wheat fields (Figure 3). This observation confirmed the hypothesis of hybridization with wheat as

neither AFLP nor calculation biases (e.g. homoplasy or deviations from the Hardy-Weinberg equilibrium (Vekemans et al. 2002; Arrigo et al. 2009)) could be expected to segregate samples according to their distance to cultivations. These results were validated by observations made under natural conditions. The majority of *Aegilops* sp. x *Triticum* sp. F1 hybrids were indeed found close to wheat cultivations (Zaharieva and Monneveux 2006), an expected result since wheat pollen mainly disperses within a 20 meters range (i.e. 30 meters was recently proposed as isolation distance between wheat cultivations to limit wheat cross-pollinations at a 0.01% level (Matus-Cadiz et al. 2004; Hanson et al. 2005; Loureiro et al. 2007b)). However these results may vary according to the cultivars and the meteorological conditions as pollen flow was reported at more than 2.75 Km from their source (De Vries 1971; Zaharieva and Monneveux 2006; Matus-Cadiz et al. 2007). This long range pollen dispersal might explain the observation of introgressed individuals in sites isolated from wheat, such as the E0613A population where a small number of introgressed samples were collected 800 meters away from a wheat field. Alternatively, cattle or human activities may also disperse seeds on long distances. For instance, this population grew in a region where an active sheep itinerant pastoralism (www.juntadeandalucia.es/agriculturaypesca/portal/www/portal/com/bin/portal/DGPAgraria/Estudios_Prospectiva/Estudios_Informes/politica_agraria/pac03111801.pdf, accessed the 13th of March 2009) is likely to promote seed exchanges between wheat field borders and adjacent pastures. Finally, remnant traces of introgression may explain the results obtained in the E0613A population. Cultivated surfaces have decreased in Andalousia by the progressive conversion of crop cultures into pastures and the E0613A population area might have been cultivated 50 years ago (Villegas Molina et al. 1990). This phenomenon was observed by Weissman et al. (2005) in *Ae. peregrina* samples collected in an abandoned wheat field.

The three *Aegilops* species showed different levels of introgression (Figure 3). No clear evidences of wheat gene transfers were observed in *Ae. geniculata.* In contrast, *Ae. neglecta* and *Ae. triuncialis* appeared to be significantly introgressed. This result was in accordance with the mating system of *Aegilops* species. Like wheat, *Aegilops* species are autogamous since they produce almost cleistogamous florets (i.e. closed flowers that mechanically prevent the allopollination). This isolation system is enhanced by self-pollination that takes place at the early stages of stigma receptivity, while the spikelets are closed (Hammer 1980). This autogamy is however facultative and may vary in magnitude between species. For instance, *Ae. neglecta* and *Ae. triuncialis*, with their developed stigma and their significant pollen production are more allogamous than *Ae. geniculata*, a poor pollen producer that displays short stigma (Hammer 1980, Arrigo et al. in prep.). These differences corroborate the results obtained in our study.

Finally, wheat assignment probabilities showed that tetraploid wheat, *T. turgidum* (AABB), was the most likely wheat genes donor of both *Ae. neglecta* (UUMM) and *Ae. triuncialis* (UUCC) introgressed samples (Figure 4). In contrast, no clear evidence of introgression involving the hexaploid bread wheat (AABBDD) was observed, except for one *Ae. geniculata* x *T. aestivum* F1 hybrid ($ABDU^{g}M^{g}$ as expected genome formula). In this case, the D genome of bread wheat probably provided many specific AFLP bands and ensured a reliable identification of the paternal parent. Furthermore, homoeologous pairings in *Aegilops* x *T. aestivum* F1 hybrids generally involve wheat chromosomes of

the A and D genomes (Fernández-Calvín and Orellana 1992; Cifuentes et al. 2006). As a result, the introgression of bread wheat genes into the *Aegilops* pool would have been detected at least with AFLP bands specific to the D genome. Our results suggested an enhanced gene transfer between wheat and *Aegilops* species sharing a same ploidy level. Similar results were obtained in other plant species models where aneuploid zygotes were less viable (Sabelli and Larkins 2009). In addition, this hypothesis was partly corroborated by crossing experiments where tetraploid *Aegilops* x *T. turgidum* hybrids showed better germination percentages than pentaploid *Aegilops* x *T. aestivum* hybrids (Nishiyama 1981; Katsiotis et al. 1995).

Conclusion

Overall, our study supported the hypothesis that an active introgression occurred between *T. turgidum* and *Ae. neglecta* or *Ae. triuncialis*. Such extent of introgression was unexpected since *Aegilops* and wheat species were considered as highly autogamous and that F1 hybrids are generally highly sterile. Our study proved that these reproductive barriers can be bypassed since the occurrence of F1 hybrids was probably a long lasting and recurrent phenomenon that took place over large geographical scales. Indeed, wheat cultivations and *Aegilops* species share a long common history as they probably grow sympatrically in Southern Europe since the earliest settlements of agriculture. In addition, our study suggested that cattle and human activities might have promoted the dispersal of introgressed individuals and consequently long distance spread of wheat genes towards areas isolated from cultivations. As a consequence, the commercial release of transgenic wheats in southern Europe should take into account the possible introgression of transgenes within wild *Aegilops* populations. A special attention should be given to *Aegilops* species that are facultative autogamous and to tetraploid wheat species. In the European context, only containment strategies of transgenes based on fitness mitigation might be effective by limiting the survival and the reproduction of a potential transgenic *Aegilops* x *Triticum* descent.

Acknowledgements

The authors are grateful to the team of the Botanical Garden of Neuchâtel for their help in plants cultivation. Marta Cifuentes, Jacques David, Nadir Alvarez, Geert Kleier, Philippe Küpfer, Mélanie Thobor, Anouk Beguin, Celia Bueno, Graham Chapman and Ilhan Kaya provided useful comments and technical assistance. This project was funded by the National Centre of Competence in Research (NCCR) Plant Survival and the PNR-59, two research programs of the Swiss National Science Foundation.

Additional File 1. Distribution of P_{wheat}, for each population included in the analysis. The absolute frequency and the proportion of samples are indicated for seven P_{wheat} intervals (0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5, 0.5-0.6, and 0.6-0.7). Summary statistics are provided at the end of each table. In addition, information regarding populations (country and distance to wheat cultivations; D.wheat) are provided.

				Pwheat (occurrences)								Pwheat (proportion)						
	Country	Population	D.wheat (m)	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	Total	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7
Ae. geniculata	*Spain*	E0609B	1	23							23	100%						
		E0610B	1	15							15	100%						
		E0615A	1	5							5	100%						
		E0618A	1	21						1	22	95%						5%
		E0610E	10	5							5	100%						
		E0616D	20	11							11	100%						
		E0617A	20	14							14	100%						
		E0613A	800	3							3	100%						
		E0610C	2400	25							25	100%						
		E0613B	10400	7							7	100%						
		E0611B	>15 km	14							14	100%						
		E0611D	>15 km	5							5	100%						
	France	F070606D	1	17							17	100%						
		F1207A	50	11							11	100%						
		F1307A	600	12	1						13	92%	8%					
		F070406D	1300	16							16	100%						
	Italy	I0624C	1	11							11	100%						
		I0624A	120	17							17	100%						
		I0623D	400	4							4	100%						
		I0623A	1200	12							12	100%						
			Total	248	1					1	250	mean: 99%	8%					5%
Ae. neglecta	*Croatia*	C0723C	30	4	3		1				8	50%	38%		13%			
	Spain	E0616B	1	4	1	1	1				7	57%	14%	14%	14%			
		E0616A	1	9		2					11	82%		18%				
		E0615C	200	14							14	100%						
		E0613A	800	11	1						12	92%	8%					
	Italy	I0623B	1	6		1	2				9	67%		11%	22%			
		I0624B	1	15							15	100%						
		I0624C	1	1							1	100%						
		I0623D	400	1							1	100%						
		Orb1	>15 km	12							12	100%						
			Total	77	5	4	4				90	mean: 85%	20%	15%	16%			
Ae. triuncialis	*Croatia*	C0724A	10	6	1	2	1				10	60%	10%	20%	10%			
		C0723C	30	8	1						9	89%	11%					
	Spain	E0615A	1	6	2	3	1	1			13	46%	15%	23%	8%	8%		
		E0609B	1	3				1			4	75%				25%		
		E0618A	1	7		1	1				9	78%		11%	11%			
		E0610B	1	5	1						6	83%	17%					
		E0610E	10	7	1	1					9	78%	11%	11%				
		E0612A	10	6	1						7	86%	14%					
		E0616D	20	6	1						7	86%	14%					
		E0615C	200	7							7	100%						
		E0613A	800	5	1	2					8	63%	13%	25%				
		E0612B	2800	14							14	100%						
		E0611B	>15 km	7							7	100%						
	Italy	I0626B	100	12							12	100%						
	Turkey	Turk6	>15 km	11							11	100%						
			Total	110	9	9	3	2			133	mean: 83%	13%	18%	10%	16%		
			Grand total	**435**	**15**	**13**	**7**	**2**	**0**	**1**	**473**							

References

Arrigo N, Tuszynski JW, Ehrich D, Gerdes T, Alvarez N (2009) Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. BMC Bioinformatics 10.

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Molecular Ecology 16: 3737-3758.

Bonin A, Bellemain E, Bronken Edeisen P, Pompanon F, Brochmann C et al. (2004) How to track and assess genotyping errors in population genetics studies. Molecular Ecology 13: 3261-3273.

Buerkle CA (2005) Maximum-likelihood estimation of a hybrid index based on molecular markers. Mol Ecol Notes 5: 684-687.

Chee PW, Lavin M, Talbert LE (1995) Molecular analysis of evolutionary patterns in U genome wild wheats. Genome 38: 290-297.

Chen DH, Ronald PC (1999) A rapid DNA minipreparation method suitable for AFLP and other PCR applications. Plant Mol Biol Rep 17(1): 53-57.

Cifuentes M, Blein M, Benavente E (2006) A cytomolecular approach to assess the potential of gene transfer from a crop (*Triticum turgidum* L.) to a wide relative (*Aegilops geniculata* Roth.). Theoretical And Applied Genetics 112: 657-664.

David JL, Benavente E, Bres-Patry C, Dusautoir JC, Echaide M (2004) Are neopolyploids a likely route for a transgene walk to the wild? The *Aegilops ovata* x *Triticum turgidum durum* case. Biol J Linnean Soc 82(4): 503-510.

De Vries AP (1971) Flowering biology of wheat, particularly in view of hybrid seed production - a review. Euphytica 20(2): 152-170.

Doebley J (1990) Molecular evidence for gene flow among *Zea* species. BioScience 40(6): 443-448.

Fernández-Calvín B, Orellana J (1992) Relationship between pairing frequencies and genome affinity estimations in *Aegiops ovata* x *Triticum aestivum* hybrid plants. Heredity 68: 165-172.

Guadagnuolo R, Savova-Bianchi D, Felber F (2001) Gene flow from wheat (*Triticum aestivum* L.) to jointed goatgrass (*Aegilops cylindrica* Host.), as revealed by RAPD and microsatellite markers. Theoretical And Applied Genetics 103(1): 1-8.

Gugerli F, Englisch T, Niklfeld H, Tribsch A, Mirek Z et al. (2008) Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation - a project synopsis. Perspect Plant Ecol 10(4): 259-281.

Hammer K (1980) Vorarbeiten zur Monographischen Darstellung von Wildpflanzensortimenten: *Aegilops* L. Kulturpflanze 28: 33-180.

Hanson BD, Mallory-Smith CA, Shafii B, Thill DC, Zemetra R (2005) Pollen-mediated gene flow from blue aleurone wheat to other wheat cultivars. Crop Science 45: 1610-1617.

Karagöz A (2006) Hybridization in Turkish *Aegilops* L. species. Pakistan Journal of Biological Science 9(12): 2243-2248.

Katsiotis A, Hanneman RE, Forsberg RA (1995) Endosperm Balance Number and the polar-nuclei activation hypotheses for endosperm development in interspecific crosses of *Solanaceae* and *Gramineae*, respectively. Theoretical And Applied Genetics 91: 848-855.

Loureiro I, Escorial MC, Garcia-Baudin JM, Chueca MC (2006) Evidence of natural hybridization between *Aegilops geniculata* and wheat under field conditions in central Spain. Environ Biosafety Res 5: 105-109.

Loureiro I, Escorial MC, Garcia-Baudin JM, Chueca MC (2007a) Hybridization between wheat (*Triticum aestivum*) and the wild species *Aegilops geniculata* and *A. biuncialis* under experimental field conditions. Agric Ecosyst Environ 120(2-4): 384-390.

Loureiro I, Escorial MC, Garcia-Baudin JM, Chueca MC (2008) Hybridization, fertility and herbicide resistance of hybrids between wheat and *Ae. biuncialis*. Agron Sustain Dev.

Loureiro I, Escorial MC, Gonzáles-Andujar JL, Garcia-Baudin JM, Chueca MC (2007b) Wheat pollen dispersal under semiarid field conditions: potential outcrossing with *Triticum aestivum* and *Triticum turgidum*. Euphytica 156: 25-37.

Matus-Cadiz MA, Hucl P, Dupuis B (2007) Pollen-mediated gene flow in wheat at the commercial scale. Crop Science 47(2): 573-581.

Matus-Cadiz MA, Hucl P, Horak MJ, Blomquist LK (2004) Gene flow in wheat at the field scale. Crop Science 44: 718-727.

Meimberg H, Hammond JI, Jorgensen CM, Park TW, Gerlach JD et al. (2006) Molecular evidence for an extreme genetic bottleneck during introduction of an invading grass to California. Biological Invasions 8(6): 1355-1366.

Monte JV, De Nova PJG, Soler C (2001) AFLP-based analysis to study genetic variability and relationships in the Spanish species of the genus *Aegilops*. Hereditas 135(2-3): 233-238.

Morrison LA, Riera-Liazrazu O, Cremieux L, Mallory-Smith CA (2002) Jointed goatgrass (*Aegilops cylindrica* host) x wheat (*Triticum aestivum* L.) hybrids: Hybridization dynamics in oregon wheat fields. Crop Science 42(6): 1863-1872.

Nishiyama I (1981) *Aegilops-Triticum* hybridization in view of PNA hypothesis. Wheat Inf Serv 52: 39-51.

O'Hanlon PC, Peakall R, Briese DT (1999) Amplified fragment length polymorphism (AFLP) reveals introgression in weedy *Onopordum* thistles: hybridization and invasion. Molecular Ecology 8: 1239-1246.

Pester TA, Ward SM, Fenwick AL, Westra P, Nissen SJ (2003) Genetic diversity of jointed goatgrass (*Aegilops cylindrica*) determined with RAPD and AFLP markers. Weed Science 51(3): 287-293.

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

Sabelli PA, Larkins BA (2009) The development of endosperm in grasses. Plant Physiology 149: 14-26.

Schneider A, Molnar I, Molnár-Láng MM (2008) Utilisation of *Aegilops* (goatgrass) species to widen the genetic diversity of cultivated wheat. Euphytica 163(1): 1-19.

Schoenenberger N, Felber F, Savova-Bianchi D, Guadagnuolo R (2005) Introgression of wheat DNA markers from A, B and D genomes in early generation progeny of *Aegilops cylindrica* Host x *Triticum aestivum* L. hybrids. Theoretical And Applied Genetics 111(7): 1338-1346.

Schoenenberger N, Guadagnuolo R, Savova-Bianchi D, Kupfer P, Felber F (2006) Molecular analysis, cytogenetics and fertility of introgression lines from transgenic wheat to *Aegilops cylindrica* host. Genetics 174(4): 2061-2070.

Sears ER (1976) Genetic-control of chromosome-pairing in wheat. Annual Review Of Genetics 10: 31-51.

van Slageren MW (1994) Wild wheats: A monograph of *Aegilops* L.and *Ambylopyrum* (Jaub and Spach) Eig (*Poaceae*). Wageningen: Agricultural Univ., Wageningen, The Netherlands and Int. Cent. for Agric. Res. in Dry Areas, Aleppo, Syria.

Vekemans X, Beauwens T, Lemaire M, Roldan-Ruiz I (2002) Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Molecular Ecology 11: 139-151.

Villegas Molina F, Ortega Alba F, Sánchez del Árbol MA (1990) Le secteur méridional de la haute Andalousie. Déprise agricole et autres changements visibles. Géographie d'une Espagne en mutation Prospections aériennes II. Madrid: Casa de Velásquez: Ciudad Universitaria. pp. 95-114.

Waines JG, Barnhart D (1992) Biosystematic research in *Aegilops* and *Triticum*. Hereditas 116(3): 207-212.

Wang ZN, Hang A, Hansen J, Burton C, Mallory-Smith CA et al. (2000) Visualization of A- and B-genome chromosomes in wheat (*Triticum aestivum* L.) x jointed goatgrass (*Aegilops cylindrica* Host) backcross progenies. Genome 43(6): 1038-1044.

Watanabe N, Kawahara T (1999) *Aegilops* species collected in California and Oregon, USA. Wheat Inf Serv 89: 33-36.

Weissmann S, Feldman M, Gressel J (2005) Sequence evidence for sporadic intergeneric DNA introgression from wheat to a wild *Aegilops* species. Molecular Biology And Evolution 22(10): 2055-2062.

Whitlock R, Hipperson H, Mannarelli M, Butlin R, Burke T (2008) An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error. Mol Ecol Resour 8: 725-735.

Zaharieva M, Monneveux P (2006) Spontaneous hybridization between bread wheat (*Triticum aestivum* L.) and its wild relatives in Europe. Crop Science 46(2): 512-527.

Chapter 5

Ancient and recent evolutionary history of *Aegilops geniculata* Roth, a Mediterranean wild relative of the wheat

Arrigo N[1], Felber F[1], Buerki S[1], Alvarez N[1], David J[2], and Guadagnuolo R[1]

[1]Laboratory of Evolutionary Botany, Institute of Biology, University of Neuchâtel, 11 rue Emile-Argand, 2000 Neuchâtel, Switzerland

[2]UMR Diversité et Génomes des Plantes Cultivées, INRA Domaine de Melgueil, 34130, Mauguio, France

Abstract

The present study combined AFLP genotyping and sequencing of chloroplast markers (cpDNA) to address the phylogeography of *Aegilops geniculata* Roth, a widespread Mediterranean wild relative of wheat. A total of 438 samples, belonging to 143 populations collected around the Mediterranean basin were included in the study. Globally, the polymorphism detected with AFLPs was congruent with that of cpDNA markers and revealed a basal dichotomy in the genetic diversity of *Ae. geniculata*. The species was organised in two main phylogeograpic clusters; populations collected in Northern Mediterranean areas (i.e. from Portugal to the Balkans) were genetically distinct from those occurring in Southern Mediterranean locations (i.e. from Morocco to Palestine and Bulgaria). Populations from Northern Mediterranean areas appeared to be less diversified but more structured genetically than those collected in Southern Mediterranean areas. For instance, I. A significant correlation existed between the genetic and the geographic distances between samples in populations collected in the Iberian and the Franco-Italian Peninsulas. II. Samples collected in these two regions belong to distinct genetic entities. III. Additional genetic structures were detected in Spain, France and Italy. In contrast, samples collected in Southern Mediterranean areas did not show comparable genetic patterns. These results are discussed according to hypotheses that consider either I. a natural expansion or II. an agriculture-mediated colonization of the Mediterranean basin by *Ae. geniculata*.

Finally, our study revealed recent migration events with samples collected in Northern Mediterranean areas that showed genotypes specific to Southern Mediterranean areas. It appears that several of these migrants showed evidence of outcrossing with the local populations, a result that contrasted with the commonly assumed autogamy of *Ae. geniculata*.

Introduction

A large of number of studies investigated the distribution of glacial refugia and the colonization routes followed by European and Mediterranean species during post-glaciary times (e.g. Hewitt (2004a) or Schmitt (2007)). Among them, several studies focused on plants of agronomic importance and on associated species from agroecosystems (Balfourier et al. 2000; Sharbel et al. 2000; Besnard et al. 2002; Bittkau and Comes 2005; Fjellheim et al. 2006; Francois et al. 2008). It appeared that colonization patterns followed by these species were generally blurred by ancient and recent human activities that complicated the reconstruction of their expansion history (Fineschi et al. 2000; Parducci et al. 2001; Conedera et al. 2004; de Casas et al. 2006).

The present study investigates *Aegilops geniculata* Roth, an allotetraploid species (2n = 4x = 28 chromosomes, with genome formula MMUU), whose diploid ancestors are assumed to be: *Ae. comosa* Sm. In Sibth. & Sm. var. comosa (genome MM) and *Ae. umbellulata* Zhuk. (genome UU) (van Slageren 1994; Friebe et al. 1999). *Aegilops geniculata* is a wild relative of wheat (*Triticum* L.) and represents a source of agronomical traits (e.g. resistances to fungi or nematodes) in wheat improvement programs (Zaharieva et al. 2001b; Zaharieva et al. 2001a; Zaharieva et al. 2003). *Aegilops geniculata* is a pioneer Mediterranean species that naturally occurs in agricultural landscapes (e.g. in field edges, along roadsides, in pastures or in disturbed areas) of Southern Europe, North Africa and western Asia (van Slageren 1994). Recent adventive populations also occur in the Canary Islands, the United States and the North-West of Europe (van Slageren 1994). As a result, *Ae. geniculata* represents one of the most widely distributed species of the genus. The dispersal strategy of *Ae. geniculata* relies on four main biological features. I. This annual species grows in tufts and produces a large number of seeds; Zaharieva and Monneveux (2006) report a production of 1400 seeds per plant in the Mediterranean area. II. The production of seeds is generally guaranteed as *Ae. geniculata* is partly autogamous with anthers releasing the pollen before the opening of the spikelets (Hammer 1980). III. The spikes do not disarticulate at maturity and possess numerous hairy awns that ensure an efficient zoochorous dispersal. This allows the settlement of several siblings (five to six seeds per spike) in newly colonized areas. IV. The presence of awns and appressed hairs on the glumes help spikes to sink into the soil in rainy conditions (David et al, personal communication). Spikes remain intact and provide an efficient soil seed bank by protecting seeds for at least two years. For instance, germinated plants emerging from two years old buried spikes were observed (Arrigo et al, unpublished data). Long range seed dispersal is mostly linked to human activities, such as pastoralism of goats and ovine livestock, the cereal cultivation or even the tires of agricultural engines (Zaharieva et al. 2001a). Local dispersal at a much smaller scale is dependent on gravity and zoochory, as ants that store spikes as food resources (Hensen 2002).

The present study combines AFLP genotyping and chloroplast (hereafter Cp) DNA sequencing, an approach already successfully applied in several phylogeography studies (e.g. Eidesen et al. (2007) or Paun et al. (2008)). AFLPs represent a valuable tool to unravel biogeographic patterns, highlight migrant individuals and detect gene flow (Viard et al. 2004; Eidesen et al. 2007). They were extensively applied in the *Aegilops-Triticum* species complex (Monte et al. 2001; Pester et al. 2003; Sasanuma et al. 2004; Medini et al. 2005; Meimberg et al. 2006; Naghavi et al. 2007). At a different scale, Cp DNA sequencing provides complementary information regarding intra- and inter-specific relationships. For instance,

intergenic spacers (hereafter IGS) located in the small sub-unit of the chloroplast (i.e. the *ndhF-rpl32* and *rpl32-trnL* IGSs) were shown to be highly variable at the intrageneric level (Shaw et al. 2005; 2007). Chloroplast genomes were extensively investigated in *Aegilops* species (Provan et al. 2004; Sasanuma et al. 2004; Guo and Terachi 2005; Yamane and Kawahara 2005; Meimberg et al. 2006; Haider and Nabulsi 2008; Matsuoka et al. 2008).

The present study aims to investigate:

I. The ancient evolutionary history of *Ae. geniculata*. We try to identify the geographical origin and the colonization routes followed by *Ae. geniculata* during its expansion in the Mediterranean basin. In this framework, we propose several scenarios to explain the phylogeographic structures that are observed.

II. The recent history of *Ae. geniculata.* We outline the occurrence of migration and gene flow and discuss the influences of human activities in shaping the present genetic diversity of *Ae. geniculata*.

Material and Methods

Sampled area and plant material

The sampling includes collection sites from the whole Mediterranean basin (Figure 1), where six geographic areas were defined on the basis of natural frontiers (Pyrenees, Alps, Mediterranean Sea, Taurus Mounts) or disjunctions in the sampling area (e.g. Egypt). These areas represent comparable surfaces (i.e. most of the areas are 1000 km long, except for the Balkans) and were designated as: the Iberian Peninsula (IBP), the Franco-Italian Peninsula (FIP), the Balkans (HR), the Bosporus Strait (BST), the Levantine (LEV) and finally the Northern Africa (NAF).

A total of 411 *Ae. geniculata* samples, belonging to 146 populations were included in the present study. Half of the samples were collected as seeds in Spain, France, Italy, Croatia and Turkey (194 samples, 22 populations, with nine samples per population in average), cultivated at the Botanical Garden of the city and University of Neuchâtel and provided a large collection of silica-dried leaves. This sampling was completed with DNAs provided by the INRA of Montpellier (217 DNA samples from 121 populations, with two samples per population in average). Finally, samples from other *Aegilops* and *Triticum* species were used as outgroups: *Ae. triuncialis* L. individuals (27 samples, eight populations), *Ae. neglecta* Req. ex Bertol. (four samples, one population), *Ae. biuncialis* Vis. (one sample), *Triticum aestivum* L. (one sample) and *Triticum turgidum* L. (one sample). Outgroup samples were collected as seeds in the complete sampling area and cultivated at the Botanical Garden of the city and University of Neuchâtel (see below). Refer to Additional file 1 for a complete description of the AFLP and Cp datasets.

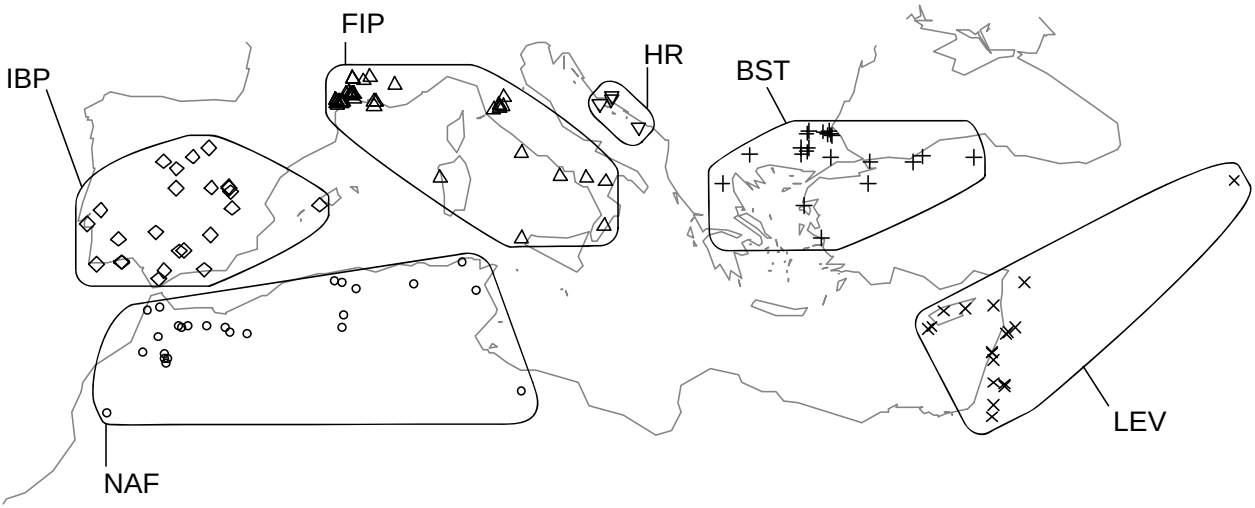

Figure 1. Sampling area. Each population is labelled according to its geographic area (IBP – Iberian Peninsula, FIP – Franco-Italian Peninsula, HR – Balkans, BST – Bosporus Strait, LEV – Levantine, NAF – Northern Africa).

Molecular analyses

AFLP amplification conditions and scoring

DNAs were extracted from 10 mg of silica-dried leaves by using a CTAB protocol (Chen and Ronald 1999). DNA concentrations were standardized at 10 ng/ l. The AFLP analysis was achieved using the complete set of *Ae. geniculata* samples and completed with 27 *Ae. triuncialis* outgroup samples. We followed the AFLP protocol developed by Gugerli et al. (2008). Reactions were conducted with a PTC-100 thermocycler (MJ Research), using 96-wells plates where samples were randomly distributed. We used two PCR selective primer pairs (EACT-MCTG and EAGT-MCAT). Primers and adaptors sequences are indicated in Table 1. PCR products were recorded using the GeneScan technology (primers flanking *Eco-RI* restriction site were fluorescently labelled with 5' 6-carboxy-fluorescein [FAM]) with a capillary sequencer (ABI 3730XL, service provided by Macrogen Inc. Seoul, Korea). Raw electropherograms were analysed with PeakScanner (peak detection parameters: default parameters except a light peak smoothing) in order to detect and calculate the size of AFLP bands. The scoring (consisting in recording the presence / absence of AFLP bands for each sample) was performed using an automated scoring R CRAN package (RawGeno (Arrigo et al. 2009)). The library was settled as follows: scoring range = 100 – 400 bp, minimum intensity = 100 rfu, minimum bin width = 0 bp, maximum bin width = 2 bp and closely sized bins [5%] were removed). Finally, the scored matrices of both primer pairs were concatenated in a single binary matrix where individuals and bands were stored as lines and columns respectively.

We replicated several AFLP reactions: five samples and one blank were distributed on each 96-wells plate to check the between plate variability and the purity of templates. Additionally, 11 samples (chosen randomly from each plate and representing 14% of the final dataset) were replicated on another plate in order to calculate the error rate (Bonin et al. 2004) and subsequently eliminate non-reproducible bands. We discarded bands that were not reproducible at a 95% threshold (representing 17% and 15% of the total number of detected bands for EACT-MCTG and EAGT-MCAT respectively).

Diversity patterns

The genetic diversity patterns of *Ae. geniculata* were investigated using two statistics:

I. The percentage of different genotypes. AFLPdat (Ehrich 2006) was used to compute the percentage of different genotypes ($P_{DIffGen}$) considering a 100 Km grid that covered the whole sampling area. The AFLPdat analysis included samples that were located within a 150 Km perimeter of each point of the grid. Genotypes were considered as non-identical if they differed fom each other with more than 5.8% of mismatching AFLP bands. Finally, in order to provide unbiased $P_{DiffGen}$ estimates due to unequal sampling effort between the different sampling areas, computations were bootstrapped by resampling 1000 times a constant number of samples per grid point (10 samples). The average percentage of different genotypes was finally computed for each sampling area (i.e. IBP, FIP, HR, BST, LEV and NAF).

II. The correlation between the genetic and the geographic distances was calculated within each sampling area. R CRAN was used to calculate genetic distances between samples by applying the Jaccard distance on the AFLP binary matrix (using the "vegan" R package).

Geographic distances were calculated between individuals by computing the Euclidean distance on projected geographical coordinates (UTM spherical projection centred on the sampling area, kilometric scale, using the "GEOmap" R package). The correlations between the geographic and the genetic distance matrices were tested using Mantel tests (Pearson's correlation, 1000 permutations).

STRUCTURE analysis

We investigated the biogeographic structure of *Ae. geniculata* using Bayesian Inference methods (STRUCTURE 2.2, (Pritchard et al. 2000), version adapted for dominant datasets) to assign individuals within a user-defined number (hereafter K) of genetic groups defined in a way to optimize the Hardy-Weinberg equilibrium within the K groups. The algorithm was not sensitive to the sampling scheme and groups can include a different number of individuals. We assumed the non-admixture of populations (since we worked at a large scale and were not interested in detecting potential hybrids between groups) with 100000 steps of burn-in and 400000 steps for data acquisition. The number of groups was tested for values ranging between one and eleven. A total of five replicates were performed for each value of K and only runs that obtained the highest maximum-likelihood values were considered for further analyses. The STRUCTURE analysis provided the best results when nine groups were considered, as the maximum-likelihood of runs followed an asymptotic curve and stabilized when reaching K=9. This value was used as further K values only provided local sub-groups. One of these groups included the *Ae. triuncialis* samples that were used as outgroups and was subsequently removed from the analysis. Individuals were assigned to the eight remaining AFLP groups, if their respective assignment probability was higher than 0.75. The consistency of STRUCTURE analyses was checked with a principal coordinates analysis, computed on a Jaccard distance matrix (see above). Finally, results of the STRUCTURE analysis were displayed on a map where populations were symbolized as pie-charts representing the number of samples assigned to each AFLP group.

Chloroplast dataset

Amplification conditions

We sequenced a subset of 125 individuals (Additional File 1). This subset included one to three *Ae. geniculata* sample per population and covered the whole sampling area. The sampling was completed with six outgroup species (*Ae. biuncialis*, *Ae. neglecta*, *Ae. triuncialis*, *Triticum aestivum* and *T. turgidum*). We investigated the small sub-unit of the chloroplast, an informative region in *Aegilops* species (Meimberg et al. 2006) composed by three molecular markers: the *ndhF-rpl32* intergenic spacer (IGS), the *rpl32* gene and the *rpl32-trnL* IGS. We amplified the whole region using the trnL(UAG) and ndhF primers (Shaw et al. 2005; 2007) and sequenced it using two flanking internal primers: Intern-F and Intern-R (Table 1). DNA amplifications were performed in a 40 l reaction volume, with 30 ng of genomic DNA, 1x PCR buffer, 0.2 mg/ml BSA, 1.5 M $MgCl_2$, 2 U Taq polymerase (Promega), 200 M dNTPs, and 0.2 M primers. We used the "rbcl" program proposed by Shaw et al. (2005; 2007) with a PTC-100 thermocycler (MJ Research). The sequencing was performed on an ABI 3730XL (service provided by Macrogen Inc. Seoul, Korea) and the base calling was checked using ChromasPro (version 1.34, Technelysium Ltd.,

Helensvale, Queensland, Australia). GenBank accession numbers are FGxxx to FGyyy. The alignment of sequences was achieved using initially ClustalX (Thompson et al. 1997) and adjusted manually using the similarity criterion (Morrison 2006) with the program BioEdit (Hall 1999).

Table 1. Primer sequences. A. The AFLP analysis needed the use of AE and AM adaptators during the ligation step (AE is built by annealing AET with AEB while AMT and AMB are used to build AM). The preselective and selective primers are indicated. Selective primers flanking the *Eco-RI* restriction site were labelled fluorescently using 5' 6-carboxy-fluorescein (FAM). B. The *ndhF-rpl32* intergenic spacer (IGS), the *rpl32* gene and the *rpl32-trnL* IGS are contiguous on the chloroplast genome and were amplified with a long range PCR using standard primers: trnL(UAG) and ndhF (Shaw et al. 2005; 2007). The sequencing needed two internal primers that flanked the region of interest (Intern-F and Intern-R).

Analysis	Primer name	Sequence (5'-3')
A. AFLP		
AFLP Ligation	Adaptator AET	CTCGTAGACTGCGTACC
AFLP Ligation	Adaptator AEB	AATTGGTACGCAGTCTAC
AFLP Ligation	Adaptator AMT	GACGATGAGTCCTGAG
AFLP Ligation	Adaptator AMB	TACTCAGGACTCAT
AFLP Pre-Selective PCR	Primer EcoRI-A	GACTGCGTACCAATTCA
AFLP Pre-Selective PCR	Primer MseI-C	GATGAGTCCTGAGTAAC
AFLP Selective PCR 1	Primer EcoRI-AGT	5'FAM-GACTGCGTACCAATTCAGT
AFLP Selective PCR 1	Primer MseI-CAT	GATGAGTCCTGAGTAACAT
AFLP Selective PCR 2	Primer EcoRI-ACT	5'FAM-GACTGCGTACCAATTCACT
AFLP Selective PCR 2	Primer MseI-CTG	GATGAGTCCTGAGTAACTG
B. Chloroplast		
Cp Long range PCR	trnL(UAG)	CTGCTTCCTAAGAGCAGCGT
Cp Long range PCR	ndhF	GAAAGGTATKATCCAYGMATATT
Cp Sequencing intern	Intern-F	CCAATTCTTATCTCTTTCTGAAAG
Cp Sequencing intern	Intern-R	GCTTTGCCCAATAGAAACACA

Haplotypes tree

Since we worked at the intra-specific level, the dataset included a limited amount of phylogenetic information and we assumed that the three molecular markers reflected a congruent evolutive history. We thus performed total evidence trees (Kluge 1989). Phylogenetic analyses and their corresponding boostrap analyses were performed using the Maximum Likelihood (ML) and the Maximum Parsimony (MP) criteria:

- ML analysis. Model selection for each partition was assessed using Modeltest version 3.7 (Posada and Crandall 1998) and the Akaike information criterion (Akaike 1973). ML analyses were performed using RAxML version 7.0.0 (Stamatakis 2006; Stamatakis et al. 2008) with a 1000 rapid bootstrap analyses followed by the search of the best-scoring ML tree in one single run. This analysis was done using the facilities offered by the CIPRES portal in San-Diego, U.S.A. (http://8ball.sdsc.edu:8888/cipres-web/home, accessed the 13th of March 2009).
- MP analysis. The combined dataset was analyzed using parsimony ratchet (Nixon 1999) as implemented in PAUPrat (Sikes and Lewis 2001). Based on recommendations by Nixon (1999), ten independent searches were performed with 200 iterations and 15% of the parsimony

informative characters perturbed. To assess the support at each node, non parametric bootstrap analyses (Felsenstein 1985) were performed using PAUP* version 4.0b10 (Swofford 2002) with 1000 replicates, TBR branch swapping, simple sequence addition, MULTREES and holding 10 trees per replicate.

The congruence of topological groupings in analyses obtained from different algorithms was evaluated using TreeJuxtaposer (Munzner et al. 2003) and bootstrap supports of each main clade were compared.

Results

AFLP data

The AFLP analysis produced a total of 649 bands (352 and 297 bands for EACT-MCTG and EAGT-MCAT respectively), with an average of 170 bands per individual. AFLP bands with small sizes (i.e. < 150 bp) are as frequent as larger bands. The measured error rate of non-matching bands per sample-replicate pair is 5.8%. The use of different DNA sources does not bias the analysis since genotypes are discriminated according to their geographical location and not to their DNA source.

Genetic diversity patterns, based on geographic areas

Patterns of genetic diversity highlight less diversity in Northern Mediterranean areas than in the Southern Mediterranean ones (Figure 2 A and B). The Iberian Peninsula (IBP), the Franco-Italian Peninsula (FIP) and the Balkans (HR) possess the lowest diversities, with averages of 27% to 31% of different genotypes observed in each grid cell. In contrast, the Bosporus Strait (BST), the Levantine (LEV) and the Northern Africa (NAF) are more diversified with 41% to 64% of different genotypes observed on average per grid cell. The Levantine and the Northern Africa are the most diversified areas. Interestingly, a significant correlation exists between the geographic and the genetic distances (Figure 2 C) in the Northern Mediterranean areas (IBP = 0.53***, FIP = 0.29***), while no correlation is outlined in the Southern Mediterranean ones (LEV = 0.08 NS and NAF = 0.11 NS), except for the BST area which is significantly structured (BST = 0.22**).

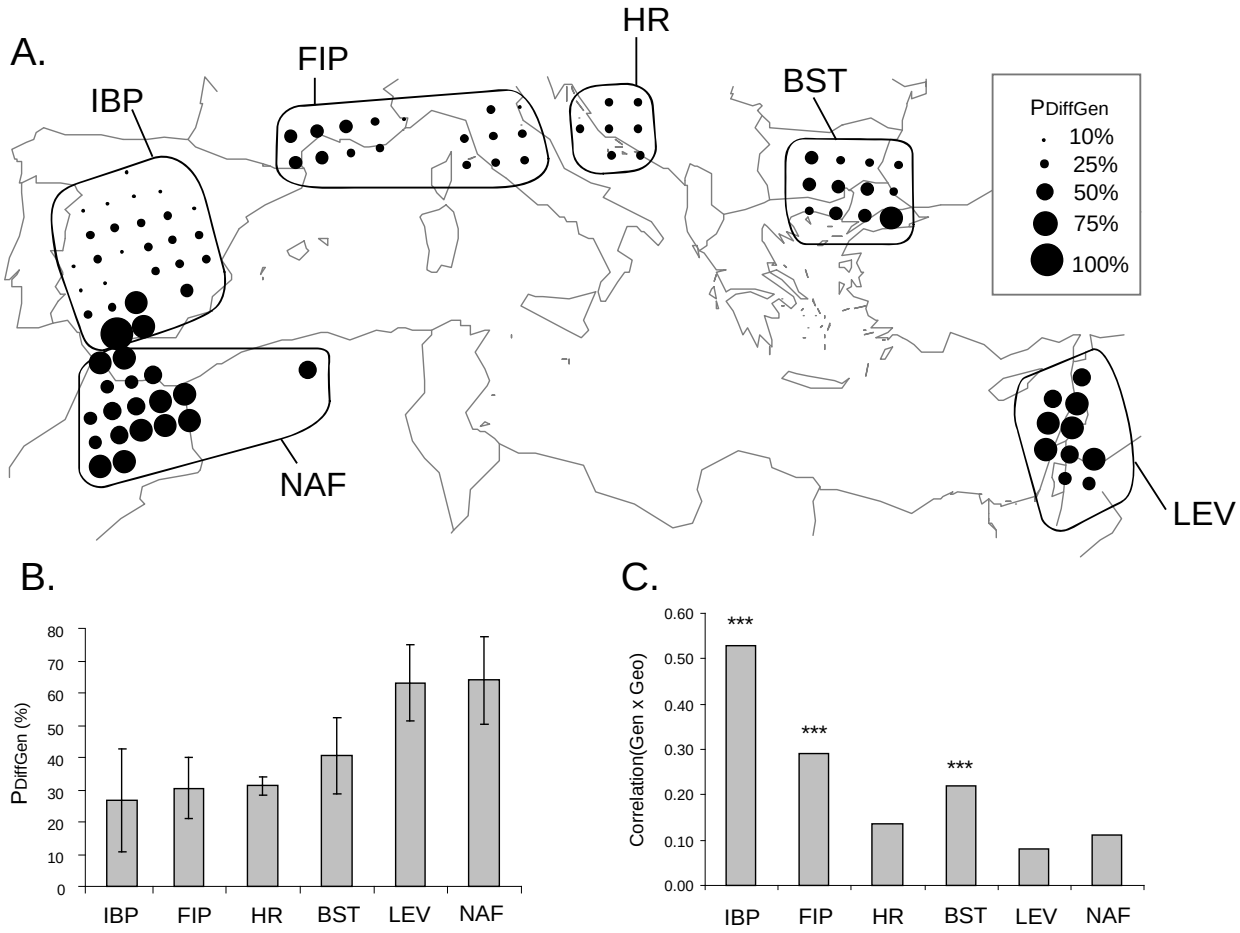

Figure 2. Genetic diversity patterns. A. Percentage of different genotypes ($P_{DiffGen}$), considering a 100 Km grid on the whole sampling area. Calculations involve samples that are included in a 150 Km perimeter around each grid point. Diversity measures were bootstrapped 1000 times including a constant number of samples per grid point. B. Summary statistics for each sampling area. The average percentage of different genotypes ($P_{DiffGen}$) per grid points are indicated for the Iberian Peninsula (IBP), the Franco-Italian Peninsula (FIP), the Balkans (HR), the Bosporus Strait (BST), the Levantine (LEV) or the Northern Africa (NAF). Error bars represent standard deviations. C. Pearson's correlations between the inter-sample genetic and geographic distances, computed with subsets of samples from each sampling area. The significance levels are indicated (Mantel test, 1000 permutations, *** Pvalue < 0.001).

STRUCTURE analysis

The eight AFLP groups are coherently distributed in the sampling area (Figure 3). Populations generally contain only one genotype, but exceptions are found at contact zones between AFLP groups. The groups *A* and *B* occur in the Iberian Peninsula: *A* is largely spread in Spain and Portugal while *B* is observed in two neighbouring Northern Spanish populations (E0617A and E0618A). The groups *C*, *D* and *E* are observed in the Franco-Italian Peninsula. The group *C* is largely spread in France and Italy, *D* is local and includes only one population from southern France (F070406D). The group *E* is represented in two populations from Northern Italy (I0623A and I0624A). The group *F* has probably its origin in the Balkans. The group *G* is observed close to the Bosphorus Strait, in Bulgaria, Greece and Turkey. Finally, the group *H* is widely spread in North Africa and ranges from Morocco to Levantine.

Several discrepancies between the AFLP groups and their geographic distribution are however observed. First, the group *F* occurs in the Balkans, Spain, and North Africa (populations C0723A and C0723C, E0615A, Ifrane and Tlemcen). Second, many individuals (with a total of 27 cases) located in Southern Italy, the Iberian Peninsula as well as the majority of Mediterranean Islands clearly belong to the *H* group and are designated as H_{Im} in further analyses.

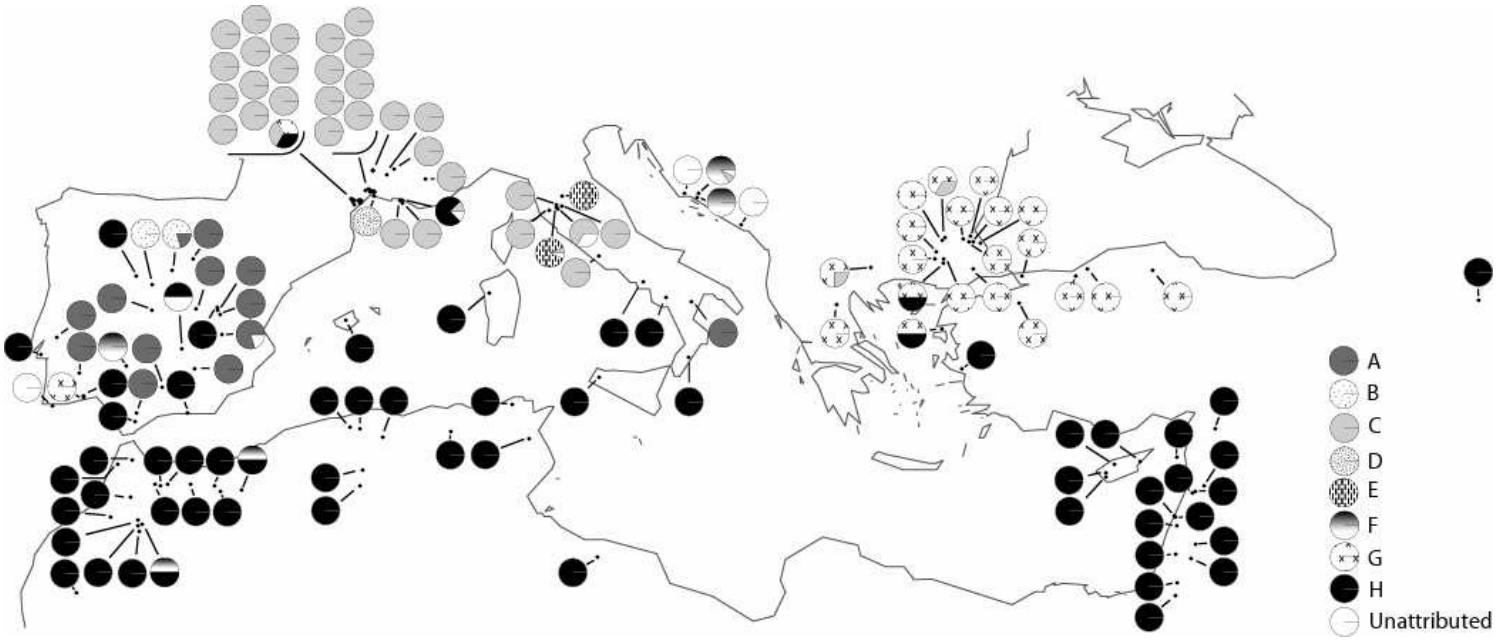

Figure 3. Geographic distribution of the AFLP groups detected in *Ae. geniculata* (*A* – Iberian Peninsula, *B* – Spanish local group, *C* – Franco-Italian Peninsula, *D* - French local group, *E* – Italian local group, *F* – Balkans, *G* – Bosporus Strait and *H* – Northern Africa and Levantine). A total of 143 populations are displayed as pie-charts that represent the proportion of individuals belonging to each of the eight AFLP groups detected by STRUCTURE.

Genetic structure, as observed from the AFLP dataset, has a strong geographical component (Figure 4). First, the main genetic gradient revealed by the first axis of the principal coordinates analysis, corresponds to an East-West gradient and ranges from an Italian group (*E*) to samples located in the Bosphorus Strait (*G* group). Second, three main genetic entities can be defined when considering genetic similarities between the eight AFLP groups. These three entities coherently reflect phylogeographic patterns:

I. The groups *A* and *B* represent the first entity and occur in the Iberian Peninsula.
II. The groups *C*, *D* and *E* belong to the second genetic entity and occur in the Franco-Italian peninsula.
III. The groups *G* and *H* represent a third entity located in the Southern Mediterranean areas.

Finally, the *F* group has an intermediate position between these three main entities and corresponds to populations from different areas (Balkans, Spain and North Africa).

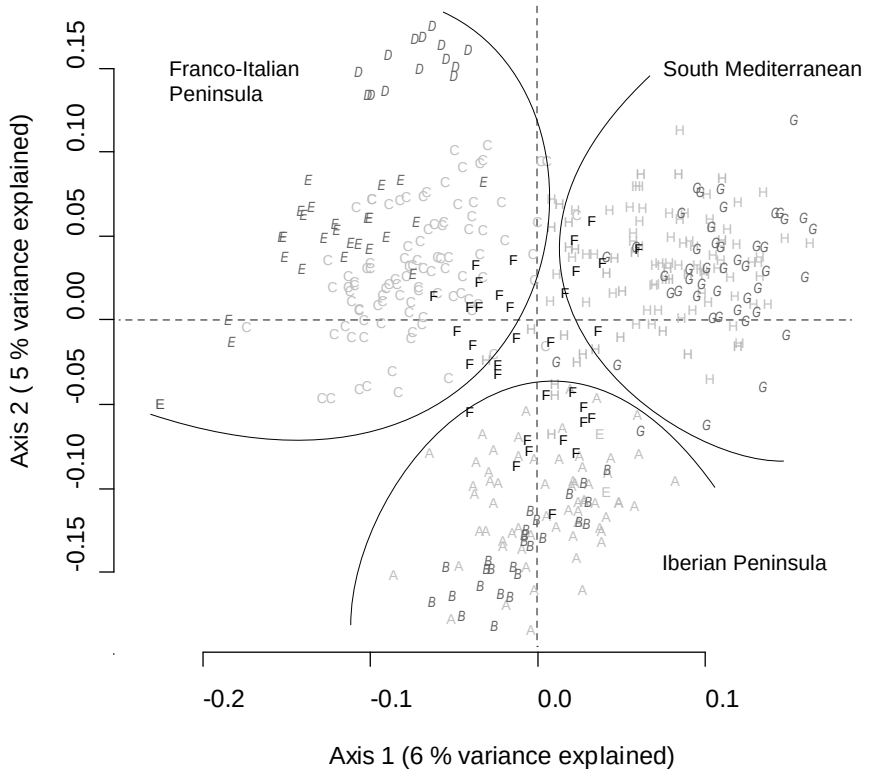

Figure 4. Principal coordinates analysis of the AFLP dataset (calculated on a Jaccard distance). Individuals are labelled according to AFLP groups detected by STRUCTURE (*A* – Iberian Peninsula, *B* – Spanish local group, *C* – Franco-Italian Peninsula, *D* - French local group, *E* – Italian local group, *F* – Balkans, *G* – Bosporus Strait and *H* – Northern Africa and Levantine). The congruence between the AFLP dataset and the geographical distribution of samples are outlined as three main phylogeographical entities: the Iberian Peninsula, the Franco-Italian Peninsula and the Southern Mediterranean. Local AFLP groups (*B, D*, *E*, and *G*) are displayed in italics. The percentage of explained variance is indicated for each axis.

Chloroplast DNA data

Ae. geniculata *haplotypes*

The aligned matrix is 1483 bp long (522 bp for the *rpl32-trnL* intergenic spacer, 192 bp for the *rpl32* gene and 769 bp for the *ndhF-rpl32* intergenic spacer), with a total of 89 variable characters with 70 potentially parsimony informative (hereafter PPI) characters (i.e. 4.7 % of the total number of characters) when the outgroups are included. The *Ae. geniculata* samples include 27 variable characters with 15 PPI characters (i.e. 1 % of the total number of characters). Most of the variability within *Ae. geniculata* occurs in the *rpl32-trnl* intergenic spacer (15 PPI characters out of 20 variable characters) while the *ndhF-rpl32* intergenic spacer (11 PPI out of 19 variable characters) mainly contributes to identify the outgroups. Finally, the *rpl32* gene shows almost no variability (one PPI out of three variable characters).

Maximum parsimony analysis results in 1038 trees having 55 steps and a CI and RI of respectively 0.763 and 0.959. Under the ML criterion, the best-fit model for the combined matrices is the general time reversible model with an alpha parameter for the shape of the gamma distribution to account for among-site rate heterogeneity (Yang, 1993). This model is used to perform the ML search (log likelihood was -2203.642) followed by rapid bootstrap analyses.

Both MP and ML algorithms provide congruent topologies: the chloroplast diversity of *Ae. geniculata* is represented by two main haplotypes families that differ from each other by a six bp long inversion (i.e. the N and S families, Figure 5 B). The S haplotypes are basal since they are also observed in the outgroup species. Each haplotype family contains a frequently occurring haplotype (i.e. N1 and S1), as well as several secondary haplotypes (i.e. N2, N3 and S2 that are deriving from N1 and S1 respectively).

Haplotypes are not uniformly distributed throughout the sampling distribution (Figure 5 A and C): N haplotypes occur more frequently than S haplotypes in populations from the IBP, FIP and HR areas. The pattern is inverted for populations located in the LEV and NAF areas. The BST area represents a contact zone and shows equivalent frequencies for N and S haplotypes (Figure 5 C). The derived haplotypes provide valuable information in revealing unexpected patterns. For instance, S2 is found in two populations located in the IBP area (E0617A and E0618A) but derives from S1; a haplotype that occurs more frequently in the NAF, LEV and BST areas. Similarly, N2 and N3 haplotypes are observed out of their putative area of origin since N2 is observed in Spain, Italy and Bulgaria while S3 is observed in France (harbour of Marseille), Morocco and Syria.

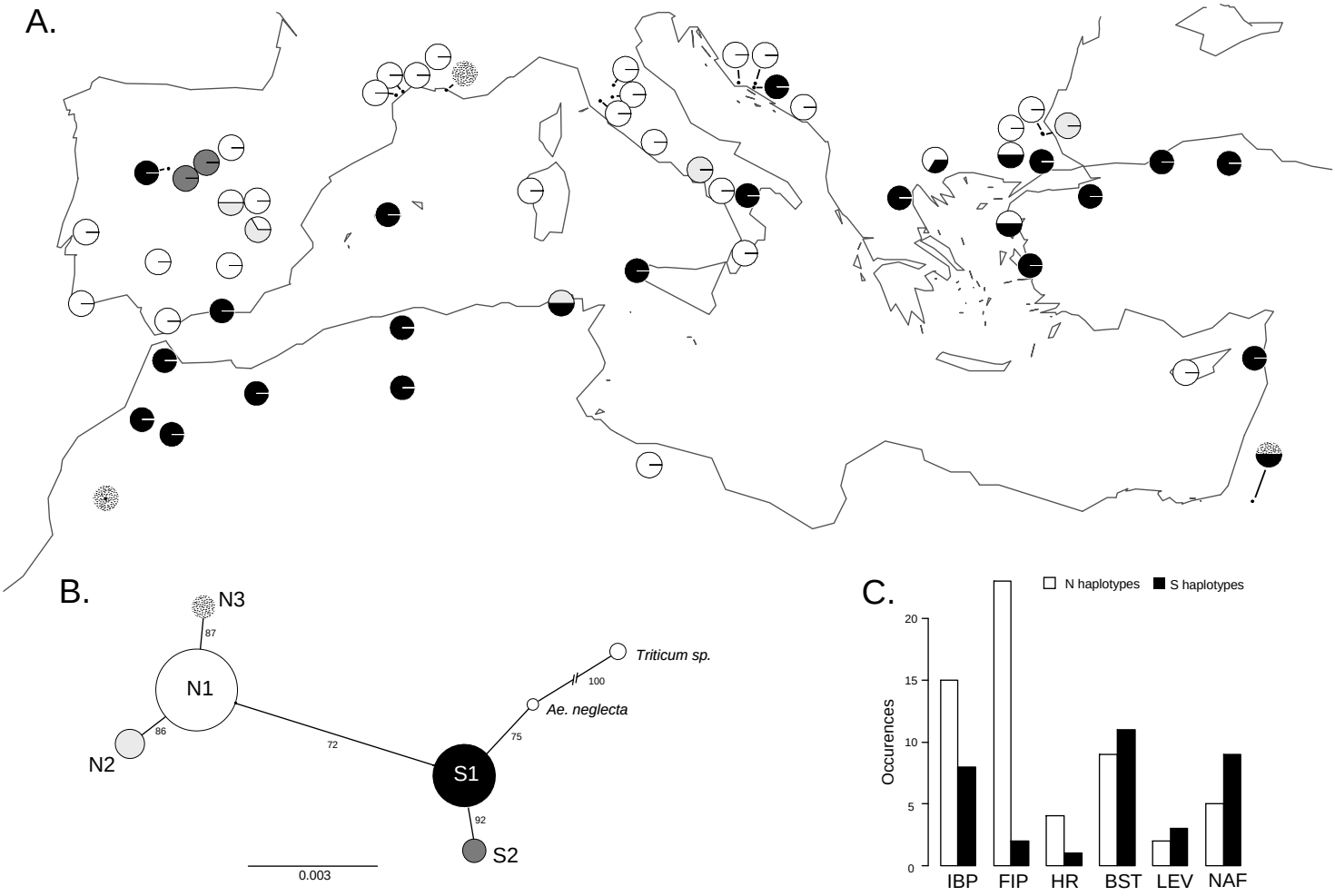

Figure 5. Chloroplast haplotypes distribution. A. Proportions of haplotypes observed in *Ae. geniculata* populations. B. Best maximum likelihood tree of haplotypes. The colors and nomenclature of haplotypes correspond to the map in point A. The diameter of circles represents the number of samples belonging to each haplotype. Only nodes supported with more than 70% are displayed (non supported nodes are pooled in their parent nodes). C. Number of observed haplotypes, belonging either to N or S clade according to the sampling area.

Nuclear and chloroplast datasets comparison

The nuclear and chloroplast datasets are globally congruent (Figure 6). Samples collected in Northern Mediterranean areas mostly belong to the *A, C*, *E*, and *F* groups and are clearly associated to N haplotypes. In contrast, equivalent frequencies of N and S haplotypes are observed in Southern Mediterranean areas (*G* and *H* groups). Interestingly, the *B* group, a Northern Mediterranean entity that occurred in two Spanish populations (E0617A and E0618A) is associated to a specific S haplotype (S2). Finally, samples of the *H* group show different haplotypes according to their geographical location: *H* samples collected in Southern Mediterranean areas (H_{Smed}) are associated to S haplotypes while *H* samples located in Northern Mediterranean areas (H_{Im}) are more frequently associated to N haplotypes.

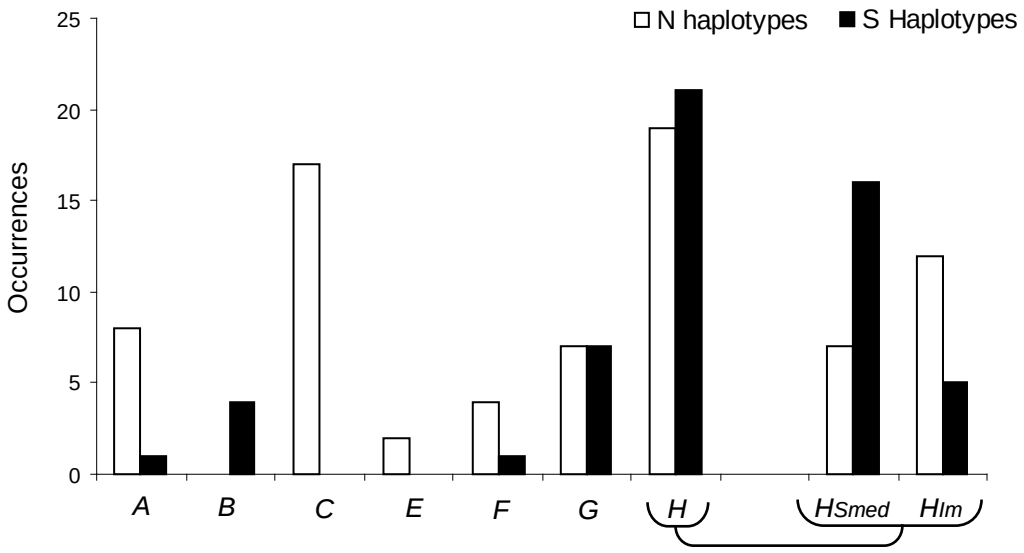

Figure 6. Comparison of AFLP and chloroplast datasets. The barplot displays a contingency table comparing AFLP groups to chloroplast haplotypes. The number of observed haplotypes (occurrences) belonging either to N or S clade is indicated according to the corresponding AFLP groups (*A* – Iberian Peninsula, *B* – Spanish local group, *C* – Franco-Italian Peninsula, *D* - French local group, *E* – Italian local group, *F* – Balkans, *G* – Bosporus Strait and *H* – Northern Africa and Levantine). More detailed statistics are provided for the *H* group since samples were collected out of its native area (H_{Im}) and therefore considered separately from the other samples (H_{Smed}).

Discussion

Datasets quality

The AFLP dataset is highly informative as a large number of AFLP bands were produced. However, the robustness and the consistency of our results cannot exclude the occasional occurrence of two main biases:

I. Difficulties in detecting AFLP bands. This bias occurs since AFLP bands are recorded only if they exceed an absolute fluorescence threshold. This strategy has two consequences. On one hand, the final dataset may either include an exaggerated amount of background noise (that appears as false-positives) if a low fluorescence threshold is applied (e.g. 50 rfu). On the other hand, the use of a high fluorescence threshold (e.g. 100 rfu as in the present study) may discard a large number of AFLP bands that would appear as false-negatives in the final dataset. As a result, this setting would especially disfavour medium-quality samples that produced weaker AFLP bands due to suboptimal reactions. In the present study, we tackle the problem by removing samples that produced significantly fewer AFLP bands during preliminary analyses and we measured the error rate. This statistic is a good estimator of the occurrence of false-negatives in the final dataset. Our results (5.8% in average of error between a sample and its replicate) are in accordance to other studies that also used automated scoring procedures (Holland et al. 2008; Whitlock et al. 2008).

II. Size homoplasy. The AFLP reaction produces a large number of anonymous bands that are identified according to their size. Since the recording of sizes is achieved trough electrophoresis, a significant proportion of AFLP bands (15% to 50% according to Caballero

et al. (2008)) may comigrate at a same position in the electrophoresis without sharing the same sequence. Several studies (Vekemans et al. 2002; Arrigo et al. 2009) showed that this problem is more frequent in inter-specific studies and prevails in short AFLP bands. Since we investigate intra-specific diversity and that short fragments are not prevailing in our AFLP reactions, we assume that size homoplasy occurs at a reasonable level in our dataset. Finally the downstream analyses that we use (i.e. STRUCTURE and distance based analyses) were shown to be relatively robust to this bias (Holland et al. 2008; Arrigo et al. 2009).

The chloroplast dataset provided an important amount of phylogenetic information, considered the low investigated taxonomical level (intraspecific). Robust results are obtained since MP and ML algorithms provide highly congruent topologies. Finally, highly supported clades are in agreement with the AFLP results.

Phylogeography of Ae. geniculata

Globally, populations of the Northern Mediterranean are different from those of Southern Mediterranean, based on both nuclear and chloroplast datasets. It clearly appears that populations occurring in the Northern Mediterranean areas are more structured while less diversified than those located South of the Gibraltar and the Bosporus Straits. These results confirm and extend those previously obtained with RFLP markers (Zaharieva et al. 2001a) and are in accordance with the hypothesis of an origin from Eastern Mediterranean (i.e. the BST area, Turkey and Greece), as proposed by Van Slageren (1994). Several evidences corroborate it:

I. The diploid ancestors of *Ae. geniculata* naturally co-occur in this region (Chennaveeraiah 1960; Waines and Barnhart 1992; Resta et al. 1996).

II. The genetic gradient outlined by the principal coordinate analysis indicate that the expansion of *Ae. geniculata* is oriented East - West.

III. Turkey and Greece include the majority of the chloroplast diversity and represents a contact zone between N and S haplotypes (i.e. these haplotypes are observed in equal frequencies in the BST area).

IV. Areas that are either genetically or geographically distant to the BST area show high differences in haplotype frequencies (e.g. the FIP and the NAF areas show unbalanced frequencies between the N and S haplotypes).

We then assume that *Ae. geniculata* followed two expansion routes from the BST area. The first route colonized the Levant, and Northern Africa. It is characterized by a specific AFLP group (*H* group, Figure 3). Both N and S haplotypes are present in these populations although S haplotypes are more frequent (Figure 5). This result may indicate that the expansion of *Ae. geniculata* in North Africa has kept S haplotypes, while both haplotypes are still present in the Levantine. Interestingly, this expansion route is poorly structured as only one AFLP group is detected and no significant correlations are observed between the geographic distances and the genetic ones. In addition, this expansion route is characterized by high levels of genetic diversity. These results remain difficult to explain since they can originate from several non-exclusive processes. The high genetic diversity suggests an ancient origin, at least in the Levantine area but the absence of clear genetic structures may be in contradiction with such hypothesis.

Alternatively, the recent introduction of agriculture by maritime pioneers in this area, 7000 years before present (Zeder 2008), may have promoted a rapid expansion of ruderal and pasture species. In this framework, pastoralism and especially seasonal transhumance between summer and winter pastures may have homogenised the genetic diversity of *Ae. geniculata* populations, even over long distances. Interestingly, the same pattern is observed in *Lolium* species (Balfourier et al. 2000), where no differentiation was reported between populations growing in North Africa and the Levantine.

The second route colonized the Southern Europe and is characterized by the frequent occurrence of N haplotypes. As a consequence, we assume that this expansion route started westwards of the Bosporus Strait area or in the Ionian Sea. This hypothesis is corroborated by the chloroplast dataset where N haplotypes prevail in populations located on Greek and Bulgarian coasts. European populations of *Ae. geniculata* significantly diverge from the Levantine and North African populations. Their genetic diversity is low but highly structured since significant correlations are observed between the geographic and genetic distances in the Iberian Peninsula (IBP) and the Franco-Italian areas (FIP). In addition, clear phylogeographic patterns are observed since the Iberian (*A* and *B*) and Franco-Italian Peninsula (*C*, *D* and *E* groups) belong to distinct genetic entities. Fine structures are also outlined with the presence three local AFLP groups in Spain, France and Italy (*B, D*, and *E* groups).

Two scenarios may explain these results. The first scenario places the colonization of Europe by *Ae. geniculata* before the glaciations period of the Würm (110'000 to 15'000 years before present). This hypothesis is supported by previous chloroplast data revealing *Ae. geniculata* as one of the oldest polyploid species of the genus (Terachi et al. 1984; Resta et al. 1996; Wang et al. 2000). In a second time, a post-glacial recolonization of Europe, starting from refugia located in the South of the Iberian Peninsula, Italy and Southern Balkans may explain the observed patterns of genetic diversity. These potential refugia and recolonization routes have already been outlined with other plant and animal species, including several diploid *Aegilops* species growing in the Aegan Islands and in Turkey (Ohta 2000; Zaharieva et al. 2001a; Hewitt 2004b; Raskina et al. 2004; Matsuoka et al. 2008). The recolonisation of newly deglaciated areas, achieved through slow expansion rates, may explain the low genetic diversities and the significant correlations that are observed between the geographic distance and the genetic in the IBP and FIP areas. Pyrenees may then appear as an effective barrier between different recolonization routes, while the Alps may have been bypassed by colonization routes that followed the Mediterranean coasts to reach Southern France.

The second scenario considers the expansion of agriculture through Europe as a vector for the spread of segetal vegetation. According to recent studies (Wainwright and Thornes 2004; Zeder 2008), the expansion of agriculture was promoted by maritime colonists that brought crops, cattle and agricultural techniques from one area to the other. Their spread started from the Levant where the earliest evidences of agricultural practices were discovered (ca 11'500 years B.P.) and first reached the Greece and Aegan Islands (ca 9000 – 8000 years B.P.). From there, they quickly reached the Italian Peninsula (8100 years B.P.) and only 500 years were needed before the Iberian Peninsula and Southern France were occupied (7600 BP). Interestingly, Italy acted as a platform to colonize Southern France (Zeder 2008): this may explain the genetic proximity observed between these two areas (as revealed by the *C* group). The

Iberian Peninsula was colonized at the same time from Italy via coastal settlements located on the Eastern and Southern coasts of Spain. This colonisation pattern is corroborated by the chloroplast dataset as the Franco-Italian and the Iberian Peninsula are both associated to N haplotypes. In addition, the genetic divergence observed between populations from the Iberian (*A* and *B*) and the Franco-Italian Peninsula (*C*, *D* and *E* groups) may be explained by distinct colonization events. From these, agriculture expanded progressively within each area. As a result, the Pyrenees, by delaying human expansion, may appear as a barrier between the genetic and cultural pools derived from the Franco-Italian and the Spanish colonization event. Several studies reported suture zones derived from agriculture expansion processes. For instance, both the Iberian cattle (Beja-Pereira et al. 2006) and European *Lolium* species (Balfourier et al. 2000) followed distinct colonisation routes (i.e. the Mediterranean and the Danubian routes) that met respectively in the Pyreneans or in Eastern and Northern Europe. This scenario is partially supported by the low values of genetic diversity measured in the IBP and FIP areas.

At the end, both scenarios remain likely and definitive conclusions cannot be provided. At least can we assume an ancient origin of *Ae. geniculata* in the Mediterranean Basin, with a genetic diversity initially shaped by large scale processes, such as post-glacial recolonization or the early spread of agriculture. Unexpectedly, these diversity patterns can still be detected, despite recent and intensive exchanges caused by human activities in the Mediterranean Basin.

Recent human-driven migration events and recombination

Aegilops geniculata is a colonizing species that frequently occurs at field borders or in pastures. It is likely that human activities have contributed to dispersal in its recent history. For instance, the continuous presence of trade, pastoralism and transhumance (i.e. the introduction of goats in Europe occurred 7700 years BP) might have promoted recurrent and independent migration events (Wainwright and Thornes 2004; Zeder 2008). This hypothesis is corroborated by several discrepancies between the AFLP groups and their geographic distribution that are observed in the dataset (Figure 3). For example, the same genotypes are observed in the IBP and the FIP areas (*A* group samples), the FIP and the BST areas (*C* group samples) or the BST and the IBP areas (*G* group samples). The same pattern is observed for the chloroplast haplotypes (e.g. the N2 and N3 haplotypes), supporting exchanges between Spain, Algeria, Italy and Bulgaria, or France, Morocco and Syria. Seed exchanges or trade routes may explain these patterns. The *F* group has a more complicated structure as it is represented occasionally in Spain and in North Africa. Its most likely origin is located in the Balkans, as Croatian populations mainly belong to this group. These exchanges generally involve a limited number of individuals, but their magnitude may be underestimated due to limitations of our sampling.

The most striking anomalies involve individuals geographically located in Northern Mediterranean areas that are genetically close or identical to individuals that belong to the Southern Mediterranean groups (i.e. samples from the Northern Mediterranean *B* group and samples from the *H* group). We consider that these individuals are migrants. Most of these transfers apparently occurred northwards and originated from Southern Mediterranean areas. While those migration events are clearly related to human activities, it remains difficult to identify their origin and colonization routes. For instance, several historical events (e.g. Phoenician trade routes in the Mediterranean islands or Southern Italy, Arabian invasions in

Southern Spain) broadly overlap with the observed distribution of migrant samples. However, more recent migration events are almost certainly involved as well. Interestingly, the same colonization patterns are observed for sheep and goats in Europe (Beja-Pereira et al. 2006; Zeder et al. 2006) or in ruderal plant species as the *Senecio* complex (Comes and Abbott 2001) or *Hypochaeris radicata* (Ortiz et al. 2008) where migrations occurred from northern Africa to Southern Spain, the Baleares and Sicily.

Finally, our results outline several recombination events that involved migrants and native populations. A first example is provided with one ancient migration event involving two Spanish populations (E0617A and E0618A). Interestingly, E0617A and E0618A samples possess a chloroplast haplotype (S2) related to Southern Mediterranean areas that differs by one substitution from the original S1 haplotype (Figure 5). In addition, E0617A and E0618A samples belong to the *B* group, an AFLP genotype specific to the IBP area and closely related to the Spanish genotypes (*A* group). These results suggest an ancient Southern Mediterranean origin for the E0617A and E0618A populations, followed by an admixture with Northern Mediterranean *A* genotypes. Further examples of migration followed by outcrossing events involving migrant and native genotypes are provided by the H_{lm} samples. These latter possess a somehow "inverted" genome constitution, with AFLP indicating presumably a Southern Mediterranean origin for the nuclear genome (Figure 3) while chloroplasts mainly belong to Northern haplotypes (Figure 5). Indeed, the proportion of N haplotypes in the H_{lm} samples is twice that observed in the rest of the *H* group (Figure 6). This indicates that migrants mate with indigenous populations, an hypothesis corroborated by the AFLP-based principal coordinates analysis (Figure 4) where H_{lm} samples mostly have an intermediate position between Northern and Southern Mediterranean genotypes. These results contradict the commonly assumed autogamy for *Ae. geniculata* and also shows that the divergence between Northern and Southern Mediterranean populations did not set up a reproductive isolation between the two lines.

Conclusion

AFLP and chloroplast datasets show complementary and congruent patterns. Our results suggest an Eastern Mediterranean origin for *Ae. geniculata*. Significant genetic diversity structures are observed as Northern and Southern Mediterranean populations belong to distinct genetic entities. In addition, Northern populations present structures that may reflect the dispersal routes of *Ae. geniculata*. Although definitive conclusions regarding the early history of *Ae. geniculata* cannot be provided, our results suggest that long-term processes, such as post-glaciary recolonization patterns or the early spread of agriculture were involved in shaping the diversity of *Ae. geniculata*. Interestingly, these diversity patterns are still detected, despite long-term influences of human activities in the Mediterranean Basin. As an example, several anomalies observed in the distribution of genotypes advocate for recent and recurrent seeds exchanges over long distance. These exchanges were probably promoted by trade or agriculture. Finally, migrants appeared to have mated with indigenous individuals, a result further contradicting the common assumption that *Ae. geniculata* is a strict autogamous species.

Acknowledgments

The authors are grateful to the team of the Botanical Garden of Neuchâtel for their help in plants cultivation. Xavier Vekemans, Dorothée Ehrich, Philippe Küpfer, Jean-Claude Dusautoir, Mélanie Thobor, Anouk Beguin, Sylvain Lappe, Xavier Bernex, Celia Bueno, John Cleese and Anahí Espindola provided useful comments and technical assistance. This project was funded by the National Centre of Competence in Research (NCCR) Plant Survival and the PNR-59, two research programs of the Swiss National Science Foundation.

Additional File 1. Plant material. The table indicates for each population: the species, the country, the geographical coordinates (decimal degrees, NDD, EDD), the DNA origin and the number of samples included in the AFLP (AFLP) or in the chloroplast (Cp) analyses.

Species	Country	Population	NDD	EDD	DNAOrigin	AFLP	Cp
Ae. geniculata	Algeria	Blida	36.33	3.09	INRA	1	
		BordjBouArreridj	36.06	4.49	INRA	2	
		BouSaada	35.00	3.55	INRA	1	
		Constantine	36.22	7.01	INRA	1	
		Djelfa	34.51	3.48	INRA	1	1
		Guelma	36.22	7.47	INRA		1
		TiziOuzou	36.30	3.47	INRA	1	1
		Tlemcen	34.52	-1.15	INRA	2	
	Azerbaijan	Azerbaidjan	48.58	47.30	INRA	1	
	Bulgaria	AHTOPOL	42.07	28.00	INRA	4	1
		ELHOVO	42.15	26.33	INRA	4	
		G.DELCHEV	41.31	23.49	INRA	4	3
		GRUDOVO	42.20	27.11	INRA	2	
		HARMANLI	41.58	26.00	INRA	2	
		IVAYLOVGRAD	42.27	26.06	INRA	4	
		M.TIRNOVO	42.09	27.33	INRA	2	
		MITCHURIN	42.10	27.52	INRA	3	2
		SOZOPOL	42.26	27.39	INRA	4	
		SVILENGRAD	41.56	26.37	INRA	2	
		YAMBOL	42.25	26.40	INRA	3	1
	Croatia	C0723A	43.59	16.73	Field collections	13	2
		C0723C	43.47	16.69	Field collections	8	2
		Dalmatien	43.31	16.14	INRA	1	1
		Dubrovnik	42.39	18.02	INRA	1	2
	Cyprus	Famagusta	35.26	34.07	INRA	1	
		Kyrenia	35.18	33.04	INRA	1	
		Nikosia	34.57	32.44	INRA	1	
		Paphos	34.50	32.28	INRA	1	1
	France	Agde	43.31	3.47	INRA	1	
		Baillargues	43.66	4.01	INRA	8	
		Baillargues2	43.66	4.01	INRA	1	
		Baillargues3	43.66	4.01	INRA	2	
		Baillargues4	43.66	4.01	INRA	1	
		Castelnau	43.63	3.90	INRA	2	
		Clansaye	44.37	4.82	INRA	2	2
		F070406D	43.71	3.68	Field collections	14	
		F070606D	43.68	3.91	Field collections	11	4
		F1207A	44.05	6.06	Field collections	9	
		F1307A	43.39	5.13	Field collections	8	5
		Fontanes	43.50	4.05	INRA	2	
		Grabels	43.39	3.46	INRA	2	
		JeanVedas	43.35	3.50	INRA	2	2
		Laverune	43.35	3.48	INRA	3	
		MARSEILLE	43.39	5.24	INRA	2	
		Matelles	43.73	3.81	INRA	2	
		Montferriez	43.67	3.86	INRA	2	
		MONTPELLIER	43.40	3.39	INRA	2	
		Mudaison	43.65	4.04	INRA	2	
		Murviels	43.44	3.15	INRA	2	
		Murviels2	43.44	3.15	INRA	1	
		Murviels3	43.44	3.15	INRA	1	
		Neffies	43.32	3.20	INRA	3	
		Paillade	43.39	3.49	INRA	2	
		Pep	43.39	3.49	INRA	2	
		RabieuxClermont	43.40	3.27	INRA	2	
		RoqueSurCeze	44.19	4.52	INRA	1	
		Salagou	43.39	3.23	INRA	2	
		SalonProvence	43.39	5.04	INRA	2	
		Serviers4	44.29	3.99	INRA	4	
		Serviers5	44.29	3.99	INRA	1	
		Serviers6	44.29	3.99	INRA	2	
		Thibery	43.24	3.25	INRA	2	2
		Valette	43.39	3.53	INRA	2	
		Vendargues	43.66	3.97	INRA	2	
	Greece	Elassona	40.17	22.15	INRA	1	1
		Kozani	40.17	21.43	INRA		1
		Makedonia	40.51	22.04	INRA		1
	Italy	Basilica	40.24	16.44	INRA	1	1
		Calabria	38.52	16.34	INRA	1	1
		Chianti	43.23	11.18	INRA	1	
		Fardella	40.38	15.47	INRA	1	1
		Gullura	40.38	8.29	INRA	2	2
		I0623A	43.54	11.42	Field collections	12	2
		I0623D	43.18	11.37	Field collections	3	4
		I0624A	43.15	11.18	Field collections	16	
		I0624C	43.07	10.91	Field collections	2	4
		Napoli	40.46	14.17	INRA	1	1
		Roma	41.39	12.30	INRA	1	1
		Siena	43.13	11.23	INRA	1	
		Trapani	38.00	12.29	INRA	1	1

Additional File 1. (suite)

Species	Country	Population	NDD	EDD	DNAOrigin	AFLP	Cp
Ae. geniculata	Jordan	Amman	31.52	35.50	INRA	1	
(continued)		Irbid	32.39	35.45	INRA	1	
		Karak	31.22	35.42	INRA	2	2
		Mafraq	32.30	36.05	INRA	2	
	Lebanon	Akkar	34.30	36.10	INRA	1	
		AlBiga	33.27	35.49	INRA	2	
		Kesserouan	33.59	35.40	INRA	1	
		Metten	33.54	35.42	INRA	1	
		Qubayat	34.35	36.20	INRA	1	
	Lybia	Tripoli	32.53	13.21	INRA	2	2
	Morocco	ALHOCEIMA	34.58	-4.08	INRA	2	
		AZROU	33.12	-5.18	INRA	1	1
		BERKANE	34.54	-2.24	INRA	2	
		CHEFCHAOUEN	34.56	-4.56	INRA	1	1
		ELHAJEB	33.29	-5.23	INRA	1	
		IFRANE	33.35	-5.05	INRA	4	
		KENITRA	34.15	-5.54	INRA	2	
		KETAMA	34.54	-4.35	INRA	2	
		LARACHE	35.20	-6.04	INRA	1	
		MARRAKECH	31.28	-7.25	INRA	4	2
		MEKNES	33.48	-5.23	INRA	2	
		NADOR	34.58	-3.17	INRA	2	
		OUJDA	34.34	-2.00	INRA	2	2
		RABAT	33.56	-6.30	INRA	3	1
		TETOUAN	35.32	-5.45	INRA	2	2
	Portugal	ArrudaLisboi	38.59	-9.05	INRA	1	
		Beja	38.01	-7.53	INRA	1	
		SantaremSantarem	39.15	-8.40	INRA	1	1
		TaviraFaro	37.09	-7.38	INRA	1	
		VianadoCastelo	37.02	-8.58	INRA	1	1
		VilaReal	37.11	-7.36	INRA	1	
	Spain	849587			Field collections		1
		Baleares	40.00	3.50	INRA	4	2
		Burgos	41.57	-3.07	INRA	1	1
		Cadiz	40.04	-2.51	INRA	2	
		Cordoba	36.56	-5.16	INRA	2	2
		Cuenca	40.05	-2.12	INRA	1	1
		E0609B	40.01	-2.96	Field collections	10	4
		E0610B	39.98	-2.09	Field collections	9	
		E0610C	39.85	-2.02	Field collections	21	
		E0610E	39.22	-1.94	Field collections	5	4
		E0611B	38.15	-3.02	Field collections	2	4
		E0611D	37.57	-4.29	Field collections	2	
		E0613A	36.77	-5.27	Field collections	1	2
		E0615A	38.26	-5.70	Field collections	10	
		E0616D	39.97	-4.71	Field collections	9	
		E0617A	40.75	-4.66	Field collections	14	2
		E0618A	41.22	-3.88	Field collections	15	4
		Granada	37.00	-3.30	INRA	1	2
		Salamanca	41.05	-5.30	INRA	1	1
	Syria	Aleppo	36.30	37.00	INRA	1	
		Damaskus	33.41	36.35	INRA	3	
		Homs	34.54	36.57	INRA	1	
		Lattakia	35.39	35.50	INRA	2	2
	Turkey	Bolu	41.02	31.52	INRA	1	
		Bursa	40.23	29.35	INRA	2	2
		Canakkale	39.51	26.37	INRA	2	2
		Edirne	41.44	26.37	INRA	2	2
		Istanbul	41.05	29.44	INRA	2	
		Izmir	38.26	27.10	INRA	2	2
		Kirklareli	41.30	27.50	INRA	2	2
		Sinop	41.20	34.53	INRA	1	1
		Zonguldak	41.24	32.00	INRA	1	
	Tunisia	Bizerte	37.06	9.53	INRA	2	2
		Kairouan	36.00	10.10	INRA	1	
	Ukraine	Jalta	48.58	34.10	INRA	1	1
Ae. triuncialis	Bulgaria	PLOVDIV	42.04	26.04	INRA	3	
	Croatia	Dalmatien	43.10	16.15	INRA	1	
	Greece	GRE	38.52	21.91	INRA	3	
		Votonossi	39.47	21.05	INRA	1	2
	Iran	Iran	32.43	53.69	INRA	1	
	Lybia	JabalAlAkhdar	32.37	22.15	INRA	2	
		Tripoli	32.50	13.00	INRA	1	
		Turk10	38.66	29.14	Field collections	15	
	Spain	757282			Field collections		1
Ae. neglecta	Spain	E0615A	38.26	-5.70	Field collections		4
Ae. biuncialis	Spain	E0610C	39.85	-2.02	Field collections		1
Triticum aestivum					Field collections		1
Triticum turgidum					Field collections		1
					Total	438	125

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Kiado A, editor. Second International Symposium on Information Theory. Budapest. pp. 267–281.

Arrigo N, Tuszynski JW, Ehrich D, Gerdes T, Alvarez N (2009) Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. BMC Bioinformatics 10.

Balfourier F, Imbert C, Charmet G (2000) Evidence for phylogeographic structure in *Lolium* species related to the spread of agriculture in Europe. A cpDNA study. Theoretical And Applied Genetics 101: 131-138.

Beja-Pereira A, Caramelli D, Lalueza-Fox C, Vernesi C, Ferrand N et al. (2006) The origin of European cattle: Evidence from modern and ancient DNA. Proceedings Of The National Academy Of Sciences Of The United States Of America 103(21): 8113-8118.

Besnard G, Khadari B, Baradat P, Berville A (2002) *Olea europaea* (Oleaceae) phylogeography based on chloroplast DNA polymorphism. Theoretical And Applied Genetics 104(8): 1353-1361.

Bittkau C, Comes HP (2005) Evolutionary processes in a continental island system: molecular phylogeography of the Aegean *Nigella arvensis* alliance (*Ranunculaceae*) inferred from chloroplast DNA. Molecular Ecology 14(13): 4065-4083.

Bonin A, Bellemain E, Bronken Edeisen P, Pompanon F, Brochmann C et al. (2004) How to track and assess genotyping errors in population genetics studies. Molecular Ecology 13: 3261-3273.

Caballero A, Quesada H, Rolan-Alvarez E (2008) Impact of amplified fragment length polymorphism size homoplasy on the estimation of population genetic diversity and detection of selective loci. Genetics 179: 539- 645.

Chen DH, Ronald PC (1999) A rapid DNA minipreparation method suitable for AFLP and other PCR applications. Plant Mol Biol Rep 17(1): 53-57.

Chennaveeraiah MS (1960) Some aspects of the cytotaxonomy of *Aegilops*. Proceedings summer school of botany: 43-50.

Comes HP, Abbott RJ (2001) Molecular phylogeography, reticulation, and lineage sorting in Mediterranean *Senecio* sect. *Senecio* (*Asteraceae*). Evolution 55(10): 1943-1962.

Conedera M, Krebs P, Tinner W, Pradella M, Torriani D (2004) The cultivation of *Castanea sativa* (Mill.) in Europe, from its origin to its diffusion on a continental scale. Vegetation History And Archaeobotany 13(3): 161-179.

de Casas RR, Besnard G, Schonswetter P, Balaguer L, Vargas P (2006) Extensive gene flow blurs phylogeographic but not phylogenetic signal in *Olea europaea* L. Theoretical And Applied Genetics 113(4): 575-583.

Ehrich D (2006) AFLPdat: a collection of R functions for convenient handling of AFLP data. Mol Ecol Notes 6: 603-604.

Eidesen PB, Alsos IG, Popp M, Stensrud O, Suda J et al. (2007) Nuclear vs. plastid data: complex Pleistocene history of a circumpolar key species. Molecular Ecology 16(18): 3902-3925.

Felsenstein J (1985) Confidence-limits on phylogenies - an approach using the bootstrap. Evolution 39: 783-791.

Fineschi S, Taurchini D, Villani F, Vendramin GG (2000) Chloroplast DNA polymorphism reveals little geographical structure in *Castanea sativa* Mill. (*Fagaceae*) throughout southern European countries. Molecular Ecology 9(10): 1495-1503.

Fjellheim S, Rognli OA, Fosnes K, Brochmann C (2006) Phylogeographical history of the widespread meadow fescue (*Festuca pratensis* Huds.) inferred from chloroplast DNA sequences. Journal Of Biogeography 33(8): 1470-1478.

Francois O, Blum MGB, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. Plos Genetics 4(5).

Friebe BR, Tuleen NA, Gill BS (1999) Development and identification of a complete set of *Triticum aestivum Aegilops geniculata* chromosome addition lines. Genome 42(3): 374-380.

Gugerli F, Englisch T, Niklfeld H, Tribsch A, Mirek Z et al. (2008) Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation - a project synopsis. Perspect Plant Ecol 10(4): 259-281.

Guo CH, Terachi T (2005) Variations in a hotspot region of chloroplast DNAs among common wheat and *Aegilops* revealed by nucleotide sequence analysis. Genes & Genetic Systems 80(4): 277-285.

Haider N, Nabulsi I (2008) Identification of *Aegilops* L. species and *Triticum aestivum* L. based on chloroplast DNA. Genetic Resources And Crop Evolution 55(4): 537-549.

Hall TA (1999) Bioedit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41: 95-98.

Hammer K (1980) Vorarbeiten zur Monographischen Darstellung von Wildpflanzensortimenten: *Aegilops* L. Kulturpflanze 28: 33-180.

Hensen I (2002) Seed predation by ants in south-eastern Spain (Desierto de Tabernas, Almería). Anales de Biología 24: 89-96.

Hewitt GM (2004a) Genetic consequences of climatic oscillations in the Quaternary. Philosophical Transactions Of The Royal Society Of London Series B-Biological Sciences 359(1442): 183-195.

Hewitt GM (2004b) Genetic consequences of climatic oscillations in the
Quaternary. Phil Trans R Soc Lond B 359: 183–195.

Holland BR, Clarke AC, Meudt HM (2008) Optimizing automated AFLP scoring parameters to improve phylogenetic resolution. Syst Biol 57(3): 347-366.

Kluge AG (1989) A Concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst Zool 38: 7-25.

Matsuoka Y, Takumi S, Kawahara T (2008) Flowering time diversification and dispersal in central eurasian wild wheat *Aegilops tauschii* Coss.: genealogical and ecological framework. Plos One 3(9).

Medini M, Hamza S, Rebai A, Baum M (2005) Analysis of genetic diversity in Tunisian durum wheat cultivars and related wild species by SSR and AFLP markers. Genetic Resources And Crop Evolution 52(1): 21-31.

Meimberg H, Hammond JI, Jorgensen CM, Park TW, Gerlach JD et al. (2006) Molecular evidence for an extreme genetic bottleneck during introduction of an invading grass to California. Biological Invasions 8(6): 1355-1366.

Monte JV, De Nova PJG, Soler C (2001) AFLP-based analysis to study genetic variability and relationships in the Spanish species of the genus *Aegilops*. Hereditas 135(2-3): 233-238.

Morrison DA (2006) Multiple sequence alignment for phylogenetic purposes. Aust Syst Bot 19: 479-539.

Munzner T, Guimbretiere F, Tasiran S, Zhang L, Zhou Y (2003) TreeJuxtaposer: scalable tree comparison using focus, plus context with guaranteed visibility. ACM Trans Graph 22: 453-462.

Naghavi MR, Mardi M, Pirseyedi SM, Kazemi M, Potki P et al. (2007) Comparison of genetic variation among accessions of *Aegilops tauschii* using AFLP and SSR markers. Genet Resour Crop Evol 54: 237–240.

Nixon KC (1999) The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics 15: 407-414.

Ohta S (2000) Genetic differentiation and post-glacial establishment of the geographical distribution of *Aegilops caudata* L. Genes & Genetic Systems 75: 189-196.

Ortiz MA, Tremetsberger K, Terrab A, Stuessy TF, Garcia-Castano JL et al. (2008) Phylogeography of the invasive weed *Hypochaeris radicata* (Asteraceae): from Moroccan origin to worldwide introduced populations. Molecular Ecology 17(16): 3654-3667.

Parducci L, Szmidt AE, Madaghiele A, Anzidei M, Vendramin GG (2001) Genetic variation at chloroplast microsatellites (cpSSRs) in *Abies nebrodensis* (Lojac.) Mattei and three neighboring *Abies* species. Theoretical And Applied Genetics 102(5): 733-740.

Paun O, Schönswetter P, Winkler M, Consortium I, Tribsch A (2008) Historical divergence vs. contemporary gene flow: evolutionary history of the calcicole *Ranunculus alpestris* group (Ranunculaceae) in the European Alps and the Carpathians. Molecular Ecology 17: 4263-4275.

Pester TA, Ward SM, Fenwick AL, Westra P, Nissen SJ (2003) Genetic diversity of jointed goatgrass (*Aegilops cylindrica*) determined with RAPD and AFLP markers. Weed Science 51(3): 287-293.

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. Bioinformatics 14: 817-818.

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

Provan J, Wolters P, Caldwell KH, Powell W (2004) High-resolution organellar genome analysis of *Triticum* and *Aegilops* sheds new light on cytoplasm evolution in wheat. Theoretical And Applied Genetics 108(6): 1182-1190.

Raskina O, Belyayev A, Nevo E (2004) Quantum speciation in *Aegilops*: Molecular cytogenetic evidence from rDNA cluster variability in natural populations. Proceedings Of The National Academy Of Sciences Of The United States Of America 101(41): 14818-14823.

Resta P, Zhang H-B, Dubcovsky J, Dvorak J (1996) The origins of the genomes of *Triticum biunciale*, *T. ovatum*, *T. neglectum*, *T. columnare*, and *T. rectum* (Poaceae) based on variation in repeated nucleotide sequences. American Journal of Botany 83(12): 1556-1565.

Sasanuma T, Chabane K, Endo TR, Valkoun J (2004) Characterization of genetic variation in and phylogenetic relationships among diploid *Aegilops* species by AFLP: incongruity of chloroplast and nuclear data. Theoretical And Applied Genetics 108(4): 612-618.

Schmitt T (2007) Molecular biogeography of Europe: Pleistocene cycles and

postglacial trends. Frontiers in Zoology 4(11).

Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Molecular Ecology 9: 2109–2118.

Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. American Journal Of Botany 94(3): 275-288.

Shaw J, Lickey EB, Beck JT, Farmer SB, Liu WS et al. (2005) The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal Of Botany 92(1): 142-166.

Sikes DS, Lewis PO (2001) PAUPRAT. University of Connecticut, Storrs.

Stamatakis A (2006) RAxML-VI-HPC: Maximum Likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web-servers. Syst Biol 57: 758-771.

Swofford DL (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates, Sunderland, Massachusetts.

Terachi T, Ogihara Y, Tsunewaki K (1984) The molecular-basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*.3. Chloroplast genomes of the M and modified-M genome-carrying species. Genetics 108(3): 681-695.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The Clustal X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876-4882.

van Slageren MW (1994) Wild wheats: A monograph of *Aegilops* L.and *Ambylopyrum* (Jaub and Spach) Eig (*Poaceae*). Wageningen: Agricultural Univ., Wageningen, The Netherlands and Int. Cent. for Agric. Res. in Dry Areas, Aleppo, Syria.

Vekemans X, Beauwens T, Lemaire M, Roldan-Ruiz I (2002) Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Molecular Ecology 11: 139-151.

Viard F, Arnaux J-F, Delescluse M, Cuguen J (2004) Tracing back seed and pollen flow within the crop–wild *Beta vulgaris* complex: genetic distinctiveness vs. hot spots of hybridization over a regional scale. Molecular Ecology 13: 1357–1364.

Waines JG, Barnhart D (1992) Biosystematic research in *Aegilops* and *Triticum*. Hereditas 116(3): 207-212.

Wainwright J, Thornes JB (2004) Environmental issues in the Mediterranean: processes and perspectives from the past and present. London: Routledge.

Wang GZ, Matsuoka Y, Tsunewaki K (2000) Evolutionary features of chondriome divergence in *Triticum* (wheat) and *Aegilops* shown by RFLP analysis of mitochondrial DNAs. Theoretical And Applied Genetics 100(2): 221-231.

Whitlock R, Hipperson H, Mannarelli M, Burke T (2008) A high-throughput protocol for extracting high-purity genomic DNA from plants and animals. Mol Ecol Resour 8(4): 736-741.

Yamane K, Kawahara T (2005) Intra- and interspecific phylogenetic relationships among diploid *Triticum-Aegilops* species (Poaceae) based on base-pair substitutions, indels, and microsatellites in chloroplast noncoding sequences. American Journal Of Botany 92(11): 1887-1898.

Zaharieva M, Monneveux P (2006) Spontaneous hybridization between bread wheat (*Triticum aestivum* L.) and its wild relatives in Europe. Crop Science 46(2): 512-527.

Zaharieva M, Santoni S, David J (2001a) Use of RFLP markers to study genetic diversity and to build a core-collection of the wild wheat relative *Ae. geniculata* Roth (= *Ae. ovata* L.). Genetics Selection Evolution 33: S269-S288.

Zaharieva M, Gaulin E, Havaux M, Acevedo E, Monneveux P (2001b) Drought and heat responses in the wild wheat relative *Aegilops geniculata* Roth: Potential interest for wheat improvement. Crop Science 41(4): 1321-1329.

Zaharieva M, Dimov A, Stankova P, David J, Monneveux P (2003) Morphological diversity and potential interest for wheat improvement of three *Aegilops* L. species from Bulgaria. Genetic Resources and Crop Evolution 50: 507–517.

Zeder M (2008) Domestication and early agriculture in the Mediterranean basin: origins, diffusion and impact. pnas 105(33): 11597-11604.

Zeder M, Emshwiller E, Smith BD, Bradley D (2006) Documenting domestication: the intersection of genetics and archaeology. Trends In Genetics 22(3): 139-155.

Chapter 6

Production and characterization of first and second generation hybrids using *Aegilops cylindrica* Host and transgenic lines of "Bobwhite" bread wheat

Arrigo N[1], Guadagnuolo R[1] and Felber F[1]

[1]Laboratory of Evolutionary Botany, Institute of Biology, University of Neuchâtel, 11 rue Emile-Argand, 2000 Neuchâtel, Switzerland

Abstract

The natural hybridization between ***Aegilops cylindrica*** Host and the bread wheat (***Triticum aestivum*** L.), as well as the spontaneous occurrence of backcrossed progenies are nowadays largely documented. This process, leading potentially to the introgression of wheat (trans)genes within ***Ae. cylindrica*** represents a major agronomical problem in the United States. Here, we simulate an introgressive series using ***Ae. cylindrica*** and six conventional or transgenic "Bobwhite" bread wheat lines. First generation hybrids were produced using emasculated ***Ae. cylindrica*** florets. The obtained hybrids were then backcrossed in "semi-natural conditions" using ***Ae. cylindrica*** as pollen donor. Fitness components of the two produced generations were estimated measuring the fertility of spikes and the survival of seedlings. First generation hybrids had a relatively high seed survival (85% of F_1 seeds survived) but were highly sterile (an average of 0.32% of fertility). As a result, a limited but reasonable number of backcrossed individuals could be produced (181 BC_1 seeds were available). Due to the pentaploid chromosome content of the F_1 generation, various chromosome numbers and a large phenotypic variation were observed in the BC_1 generation. Unfortunately, the viability of BC1 seeds was limited since only 18 seedlings survived (10% survival). Finally, we investigated the genetic transmission of the transgenes and three molecular markers specific to the "Bobwhite" line using PCR based techniques. All the F_1 hybrids possessed wheat specific markers. The same applied for F_1 samples derived from transgenic wheat, except for progenies obtained from one wheat variety (BW71) that appeared to be hemizygous for the transgene. The transmission of genetic markers to the backcrossed samples was difficult to assess due to the low number of available plants. Finally, issues regarding the production of third generation of backcrossed individuals were discussed.

Introduction

Aegilops cylindrica Host is a wild relative of the wheat (***Triticum*** L.). Although its origin is located in central-Asia, this species has adventively colonized the Mediterranean basin (including France, Nothern Italy, Spain) and the United States (van Slageren 1994). Several populations are also reported in Switzerland, in disturbed areas such as railway stations or vineyards (Guadagnuolo et al. 2001).

The species represents a major weed of wheat fields in the US, where 32 states report its presence (USDA, http://plants.usda.gov/java/profile?symbol=AECY, accessed on the 13th of February 2009). Classical control strategies were unable to limit the expansion of this weed since its introduction in the late nineteenth century. As it belongs to the Poaceae family, this weed cannot be eradicated by selective herbicides in wheat cultivations. Moreover, its habitus (i.e. the plant can be as high as a wheat plant) and especially its spike morphology (i.e. mature spikes disaggregate into spikelets that have the size of a wheat grain and cannot be mechanically filtered from harvest) make ***Ae. cylindrica*** especially prone to contaminate wheat seed stocks in the US agrosystems.

Hybridization events involving ***Ae. cylindrica*** and the bread wheat (***Triticum aestivum*** L.) occur regularly in wheat fields (Gaines et al. 2008). ***Ae. cylindrica*** is an allotetraploid plant (n = 7 chromosomes, organised in two genomes: C and D) while bread wheat is an allohexaploid (n = 7 chromosomes organised in three genomes: A, B and D). The two species share the D genome, inherited from their common ancestor ***Aegilops tauschii*** Coss. F_1 hybrids of both species are pentaploids (with genomes ABCDD). They are male sterile but display a low female fertility that allows the production of backcrossed generations (Schoenenberger et al. 2006). During meiosis, the segregation and recombination of chromosomes is complicated by the pentaploidy and the heterogeneous genome composition of the F_1 hybrids. Chromosomes belonging to the D genome generally perform normal pairings and segregate normally. In contrast, chromosomes belonging to the other genomes remain generally as univalents and occasionally undergo homoeologous crossing-overs (Wang et al. 2000), leading finally to a random segregation of chromosomes in the produced gametes (Wang et al. 2001). Despite these difficulties, backcrossed individuals occur under field conditions (Morrison et al. 2002a; Morrison et al. 2002b; Gandhi et al. 2006). These observations suggest that the transmission of a transgene from the crop to the weed must be taken into account.

The present study attempts to simulate an introgression from transgenic bread wheat to ***Ae. cylindrica***. It includes the production and characterisation (i.e. fertility, survival, chromosome number, transgene and genetic markers [SCARs] transmission) of first generation hybrids and backcrossed individuals.

Material and methods

Plant material

Aegilops cylindrica seeds were collected within a large wild population growing along a railway near the station of Sierre (Switzerland).

A total of six transformed Bobwhite wheat lines (Stoger et al. 1998; Stoger et al. 1999; Arcalis et al. 2004) and one non-transformed control were used as male parents (Table 1) in order to produce first generation hybrids (F_1 individuals, ***Ae. cylindrica*** x ***T. aestivum***). The transgenic construct differed according to the wheat line, but all transgenic varieties carried the reporter "Bar" gene, conferring a resistance to the Basta® herbicide. The transgenes were not mapped in the genomes of the used wheats.

Table 1. Transgenic wheat lines (Stoger et al. 1998; Stoger et al. 1999; Arcalis et al. 2004).

Wheats	Inserted transgenic construct
BW WT	Wild Type – Non transgenic isoline of the transformed lines
BW HSA	Human Serum Albumine production
BW 71	Bar Gene — BASTA herbicide resistance
BW 99	GNA — ***Galanthus nivalis*** agglutinin, insecticidal activity
BW 28	Bar Gene — BASTA herbicide resistance
BW 34	Bar Gene — BASTA herbicide resistance
BW 49	Bar Gene — BASTA herbicide resistance

Crossing display

Crossings were conducted indoors, at the Botanical Garden of Neuchâtel. First generation hybrids were produced with emasculated ***Ae. cylindrica*** used as female parents. The second generation hybrids were obtained under "indoor open pollination" experiment where F_1 individuals (used as female parents) were randomly displayed in an artificial ***Ae. cylindrica*** population (7000 individuals covering 25 m^2). Pollen circulation within the greenhouse was enhanced by using four timed fans displayed at the room corners.

Biological traits measures

Leaves from each produced F_1 and BC_1 individual were collected and stored in silica-gel during the different experiments in order to achieve further analyses. Root tips were collected and pre-treated with a water saturated solution of -bromonaphtalene for 180 minutes at room temperature, before being fixed in an absolute ethanol and glacial acetic acid (3:1) solution (added with acetocarmine and traces of iron acetate). The final staining step was conducted in a acetocarmine solution (Schoenenberger et al. 2006). Chromosome counts were only performed on the surviving BC_1 individuals. The number of surviving seedlings (nsurv), florets (nflorets) and retrieved seeds (Seed set) was systematically recorded. Female fertility was calculated as Fertility = 100 * Seed set / nflorets.

DNA extraction

Only F_1 individuals that produced a viable BC_1 generation were included in the analysis. DNA was extracted using 10mg of dried leaves with a CTAB protocol (Chen and Ronald 1999). DNA concentrations were standardized at 10 ng/ l after a quality and quantity check on 1% agarose gels.

PCR markers

Four molecular markers were used to characterize the offspring generation: the "Bar" reporter gene and three SCARs (i.e. DP9 – located on chromosome 6A, D1P9 – chromosome 6D and GB10 – chromosome 5B). The SCARs primers and the PCR protocols were developed by Schoenenberger et al. (Schoenenberger et al. 2005). Positive and negative controls were always included in the PCR reactions.

Results

Seedlings production results

The production of F_1 individuals showed 10.6 % of average fertility when ***Ae. cylindrica*** was used as female parent (Table 2). F_1 seedlings were clearly viable and their germination was as high as that of ***Ae. cylindrica*** seeds (with respectively an average germination rate of 85.1% and 83.8%). By contrast, the production of BC_1 individuals was complicated by the low fertility of F_1 individuals (only 0.32% of the F_1 florets were fertile, Table 3) and by a low germination rate since only 18 plants out of 181 sown seedlings emerged and survived after winter (10% survival).

Table 2. Fertility results. Each crossing is represented by a line in the table. The crossing scheme differed according to the produced generation. F_1 individuals were obtained with emasculated ***Ae. cylindrica*** as female parent whereas BC_1 individuals were produced with intact F_1 individuals. The number of florets (nflorets), the number of retrieved seeds (Seed set), the number of sown seeds (nseeds) obtained with the considered crossing and their post-winter survival (nsurv) are indicated; their associated percentages are displayed between brackets (i.e. corresponding either to the fertility or the germination rate).

Crossing scheme			Parents Fertility			Offspring survival		
Female parent		Male parent	nflorets	Seed set (%)		nseeds	nsurv (%)	
Ae. cylindrica	x	***Ae. cylindrica***	---	---	---	216	181	(83.8)
F1 crossings								
Ae. cylindrica	x	BW 28	104	16	(15.38)	16	15	(93.75)
Ae. cylindrica	x	BW 34	140	2	(1.43)	2	2	(100)
Ae. cylindrica	x	BW 49	255	17	(6.67)	17	14	(82.35)
Ae. cylindrica	x	BW 71	361	61	(16.9)	60	53	(88.33)
Ae. cylindrica	x	BW 99	944	134.5	(14.25)	60	48	(80)
Ae. cylindrica	x	BW HSA	694	87	(12.54)	60	32	(53.33)
Ae. cylindrica	x	BW WT	653	47	(7.2)	47	46	(97.87)
BC1 crossings								
F1 BW 71	x	***Ae. cylindrica***	12418	36	(0.29)	36	5	(13.89)
F1 BW 99	x	***Ae. cylindrica***	30316	139	(0.46)	60	5	(8.33)
F1 BW HSA	x	***Ae. cylindrica***	21018	25	(0.12)	25	3	(12)
F1 BW WT	x	***Ae. cylindrica***	27043	109	(0.4)	60	5	(8.33)

Genetic characterization

The whole F_1 generation was positive for the wheat specific markers (Table 3). The presence of the transgene was confirmed in each analyzed F_1 individual, except in the BW 71 descent line where only six F_1 individuals (representing 42.84% of the BW 71 descent) contained the transgene. Due to the pentaploidy of their maternal parents, BC_1 individuals had varied chromosome numbers (Table 4) and displayed a high morphologic and phenologic variability. Two BC_1 individuals recovered almost a standard chromosome number and contained wheat specific markers. The transgene transmission from the F_1 parent to the BC_1 generation was effective on three BC_1 individuals out of the 13 potentially transgenic BC_1 individuals. Finally, the transmission of SCARs provided unclear results due to the low number of available samples.

Table 3. Genetic results. Features of the different crossed individuals. The ploidy level (2n), the number of analysed plants (nplants), the number of plants showing presence of the transgene (Bar +) and the three SCARs (DP9, D1P9 and GB10) are summarized (percentages are indicated within brackets).

Crossed line	2n	nplants	Bar +		DP9		D1P9		GB10	
F1 generation										
F1 BW 71	35	14	6	(42.86)	14	(100)	14	(100)	14	(100)
F1 BW 99	35	36	36	(100)	36	(100)	36	(100)	36	(100)
F1 BW HSA	35	12	12	(100)	12	(100)	12	(100)	12	(100)
F1 BW WT	35	32	0	(0)	32	(100)	32	(100)	32	(100)
BC1 generation										
BC1 BW 71	33-46	5	0	(0)	4	(80)	1	(20)	5	(100)
BC1 BW 99	35-44	5	1	(20)	4	(80)	1	(20)	2	(40)
BC1 BW HSA	27-34	3	2	(66.67)	1	(33.33)	3	(100)	3	(100)
BC1 BW WT	28-42	5	0	(0)	4	(80)	5	(100)	4	(80)

Table 4. Genetic characterisation of BC_1 individuals. Each sample is analysed. The chromosome number (2n), the presence of the "Bar" reporter gene and the three SCARs (DP9 - chromosome 6A; D1P9 – chromosome 6D and GB10 – chromosome 5B) is reported for each BC1 sample ("+ in BC1") and F1 parent ("+ in F1 parent").

	Genetic markers transmission								
	2n	Bar + in F1 parent	Bar 1+ in BC1	DP9 + in F1 parent	DP9 + in BC1	D1P9 + in F1 parent	D1P9 + in BC1	GB10 + in F1 parent	GB10 + in BC1
BW 71									
BC1_14	33	1		1	1	1	1	1	1
BC1_15	38-40			1		1		1	1
BC1_16	40	1		1	1	1		1	1
BC1_17	46			1	1	1		1	1
BC1_18	43			1	1	1		1	1
BW 99									
BC1_9	44	1	1	1	1	1		1	1
BC1_10	---	1		1	1	1	1	1	
BC1_11	42	1		1	1	1		1	1
BC1_12	33-34	1		1		1		1	
BC1_13	35	1		1	1	1		1	
BW HSA									
BC1_1	---	1	1	1		1	1	1	1
BC1_2	34-35	1		1		1	1	1	
BC1_3	26-27	1	1	1	1	1	1	1	1
BW WT									
BC1_4	42	---		1	1	1	1	1	1
BC1_5	---	---		1	1	1	1	1	1
BC1_6	---	---		1	1	1	1	1	1
BC1_7	33-34	---		1	1	1	1	1	
BC1_8	28	---		1		1	1	1	1

Discussion

The fertility of ***Ae. cylindrica*** and F_1 samples corresponded to the values obtained by Schoenenberger et al. (2006). However, the F_1 fertility (0.32%) was lower than values reported in open field experiments (ranging between 0.87% and 3.8% (Zemetra et al. 1998; Snyder et al. 2000; Wang et al. 2001; Morrison et al. 2002b)). The especially low germination rate of BC_1 samples (10%) was similar to that obtained by Schoenenberger et al. (2006) (17.8%). These results confirmed that hybridization events involving ***Ae. cylindrica*** and ***T. aestivum*** can be simulated in a greenhouse experiment. We outlined that wheat pollen was able to germinate and fertilize ***Ae. cylindrica*** ovules and that F_1 individuals had a low but non-null female fertility.

The transmission of wheat genetic markers to the backcrossed generations can be explained by two distinct mechanisms. First, surnumerary chromosomes were observed in a large proportion of backcrossed individuals. Those chromosomes could act as a temporary pool of genes since they allow the presence of wheat genes in an ***Aegilops*** lineage during the initial generations of the introgression line. However we could not properly evaluate the persistence of surnumerary chromosomes in further backcrossed individuals because of two main limitations:

I. The production of seeds stopped at the BC_1 generation. Due to the limited time allowed for this PhD study, we could not restart the production of new introgression lines from the beginning and we decided to focus on the other topics of this PhD study.
II. Chromosome counting using aceto-carmine staining cannot provide information regarding the genome origin of chromosomes. For instance, at least two BC_1 individuals present an "***Aegilops***-like" chromosome number (BC1_3 and BC1_8 with 26-27 and 28 chromosomes respectively). The expected chromosome composition of these samples is $A_?B_?C_{7+?}D_{14}$, meaning that not less than 6-7 chromosomes can belong to the A, B or C genomes. In this context, only more complicated and difficult to set-up techniques such as genomic in situ hybridization (GISH) would have provided an accurate identification of surnumerary chromosomes.

Second, only crossing-over and translocations are likely to stabilize wheat genetic material in the ***Ae. cylindrica*** genome. This phenomenon depends on the chromosome location of the considered genes. Our results showed that genetic material was not systematically transmitted in subsequent generations (Table 3 and Table 4). We could however not infer hypotheses about effects due to chromosome location on the likelihood of transmission between generations. Schoenenberger et al. (2005) showed that the DP9 marker, located on the chromosome 6A, was systematically transmitted to the BC_1 generation, a result that validated previous studies outlining the prevalence of the genome A in homoeologous chromosome pairings in artificial F_1 hybrids (Cifuentes et al. 2006).

Conclusion

Our main problem was the production of BC_1 individuals. This step was crucial since the germination of BC_1 seeds may vary according to the year. Additionally, the variable chromosome content of the BC_1 generation complicated the analyses, as different chromosome lineages showed highly differing phenotypes. As a result, gene expression or fitness experiments using this plant material would have suffered from unpredictable confound effects due to the variability of the chromosome content of the BC_1 generation. In this context, only the use of a large number of BC_1 samples may capture a significant proportion of the diversity expressed by the BC_1 generation and could provide valuable results. Alternatively, the use of samples from the BC_2, BC_3 or BC_n generations would provide a more homogeneous plant material but it implies to produce the necessary seeds. For the same amount of time and work, we suggest to enhance the pollen load used to produce the different backcrossed generations. A solution would be to expose “pollen-receiving” to “pollen-giving” spikes in an optimal way. To this respect, our design allowed to recover more seeds than Schoenenberger et al. (2006). Additionnaly, backcrossed individuals that are produced in natural conditions quickly recover an “***Aegilops***-like” chromosome content (e.g. BC_1 individuals with $2n = 29 – 30$ were frequently obtained by Guadagnuolo et al. (2001)). This would constitute a solution to decrease the observed variance of the BC_1 generation, while the pollination problem remains to be solved. Finally and if possible we would use transgenic wheat lines where the transgene was mapped.

Acknowledgments

This study was funded by the NCCR Plant Survival. The authors warmly thank Anouk Beguin and the team of the Botanical Garden of Neuchâtel (Elisabeth and Laurent Oppliger, Claire-Aline Nussbaum and Nathalie Pascotto) for their logistic help during the production of the F_1 and BC_1 generations. In addition, the authors are extremely grateful to the team of 25 friends and family members that helped in retrieving BC_1 seeds from the F_1 spikes.

References

Arcalis E, Marcel S, Altmann F, Kolarich D, Drakakaki G et al. (2004) Unexpected deposition patterns of recombinant proteins in post-endoplasmic reticulum compartments of wheat endosperm[1]. Plant Physiology 136: 3457-3466.

Chen DH, Ronald PC (1999) A rapid DNA minipreparation method suitable for AFLP and other PCR applications. Plant Mol Biol Rep 17(1): 53-57.

Cifuentes M, Blein M, Benavente E (2006) A cytomolecular approach to assess the potential of gene transfer from a crop (***Triticum turgidum*** L.) to a wide relative (***Aegilops geniculata*** Roth.). Theoretical And Applied Genetics 112: 657-664.

Gaines TA, Henry WB, Westra P, Nissen SJ, Shaner DL (2008) Jointed goatgrass (***Aegilops cylindrica***) by imidazolinone-resistant wheat hybridization under field conditions. Weed Science 56: 32-36.

Gandhi HT, Mallory-Smith CA, Watson CJW, Zemetra RS, Riera-Lizarazu O (2006) Hybridization between wheat and jointed goatgrass (***Aegilops cylindrica***) under field conditions. Weed Science 54: 1073-1079.

Guadagnuolo R, Savova-Bianchi D, Felber F (2001) Gene flow from wheat (***Triticum aestivum*** L.) to jointed goatgrass (***Aegilops cylindrica*** Host.), as revealed by RAPD and microsatellite markers. Theoretical And Applied Genetics 103(1): 1-8.

Morrison LA, Cremieux L, Mallory-Smith CA (2002a) Infestations of jointed goatgrass (***Aegilops cylindrica***) and its hybrids with wheat in Oregon wheat fields. Weed Science 50: 737-747.

Morrison LA, Riera-Liazrazu O, Cremieux L, Mallory-Smith CA (2002b) Jointed goatgrass (***Aegilops cylindrica*** host) x wheat (***Triticum aestivum*** L.) hybrids: Hybridization dynamics in oregon wheat fields. Crop Science 42(6): 1863-1872.

Schoenenberger N, Felber F, Savova-Bianchi D, Guadagnuolo R (2005) Introgression of wheat DNA markers from A, B and D genomes in early generation progeny of ***Aegilops cylindrica*** Host x ***Triticum aestivum*** L. hybrids. Theoretical And Applied Genetics 111(7): 1338-1346.

Schoenenberger N, Guadagnuolo R, Savova-Bianchi D, Kupfer P, Felber F (2006) Molecular analysis, cytogenetics and fertility of introgression lines from transgenic wheat to ***Aegilops cylindrica*** host. Genetics 174(4): 2061-2070.

Snyder JR, Mallory-Smith CA, Balter S, Hansen JL, Zemetra RS (2000) Seed production on Triticum aestivum by ***Aegilops cylindrica*** hybrids in the field. Weed Science 48(5): 588-593.

Stoger E, Williams S, Keen D, Christou P (1998) Molecular characteristics of transgenic wheat and the effect on transgene expression. Transgenic Research 7: 463-471.

Stoger E, Williams S, Christou P, Down RE, Gatehouse JA (1999) Expression of the insecticidal lectin from snowdrop (***Galanthus nivalis*** agglutinin; GNA) in transgenic wheat plants: effects on predation by the grain aphid ***Sitobion avenae***. Molecular Breeding 5: 65-73.

van Slageren MW (1994) Wild wheats: A monograph of ***Aegilops*** L.and ***Ambylopyrum*** (Jaub and Spach) Eig (***Poaceae***). Wageningen: Agricultural Univ., Wageningen, The Netherlands and Int. Cent. for Agric. Res. in Dry Areas, Aleppo, Syria.

Wang ZN, Zemetra RS, Hanson BD, Mallory-Smith CA (2001) The fertility of wheat x jointed goatgrass hybrid and its backcross progenies. Weed Science 49: 340-345.

Wang ZN, Hang A, Hansen J, Burton C, Mallory-Smith CA et al. (2000) Visualization of A- and B-genome chromosomes in wheat (***Triticum aestivum*** L.) x jointed goatgrass (***Aegilops cylindrica*** Host) backcross progenies. Genome 43(6): 1038-1044.

Zemetra RS, Hansen J, Mallory-Smith CA (1998) Potential for gene transfer between wheat (***Triticum aestivum***) and jointed goatgrass (***Aegilops cylindrica***). Weed Science 46(3): 313-317.

Conclusions

The present PhD thesis brought new aspects regarding the occurrence of wheat-to-***Aegilops*** gene flow. It focused on Mediterranean species that had been neglected from previous risk assessment studies and showed the importance of human activities in promoting sympatry with wheat cultivations and dispersing ***Aegilops*** plants at a global scale. In addition, all species were not equally involved and only case by case studies will properly circumvent risks of transgene escape at a regional scale.

The present study highlighted gene flow from wheat to ***Aegilops***. Although our results need to be confirmed with direct methods (e.g. such as the sequencing of putatively introgressed markers), they outline that cultivation of transgenic wheat will have to consider the possibility of transgene(s) escape to several ***Aegilops*** species. The consequences of this escape still need to be evaluated. Three scenarios can be proposed. The first scenario considers a transgene selectively neutral in introgressed individuals of ***Aegilops*** species. This figure case would be the "best-case" scenario since the transgene does not impact the evolutive history of ***Aegilops.*** This scenario would however not prevent the spread of the transgene within ***Aegilops*** populations (e.g. via genetic drift) and questions of transgene containment would remain of concern. The second scenario considers beneficial effects of the transgene in introgressed ***Aegilops*** samples. Possible outcomes are agronomical and ecological problems with the emergence of new weeds. This scenario is realistic, since at least two ***Aegilops*** species (***Ae. cylindrica*** and ***Ae. triuncialis***) are currently considered as major weeds in the United States of America. The first example, ***Ae. cylindrica***, is interesting. The plant spread in conventional wheat cultivations by contaminating seed stocks and used the regional connections set up by human activities to colonize most of the American continent. As a result, this species is a major weed reported in Mexico, North America and Canada and causes important economical losses by competing with conventional wheat cultivars. Since this plant is a close relative of wheat, its control with herbicides is not possible without compromising the harvest. The development of herbicide-resistant transgenic wheats is not a better solution since both selection mechanisms and wheat-to-***Aegilops*** gene flow would trigger the emergence of resistant ***Aegilops cylindrica*** weeds. The only control method that seems to be efficient is crop rotations. While this system is set up in Europe for ages, its application in infested areas of the United States would surely require a complete rethinking of the American agricultural system. ***Aegilops triuncialis*** provides another interesting example. Here again, this Eurasian plant used human-driven dispersal to reach the western Coast of the United States. In this new area, ***Ae. triuncialis*** invaded vegetal communities on serpentine soils and took advantage of sheep pastoralism to increase its colonization range. Efficient adaptations against herbivory made of this plant a noxious weed that is avoided by the cattle and therefore degrades the quality of pastures. Beyond this agronomical problem, ***Ae. triuncialis*** appears to be a serious competitor as it forms dense stands and quickly covers large areas while excluding the native flora. Only herbicide based methods were proposed to control its expansion, a strategy that can be compromised if this weed sets up herbicide resistances via selection mechanisms or obtains them through gene flow with transgenic

wheat cultivars. These two examples clearly demonstrate the potential for some ***Aegilops*** species to become serious weeds and highlight the importance of containment strategies in case of transgenic wheat cultivations. The third scenario considers detrimental effects of the transgene. This scenario might arise with the use of containment techniques based on fitness-mitigation strategies (***i.e.*** the transgenic construct is beneficial within a cultivation context, but detrimental in a natural environment). As a result, the introgression of transgenes to wild relatives would be lowered since introgressed plants perform less than their wild parents. This strategy might however lead to the disappearance of the wild species from the agricultural landscapes. This situation would correspond to a hybrid depression system, where the species that provide the lowest reproduction effort are eliminated from contact zones (i.e. set up of a tension zone between non-compatible species). As a result, cultivations of transgenic wheat would probably exclude ***Aegilops*** species (at least the most allogamous ones) from agroecosystems.

The three scenario clearly demonstrate that containment strategies of transgenes should be considered in the case of transgenic wheat cultivation. The problem is worldwide; sympatry between wheat and ***Aegilops*** occurs in the complete Mediterranean basin, in the Middle East and more recently on the American continent.

Perspectives

1. The biological aspects of the crop-to-wild gene flow in the ***Aegilops*-*Triticum*** complex were first investigated by checking the mating system of ***Ae. geniculata***, ***Ae. neglecta*** and ***Ae. triuncialis***. Two approaches were used to address the question. First, based on AFLPs, estimates of genetic diversity confirmed the autogamy of ***Ae. geniculata***, while autogamy in ***Ae. neglecta*** and ***Ae. triuncialis*** appeared to be more facultative. The second approach, based on an outcrossing experiment suggested that allo-pollen could enhance the fertility (in terms of seeds production) in ***Ae. triuncialis*** and possibly ***Ae. neglecta***. Further research on this topic includes the use of codominant markers to provide proper diversity estimations. For instance, microsatellite loci should allow the direct estimation of heterozygosities, a key statistic when investigating the mating system.

2. In a second approach, we used indirect methods to detect wheat-to-***Aegilops*** gene flow. Our investigation provided convincing indications that wheat genetic material had effectively been transferred to ***Aegilops*** species. Interestingly, ***Aegilops*** samples collected along field borders contained more wheat-specific AFLP bands than reference samples collected in isolated stations. This pattern was especially clear in ***Ae. neglecta*** and ***Ae. triuncialis***, while ***Ae. geniculata*** showed no clear evidences of having experienced crop-to-wild gene flow. In addition, the tetraploid durum wheat (***Triticum tugidum***) appeared to be the most likely wheat parent of introgressed samples. Future research on this topic:
   a) Confirm the results. We confirmed these patterns in a more recent investigation focused on Andalousian (Spain) populations of ***Ae. triuncialis***[1].
   b) Rule out several alternative explanations to our AFLP results.
      - First, biases due to the anonymous status of AFLP markers should be checked. For instance, the amplification of transposable elements that are mistakenly considered as being wheat specific bands could potentially lead to erroneous conclusions. This bias is of concern since homoplasy cannot be ruled out from AFLP profiles. In this context, the worst case scenario would be the existence of ecological variants in ***Aegilops*** species that are specific to agrosystems and weakly repressing the multiplication of transposable elements in their genome. As a result, different and misleading levels of homoplasy would occur between samples collected in agrosystems or in isolated area.
      - Provide morphologic measures for putatively introgressed samples in order to detect phenotypic variations supporting the hypothesis of gene flow from wheat.
      - Use non anonymous markers. Putatively introgressed samples will be analyzed a minimum of ten to 15 SSR loci to possibly detect wheat alleles, at least in the most introgressed samples. The SSR loci will be selected from markers developed in wheat and then transferred to ***Aegilops***. We will focus on SSRs specific to the A genome. Two reason

[1] Lappe et al. (in prep)

explain this choice: (I) this genome is ancestral in wheat species and occurs in tetraploid and hexaploid wheat; (II) homoelogous pairings observed in ***Aegilops*** x ***Triticum*** hybrids systematically involved chromosomes from the A genome with chromosomes of the U, M or C ***Aegilops*** genomes[2].

- Ultimately, use of Ultra High Throughput Sequencing (e.g. 454 sequencing) to provide definitive conclusions about wheat-to-***Aegilops*** gene flow.

c) The present study considered results obtained with conventional wheat varieties. These results were reasonably transposable to transgenic wheat varieties since the transfer of wheat transgenes to ***Aegilops*** genomes was shown to be possible, at least experimentally. The main issues that remain to be addressed are the circumstances under which a transgene might get fixed in an ***Aegilops*** population. To this respect, fitness based experiments should help to assess the evolutive advantages / drawbacks of a transgene expressed in ***Aegilops*** genomes. These experimentations are now under process through a NFP 59 project. In parallel, lessons learned from natural situations where conventional wheat lines set up long-term introgression systems might also provide valuable insights. For instance, populations where long-term introgression naturally occurs might allow the identification, via 454 sequencing, of wheat genes that get fixed in ***Aegilops*** genomes. The evolutionary significance of these genes should then be assessed to generate solid theoretical bases and further develop models to handle transgenic figure cases.

3. In a way to understand the influence of human activities on the genetic diversity of ***Aegilops*** species; we investigated the phylogeography of ***Ae. geniculata***. This species showed biogeographic patterns that reflected ancient influences such as post-glaciary recolonization patterns (on the Northern Mediterranean coasts) or possibly the spread of agriculture or trade relations in Northern Africa (Southern Mediterranean coasts). Interestingly, the study revealed that recent migration events had clearly disturbed the genetic diversity of ***Ae. geniculata***. Interestingly, these exchanges allowed the observation of recombination events that involved migrant and native samples; a result that mitigated the autogamy status of ***Ae. geniculata***. While this study provided interesting results regarding ***Ae. geniculata***, it clearly showed that human activities such as trade, agriculture and pastoralism acted as efficient dispersers in the Mediterranean landscape. At a different scale, the arrival of weedy ***Aegilops*** species in the United States of America was probably caused by grain trade. Indeed, ***Ae. cylindrica*** is now a major pest of wheat fields on the complete North American continent while ***Ae. geniculata*** and ***Ae. triuncialis*** are progressively invading road sides and pastures of California. Further investigations will include samples from the United States of America, in a way to identify the Mediterranean origins of the weedy genotypes observed in California. This question, partially addressed for ***Ae. cylindrica***[3] and ***Ae. triuncialis***[4] should be completed by the collection of samples in California and Washington states.

[2] Cifuentes et al (2006)
[3] Schoenenberger (2005)
[4] Meimberg et al. (2006)