UNIVERSITÉ DE NEUCHÂTEL
Institut de biologie

Genome evolution and mechanisms underlying reproductive isolation in the polyploid *Biscutella laevigata*

Thèse présentée à la Faculté des Sciences
Institut de Biologie
Université de Neuchâtel

Pour l'obtention du grade de docteur ès Sciences de la vie (Ph.D.)

Par

Céline Geiser

Présentée aux membres du jury:
PD Dr. Christian Parisod, directeur de these, Université de Neuchâtel
Prof. Félix Kessler, co-directeur de these, Université de Neuchâtel
Prof. Yvonne Willi, rapporteur, Université de Neuchâtel
Dr. Stéphane Joost, rapporteur, Ecole Polytechnique Fédérale Lausanne

Soutenance réalisée le 04 juillet 2014

Université de Neuchâtel
2014

UNIVERSITÉ DE NEUCHÂTEL

Faculté des sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel - Suisse
Tél: + 41 (0)32 718 2100
E-mail: secretariat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Madame Céline GEISER

Titre:

"Genome evolution and mechanisms underlying reproductive isolation in the polyploid *Biscutella laevigata*"

sur le rapport des membres du jury composé comme suit:

- PD Dr Christian Parisod, Université de Neuchâtel, directeur de thèse
- Prof. Felix Kessler, co-directeur de thèse, Université de Neuchâtel
- Prof. Yvonne Willi, Université de Neuchâtel
- Dr Stéphane Joost, EPF Lausanne

Neuchâtel, le 26 août 2014

Le Doyen, Prof. B. Colbois

B. Colbois

Key words : speciation, reproductive isolation, polyploidy, duplicated genes, next generation sequencing

Mot clés : spéciation, isolement reproductif, polyploïdie, gènes dupliqués, technologies de séquençage profond

Résumé

Malgré des avancées phénoménales en biologie évolutive, les mécanismes qui mènent à la spéciation, le berceau de la biodiversité, ne sont toujours pas complètement déchiffrés. La polyploïdie, c'est-à-dire la multiplication du patrimoine chromosomique (par exemple doublé pour la tétraploïdie), est un phénomène récurrent, en particulier chez les plantes. Toutes les plantes à fleurs actuelles, avec leur incroyable diversité, ont connu au moins deux événements de polyploïdie dans leur histoire évolutive. Malgré cela il n'a jamais été possible de formellement associer la polyploïdie à des capacités adaptives accrues et son rôle dans la diversification des espèces reste controversé. Un mécanisme qui pourrait amener à des nouveautés chez le polyploïde est lié à de nouvelles fonctions émergentes des gènes dupliqués. C'est à ce mécanisme que je m'intéresse dans la première partie de ma thèse. En recomposant des séquences transcrites grâce aux nouvelles technologies de séquençage profond j'ai étudié les différents événements de polyploïdie qui ont influencé l'évolution d'une espèce alpine, la lunetière lisse tétraploïde (*Biscutella laevigata*). En analysant les forces sélectives agissant sur les gènes dupliqués j'ai pu déterminer des groupes fonctionnels potentiellement importants dans l'évolution de cette espèce. D'une part une partie des groupes fonctionnels préférentiellement retenus en deux copies, après un événement de polyploïdie récent, sont liés à l'écologie. D'autre part des copies de gènes sous sélection positive, c'est-à-dire probablement impliqués dans un processus de diversification, ont

été mis en évidence. Ces gènes sont liés au cytosquelette et pourraient être associés à une adaptation à la polyploïdie ou au stress.

Dans la deuxième partie de ma thèse je me suis intéressée aux divers mécanismes impliqués dans la spéciation commençante entre deux groupes d'une même population alpine de lunetière lisse tétraploïde (*Biscutella laevigata*). L'étude de la spéciation consiste à étudier les mécanismes qui conduisent à une cessation de flux de gènes entre deux groupes taxonomiques, c'est-à-dire l'isolement reproductif. Pour déchiffrer les divers mécanismes qui sous-tendent l'isolement reproductif, les suivis en population naturelle ainsi que des approches expérimentales sur le terrain se sont avérés être des outils essentiels. Tout d'abord la stucture génétique de cette population a révélé une zone hybride entre deux groupes. J'ai pu démontrer qu'une barrière principale qui sous-tend la divergence génétique est un décalage de floraison entre les deux groupes. Ces deux groupes semblent être adaptés à deux environnements distincts ce qui pourrait participer à l'isolement reproductif. Néanmoins, il n'a pas été possible de démontrer l'adaptation locale expérimentalement.

Contents

Acknowledgments

This work was carried out at the Faculty of Science, Laboratoire de Botanique Évolutive at the University of Neuchâtel. I thank Christian Parisod for his supervision as well as Felix Kessler and Yvonne Willi for comments on the manuscript. I thank my mi-thesis committee, Christian Lexer, Richard Buggs, François Felber and Stéphane Joost for useful comments toward the last part of my PhD. I thank the Swiss National Science Foundation and the Velux Stiftung for founding this project. I also want to thank Christiane Bobilier and the Doctoral School of Neuchâtel for providing the opportunity for advanced training. Interactions with many colleagues of the Laboratory of Evolutionary Botany helped shaping my thoughts on this project thanks to all of them. Thanks to Romain Bessire, for his constant availability in providing support on various informatics issues. I'm very grateful to Stefan Zoller and the GDC team (Genetic Diversity Center, Zürich) who supported me in my first experience with servers and bioinformatics. Without their availability and enthusiasm this would not have been possible. I also thank Nils Arrigo for his enthusiasm and support provided in bioinformatics and R. I also want to thank Amélie Bardil for the interesting discussions about the project and thank her for the help and support during the long days spent cleaning the bioinformatic files. I would like to thank the whole team of the Rochers-de-Naye for their collaboration and help during fieldwork, especially for providing a warm yourte and warm coffee during rainy and frosty fieldwork days. I would also thank the team of the botanical garden « La Rambertia » for the interesting discussions and provided information. Thanks to Kevin Leempoel and Audrey Ueberschlag who took and post-processed the GPS points. Thanks to Kevin Leempoel and Stéphane Joost for the regular meetings and useful comments on the project and the provided data. Thanks to Benjamin Dauphin for his collaboration on the environmental dataset. A great thank to the whole team of Jardin Botanique in Neuchâtel, beside being an amazing beautiful and warm place to work etc., the support in taking care of the plants and advises in botany were essential. My warm thanks go to my colleagues and friends Olivier Bachmann and Philippa Griffin. Thanks to Olivier who shared my office, for his help and support throughout this three years. Thanks to Philippa for her fulfilling help and advises. I wish also to thank past members of the Evobot team. Especially Nina who gave me a warm welcome in her office at the beginning of my PhD and introduced me to the University of Neuchâtel. A warm thank to my friends and family who accepted my lack of availability during these three years, and offered precious support. Finally, a very warm thank to my partner Tobias, who accepted that bioinformatics and fieldwork shared our weekends and hindered plenty holidays during the past three years; thanks for his help, flexibility during the fieldwork and the constant support.

Céline Geiser, June 10th 2014

General introduction

Context and research questions

In spite of decades and great advances in evolutionary biology, major questions remain unanswered regarding the drivers of speciation and the implications of polyploidy in evolution. In particular, knowledge about the interplay of diverse types of reproductive barriers and the role of natural selection in the speciation process remain incomplete. In addition, the role of polyploidy in diversification is a topic of much current debate. This thesis has been devoted to provide first evidence about the mechanisms underlying diversification inside polyploid taxa and how polyploidy could promote diversification.

- Speciation, where do we come from and where are we now?

A fundamental question in evolutionary biology is what causes speciation and biodiversity (Darwin 1859; Coyne and Orr 1998, 2004). With the expansion of literature about speciation, the language to describe mechanisms underlying speciation as well as its definition are multifarious and debated (Hausdorf 2011; Harrison 2012). Here I address speciation as being a process that starts with genetic divergence between two lineages eventually leading to reduced gene exchange, i.e. reproductive isolation toward speciation.

Darwin was the first to put forward gradual changes driven by natural selection as the main mechanism of evolution in «the Origin of Species» (Darwin 1859). He advocated that natural selection optimized traits in regard to the local environment, which would ultimately lead to speciation. However, one puzzling question remained

with Darwin's theory of speciation: How could a disadvantageous trait such as hybrid sterility evolve under natural selection? Dobzhansky (Dobzhansky 1936, 1937) partially answered this question nearly one century later explaining the importance of genetics (i.e. postzygotic genic incompatibilities) leading to hybrid breakdown as a mechanism of reproductive isolation. These ideas, together with the biological species concept (Mayr 1942), have formed the main theoretical framework for the speciation process.

For a long time evolutionary biologists distinguished between speciation scenarios based on discrete geographic categories, e.g. sympatric and parapatric speciation (Coyne and Orr 2004; Fitzpatrick et al. 2009). The former scenario is that reproductive isolation occurs on a local scale in spite of some gene flow. The latter scenario is that species evolve in geographic isolation and accumulate genetic differences due to selection and drift potentially leading to reproductive isolation if they come into contact again. Sympatric speciation was long debated because the homogenizing effect of gene flow operates in opposition to the build-up of divergence (Smajda 2011). Even so, sympatric speciation is much more common than previously assumed (Coyne and Orr 2004). Nowadays studies often emphasize on the processes promoting divergence (e.g. pre- and postzygotic isolation) and how they are affected by geography rather than an arbitrary classification into sympatric or parapatric (Fitzpatrick et al. 2009).

Reproductive barriers can act at different stages of the life cycle and are split in two main categories: whether they act before or after zygote formation, i.e. pre- and postzygotic (Widmer et al. 2009). Prezygotic isolation prevents the formation of hybrid zygotes through barriers of various types, e.g. ecogeographic (i.e. reproductive isolation due to limited contact between individuals of the same population/species), temporal

(e.g. shifts in flowering time), and gametic incompatibility (Nosil et al. 2005; Rieseberg and Willis 2007; Lowry et al. 2008). Postzygotic reproductive isolation includes all mechanisms that lead to fitness aberration in hybrids: necrosis, weakness, sterility and lethality but can also include immigrant inviability. The majority of empirical studies showed that prezygotic barriers make a greater overall contribution to reproductive isolation than later acting postzygotic barriers (Lowry et al. 2008; Widmer et al. 2009). But often, an interplay of several barriers leads to strong reproductive isolation (Lowry et al. 2008). In addition, there is growing appreciation of the effect of divergent selection by the environment (postzygotic extrinsic barrier) as a driver toward speciation, i.e. ecological speciation (Nosil 2012).

A major goal of evolutionary biology is to quantify the relative contribution of pre- and/or postzygotic barriers, assess the implication of natural selection and understand the genetic mechanisms underlying reproductive isolation (Widmer et al. 2009; Sobel et al. 2010; Nosil 2012).

- The implications of polyploidy on diversification and speciation

Whole genome duplications, i.e. polyploidy, shaped genome evolution throughout eukaryotes (Lynch and Conery 2000; Otto and Whitton 2000; Leitch and Leitch 2008; Jiao et al. 2011; Abbott et al. 2013). This is especially true for plants given that all angiosperms have experienced at least two whole genome duplications (Jiao et al. 2011). Even plants that are nowadays genetically diploid did undergo past paleopolyploidization events, including the model organism *Arabidopsis thaliana* with its small genome (Blanc and Wolfe 2004). The omnipresence of polyploidy in plants led to a debate about the influence of whole genome duplication in diversification and in the

long-term success of polyploids (Otto and Whitton 2000).

Polyploidy can alter the physiology and morphology of organisms in a few generations, and it was postulated that polyploids are good colonizers and often found in harsh or unstable environments, e.g. at higher latitudes/altitudes or during periods of climatic change (Lewis 1980; Stebbins 1985; Otto and Whitton 2000; Leitch and Leitch 2008; Fawcett et al. 2009; Vanneste et al. 2014). But what could be the reason for their success and long-term persistence? How could polyploidy mediate ecological diversification?

A first explanation comes from the fact that certain environmental conditions favor polyploidization. A common pathway to polyploidy is an error during meiosis that produces unreduced gametes (Comai 2005). Several experimental studies showed increased unreduced gamete production (2n) following harsh and stressful conditions, e.g. at coastal sites vs. in the greenhouse, under temperature differences or nutrient stress (reviewed in Ramsey and Schemske 1998). Even so, production of polyploids under stressful conditions does not explain their establishment and success in the long-term.

When few polyploids are formed inside a diploid population they have to become established. This establishment is severely limited by reproductive isolation between diploids and polyploids because the pollen comes mainly from the diploids and hinders the production of viable or fertile polyploid progeny (Comai 2005). To explain their establishment several hypotheses were put forward ranging from niche shifts, differences in niche patterns (variance in the niche) until different dispersal ranges between polyploids and their diploid progenitors (Glennon et al. 2014). The hypothesis about higher adaptability and broader niche occupancy of polyploids has recently been

challenged because it could not be demonstrated that niches between diploids and polyploids differ systematically more than the niches between homoploids (Leitch et al. 2012).

Three main phenotypic changes conferred by polyploidy could possibly be involved in their successful establishment: differences in cell size and growth rate (e.g. larger seeds relative to diploids), effects on size and shape (larger overall size, e.g. used in crop plants) and effects on reproductive systems (e.g. shift to asexual reproduction, (Otto and Whitton 2000). What is the role of the genetic redundancy caused by polyploidization underlying these changes? A potential disadvantage is the predicted inefficiency of selection because of multiple alleles (Otto and Whitton 2000; Parisod et al. 2010). Even so, the functional redundancy of duplicated genes is predicted to enhance the ability to diversify one or both copies, allowing accelerated evolution. Thereby, whole genome duplications could be seen as providing raw material for evolution upon which natural selection can act (Rieseberg and Willis 2007). After whole genome duplication a large majority of duplicated genes are lost, but the remaining fraction represents an interesting situation to address the mechanisms and forces underlying gene retention and diversification (Lynch and Conery 2000).

The newest study in the field has just found that angiosperm paleopolyploidization across 41 plant genomes was clustered around the Cretaceous-Paleocene boundary, suggesting that paleopolyploidization could somehow have contributed to plant radiation and survival at that time (Vanneste et al. 2014). In the future it will be necessary to address differences in niches occupancy between polyploids and diploids to assess causality. It is also crucial to infer the forces behind the maintenance of duplicated gene copies after whole genome duplications and assess their

implication in creating evolutionary novelties (McGrath and Lynch 2012). Finally, to understand the causal relation between polyploidy and plant radiations, it is important to assess the timing between whole genome duplications and radiations (Schranz et al. 2012).

Speciation by polyploidization events is well known; changes in ploidy level are responsible for 15% of angiosperm speciation events (Wood, 2009). While changes in ploidy promote speciation, homoploid speciation, where hybrids sharing the same polyploidy level as their parents form a new species is considered rare (Abbott et al. 2010). Polyploids are classified in two major categories where allopolyploids are formed by the merging of interspecific genomes, and autopolyploids by the merging of intraspecific genomes. Allopolyploids are characterized by disomic inheritance and autopolyploids by tertrasomic inheritance, i.e. polysomic when more than 4x or more (Gallais 2003). Under disomic inheritance, chromosomes have pairing preferences and genes behave as duplicates with two segregating alleles each, whereas no pairing preference exists under tertrasomic inheritance and each locus has four segregating alleles.

After polyploidization the genome undergoes diploidization, i.e. gene loss, eventually leading to full disomic inheritance even in autopolyploids (Bowers et al. 2003). Therefore, the rate of tetrasomic inheritance (vs. disomic) can vary inside autopolyploid lineages, but also inside the polyploid genome (Bowers et al. 2003; Meirmans and Van Tienderen 2013). Tetrasomic populations are theoretically characterized by higher heterozygozity and by nearly doubled effective population size as compared to disomic ones, however their response to selection is predicted to be lower (Ronfort et al. 1998; Parisod et al. 2010). Even so, successful range expansion

after autopolyploidy was demonstrated as well as adaptation, putatively sustaining divergence, in autotetraploid lineages (Manton 1937; Gasser 1986). In general, many questions remain about the ecological consequences of polyploidy and its role in diversification (Weiss-Schneeweiss et al. 2013). This is especially true for autopolyploids since little is known about the role of whole genome duplications on diversification and the adaptive potential of autopolyploids (Bretagnolle et al. 1998; Leitch and Leitch 2008).

Study system

Figure 1: Tetraploid *Biscutella laevigata* in its habitat in the Swiss Alps. Top right pictures: on the left side an inflorescence and on the right side the typical didymous fruits, resembling a pair of glasses.

The Mediterranean *Biscutella* genus contains about 40 species that are easily identifiable by their typical didymous fruits resembling glasses (Figure 1, Olowokudejo 1986). The genus *Biscutella* can be divided into two clades amongst others. The first consists of mostly annual species with a basic chromosome number of n=8 (Manton 1932; Olowokudejo and Heywood 1984; Olowokudejo 1986). The second clade (series *laevigatae*) is composed of perennial species with a basic chromosome number n= 9 (one exception with n=8) and contains polyploids and diploids (Manton 1932; Schönfelder 1968). *Biscutella laevigata* is thought to have survived the Pleistocene as a diploid and recolonized the Alps as an autotetraploid after the end of the ice ages through polytopic autopolyploidy, with recolonization by independent lineages (Manton

1937; Tremetsberger et al. 2002; Parisod and Besnard 2007; Abbott et al. 2013). *Biscutella laevigata* L. is thereby the only species in the *Biscutella* genus that colonized the Alps.

The polyploid *Biscutella laevigata* L. (en: Buckler Mustard, de: Brillenschötchen, fr: lunetière lisse; Brassicaceae) is a long-lived perennial species with a sporophytic self-incompatibility system and generalist pollinators (Olowokudejo and Heywood 1984; Parisod and Bonvin 2008; Parisod and Christin 2008). Therefore, *Biscutella laevigata* represents an ideal non-model species to investigate i) the history of multiple past paleopolyploidization events in a non-model organism, ii) the putative role of polyploidy in the colonization of the Alps, and iii) diversification between homoploid lineages and the implication of natural selection in an alpine species.

Study population

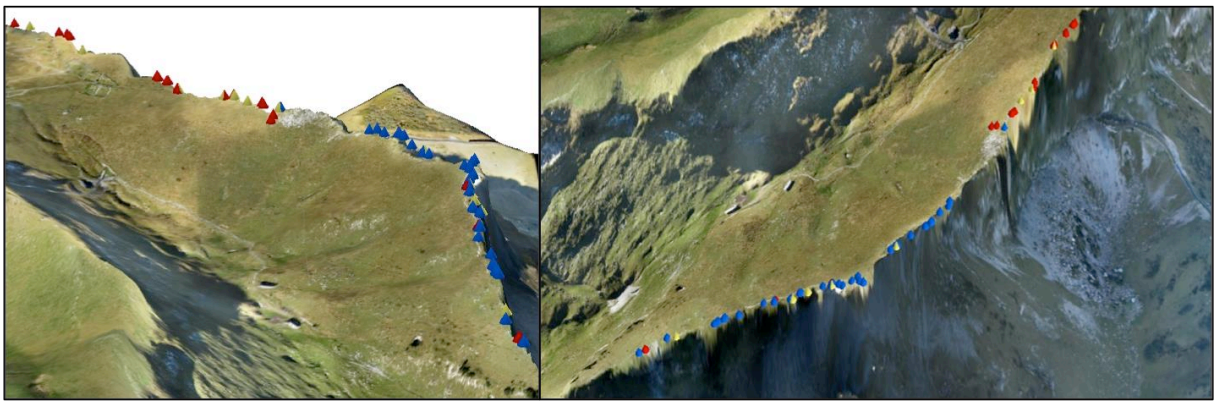


Figure 2: Study population in the Rochers-de-Naye. The picture on the left shows an eastern view of the whole ridge area. Triangles indicate occurrence of two lineages (red and blue, hybrids in yellow) of *Biscutella laevigata* according to the sampling in this PhD. The picture on the right shows a northwestern view of the central region of the ridge; the area to its left is dominated by grassland and the area to the right by a steep cliff. Aerial picture underlying the map by Kevin Leempoel.

In this PhD I studied a tetraploid population of *Biscutella laevigata* L. in the Swiss Alps (Rochers-de-Naye, N46°26'00'', E6°58'50''). The population is mostly distributed along a 1.2 km transect on the top of a ridge (Figure 2) and its altitude ranges from 1864-2043m. For the study in Chapter 2, I sampled 348 individuals and surveyed 212 individuals of this population from 2011-2013. The population is divided in two genetic groups with few hybrids (red and blue individuals in Figure 2, hybrids in yellow).

In a previous study, 124 individuals of this population were investigated (Parisod and Christin 2008). They hypothesized that gene flow should homogenize the genetic variability over such a small scale. In contrast, the authors assessed "the factors responsible for the repartition of the genetic variance and revealed a composite effect of isolation by distance, phenological divergence and local adaptation to habitats characterized by different solar radiation regimes" (Parisod and Christin 2008). In addition, a genome scan analysis on the same dataset revealed that 7% of loci were

putatively under divergent selection and significantly correlated to environmental variables, mainly solar radiation (Parisod and Joost 2010).

Accordingly, strong evolutionary forces seem to have maintained incipient speciation in spite of gene flow. The different isolating barriers underlying this divergence as well as the implication of local adaptation were addressed here with an extensive sampling and survey as well as a reciprocal transplant experiment in nature.

This thesis

This PhD thesis is divided in two main chapters and addresses the incidence of past polyploidy events in shaping the evolution of the *Biscutella* genus and mechanisms underlying reproductive isolation between two divergent genetic groups of the tetraploid *Biscutella laevigata*.

Genome duplication has long been debated as being a promoter of diversification and adaptive evolution. However, little is known about the forces underlying retention and functional gene categories preferentially retained. In Chapter 1, I assembled the first transcriptome for the tetraploid *Biscutella laevigata*. This transcriptome allowed inferring past paleopolyploidy events shaping the evolution of *Biscutella*. Two of the four assessed paleopolyploidy events were compared: one that was shared with *Arabidopsis thaliana* and one that was a more recent, lineage-specific event. I performed functional analyses of differentially retained gene categories after these events and the forces that may have triggered the retention of these groups. These analyses allowed first insights in the retention dynamics after whole genome duplication and the forces underlying the maintenance of duplicated genes.

In Chapter 2, I focused on a tetraploid population of *Biscutella laevigata* in the Swiss Alps to identify reproductive isolation barriers between tetraploid natural lineages. An extensive field survey at local spatial scale highlighted a hybrid zone between two genetic groups of *Biscutella laevigata* and offered indirect evidence of mechanisms that could have shaped this observed genetic differentiation. In the second part, these indirect observations were experimentally tested in a reciprocal transplant experiment on F1 de-novo hybrid individuals and F1 parental individuals.

Literature Cited

Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. W. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, A. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, A. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Most, S. Mullen, R. Nichols, A. W. Nolte, C. Parisod, K. Pfennig, A. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Vainola, J. B. W. Wolf, and D. Zinner. 2013. Hybridization and speciation. Journal of Evolutionary Biology 26:229-246.

Abbott, R. J., M. J. Hegarty, S. J. Hiscock, and A. C. Brennan. 2010. Homoploid hybrid speciation in action. Taxon 59:1375-1386.

Blanc, G. and K. H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667-1678.

Bowers, J. E., B. A. Chapman, J. K. Rong, and A. H. Paterson. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433-438.

Bretagnolle, F., F. Felber, F. G. Calame, and P. Kupfer. 1998. Polyploidy in plants. Botanica Helvetica 108:5-37.

Comai, L. 2005. The advantages and disadvantages of being polyploid. Nature Review Genetics 6:836-846.

Coyne, J. A. and H. A. Orr. 1998. The evolutionary genetics of speciation. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 353:287-305.

Coyne, J. A. and H. A. Orr. 2004. Speciation. Sinauer, Sunderland, MA.

Darwin, C. 1859. On the Origin of Species by Means of Natural Selection or the Preservation of Favored Races in the Struggle for Life, J. Murray, London.

Dobzhansky, T. 1936. Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics 21:113-135.

Dobzhansky, T. 1937. Genetic nature of species differences. American Naturalist 71:404-420.

Fawcett, J. A., S. Maere, and Y. Van de Peer. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event.

Proc Natl Acad Sci U S A 106:5737-5742.

Fitzpatrick, B. M., J. A. Fordyce, and S. Gavrilets. 2009. Pattern, process and geographic modes of speciation. Journal of Evolutionary Biology 22:2342-2347.

Gallais, A. 2003. Quantitative Genetics and Breeding Methods in Autopolyploid Plants, Editions Quae.

Gasser, M. 1986. Genetic-ecological investigations in *Biscutella levigata* L. Veroeffentlichungen des Geobotanischen Institutes der Eidgenoessische Technische Hochschule Stiftung Ruebel in Zuerich:7-87.

Glennon, K. L., M. E. Ritchie, and K. A. Segraves. 2014. Evidence for shared broad-scale climatic niches of diploid and polyploid plants. Ecology Letters 17:574-582.

Harrison, R. G. 2012. The language of speciation. Evolution 66:3643-3657.

Hausdorf, B. 2011. Progress toward a general species concept. Evolution 65:923-931.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, Y. Hu, H. Liang, P. S. Soltis, D. E. Soltis, S. W. Clifton, S. E. Schlarbaum, S. C. Schuster, H. Ma, J. Leebens-Mack, and C. W. dePamphilis. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97-U113.

Leitch, A. R. and I. J. Leitch. 2008. Perspective - Genomic plasticity and the diversity of polyploid plants. Science 320:481-483.

Leitch, I. J., J. Greiluber, J. Dolezel, and W. J. F. 2012. Plant Genome Diversity. Springer.

Lewis, W. H. 1980. Polyploidy in species populations. Lewis, W. H. (Ed.). Basic Life Sciences, Vol. 13. Polyploidy: Biological Relevance; Proceedings, St. Louis, Mo., USA, May 24-27, 1979. Xii+583p. Plenum Press: New York, N.Y., USA; London, England. Illus. Maps:P103-144.

Lowry, D. B., J. L. Modliszewski, K. M. Wright, C. A. Wu, and J. H. Willis. 2008. The strength and genetic basis of reproductive isolating barriers in flowering plants. Philos T R Soc B 363:3009-3021.

Lynch, M. and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Manton, I. 1932. Introduction to the general cytology of the Cruciferae. Annals of Botany 46:509-556.

Manton, I. 1937. The problem of *Biscutella laevigata* L. II. The evidence from meiosis. Annals of Botany 1:439-U414.

Mayr, E. 1942. Systematics and the Origin of Species, Columbia Univ Press, New York.

McGrath, C. L. and M. Lynch. 2012. Evolutionary significance of whole-genome duplication *in* Springer, ed. Polyploidy and Genome Evolution.

Meirmans, P. G. and P. H. Van Tienderen. 2013. The effects of inheritance in tetraploids on genetic diversity and population divergence. Heredity 110:131-137.

Nosil, P. 2012. Ecological Speciation. Oxford Series in Ecology and Evolution, Oxford Series in Ecology and Evolution.

Nosil, P., T. H. Vines, and D. J. Funk. 2005. Perspective: Reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution 59:705-719.

Olowokudejo, J. D. 1986. The infrageneric classification of *Biscutella* (Cruciferae). Brittonia 38:86-88.

Olowokudejo, J. D. and V. H. Heywood. 1984. Cyto-taxonomy and breeding system of the genus *Biscutella* (Cruciferae). Plant Systematics and Evolution 145:291-309.

Otto, S. P. and J. Whitton. 2000. Polyploid incidence and evolution. Annu Rev Genet 34:401-437.

Parisod, C. and G. Besnard. 2007. Glacial in situ survival in the Western Alps and polytopic autopolyploidy in *Biscutella laevigata* L. (Brassicaceae). Molecular Ecology 16:2755-2767.

Parisod, C. and G. Bonvin. 2008. Fine-scale genetic structure and marginal processes in an expanding population of *Biscutella laevigata* L. (Brassicaceae). Heredity 101:536-542.

Parisod, C. and P. A. Christin. 2008. Genome-wide association to fine-scale ecological heterogeneity within a continuous population of *Biscutella laevigata* (Brassicaceae). New Phytologist 178:436-447.

Parisod, C., R. Holderegger, and C. Brochmann. 2010. Evolutionary consequences of autopolyploidy. New Phytologist 186:5-17.

Parisod, C. and S. Joost. 2010. Divergent selection in trailing- versus leading-edge populations of Biscutella laevigata. Annals of Botany 105:655-660.

Ramsey, J. and D. W. Schemske. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annual Review of Ecology and Systematics 29:467-501.

Rieseberg, L. H. and J. H. Willis. 2007. Plant speciation. Science 317:910-914.

Ronfort, J. L., E. Jenczewski, T. Bataillon, and F. Rousset. 1998. Analysis of population structure in autotetraploid species. Genetics 150:921-930.

Schönfelder, P. 1968. Chromozomenzahlen einiger Arten der Gattung *Biscutella* L. Österreiches Botanikal Zeitung 115:363-371.

Schranz, M. E., S. Mohammadin, and P. P. Edger. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Current Opinion in Plant Biology 15:147-153.

Smajda, B. 2011. A framework for comparing processes of speciation in the presence of gene flow. Molecular Ecology 20:5123-5140.

Sobel, J. M., G. F. Chen, L. R. Watt, and D. W. Schemske. 2010. The biology of speciation. Evolution 64:295-315.

Stebbins, G. L. 1985. Polyploidy, hybridization and the invasion of new habitats. Annals of the Missouri Botanical Garden 72:824-832.

Tremetsberger, K., C. Konig, R. Samuel, W. Pinsker, and T. F. Stuessy. 2002. Infraspecific genetic variation in *Biscutella laevigata* (Brassicaceae): new focus on Irene Manton's hypothesis. Plant Systematics and Evolution 233:163-181.

Vanneste, K., G. Baele, S. Maere, and Y. Van de Peer. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Research.

Weiss-Schneeweiss, H., K. Emadzade, T. S. Jang, and G. M. Schneeweiss. 2013. Evolutionary Consequences, Constraints and Potential of Polyploidy in Plants. Cytogenetic and Genome Research 140:137-150.

Widmer, A., C. Lexer, and S. Cozzolino. 2009. Evolution of reproductive isolation in plants. Heredity 102:31-38.

Abbreviation list

WGD : whole genome duplication

At-β : whole genome duplication in *Arabidopsis thaliana* post-dating the Caricaceae-Brassicaceae divergence

At-α : whole genome duplication in *Arabidopsis thaliana* specific for core Brassicaceae

B. l. : *Biscutella laevigata*

A. t. : *Arabidopsis thaliana*

K_a: nonsynonymous substitution rate

K_s: synonymous substitution rate

EST: expressed sequence tags

SOLiD: *sequencing* by oligonucleotide ligation and detection, rather short reads

454: large-scale parallel pyrosequencing, rather long reads

MIRA: Assembly algorithm

bp : base pair

TAIR10 database: Most recent database for the model organism *Arabidopsis thaliana*

α-WGD: Whole genome duplication in *B. laevigata* putatively corresponding to the At-α event

B-WGD: Whole genome duplication in *B. laevigata* putatively lineage specific

DEM: digital elevation model

AFLP: amplified fragment length polymorphism

GLM: generalized linearized model

GLMM: generalized linearized mixed model

AIC : Aikaike Information criterion

Chapter 1: Paleopolyploidy and gene retention patterns of the transcriptome of *Biscutella laevigata*

Abstract

Whole genome duplication, i.e. paleopolyploidy, is a prominent feature of flowering plants. The long-term success of polyploids and their role in plant evolution has long been debated. Recently, it could be demonstrated that successful genome duplications in 41 plants could be associated with the Cretaceous-Paleogene boundary. Even so, it is still not solved how and if genetic redundancy created during whole genome duplications can promote adaptability and diversification. One key process for generating novelties, thereby producing the raw material for adaptive evolution, is imputed to gene duplications. Producing redundancy, gene duplications allow new mutations to arise, and copies eventually to diverge into novel functions. In this study I assembled a *de novo* transcriptome as a snapshot of the genome for the non-model species *Biscutella laevigata*. These new resources allowed inferring past paleopolyploidy events in the *Biscutella* genus. Assessing the evolutionary forces underlying retained genes allowed first confirming previous studies that showed that the majority of paralogous gene pairs evolve under purifying selection. Second, functional categories could by highlighted as playing a putatively important role in the adaptive evolution of the genus and are discussed in the context of the selective forces they seem to underlie.

Introduction

Whole genome duplication (WGD), also termed polyploidy, shaped genome evolution throughout eukaryotes (Lynch and Conery 2000; Otto and Whitton 2000; Zhang 2003; Leitch and Leitch 2008; McGrath and Lynch 2012). Whole genome duplications are prominent in plants, especially in angiosperms (Wood et al. 2009). Indeed, a polyploidization event occurred before the divergence of seed plants and another before the flowering plant radiation implying that all flowering plants are paleopolyploids (Jiao et al. 2011). The early analysis of *Arabidopsis thaliana* allowed revealing the existence of two paleopolyploidy events after the angiosperm radiation: At-β and At-α (Simillion et al. 2002; Bowers et al. 2003; Franzke et al. 2011). While the older At-β post-dates the Caricaceae-Brassicaceae divergence, the most recent At-α is specific for core Brassicaceae (Franzke et al. 2011) and coincides with the important radiation of the core Brassicaceae (Franzke et al. 2011). A central question remains in linking and explaining the repeated coincidence of whole genome duplications with cladogenesis and ecological diversification (Schranz et al. 2012; Vanneste et al. 2014).

Following whole genome duplications, most gene duplicates are lost at an initially high rate by pseudogenization (Lynch and Conery 2000; Zhang 2003). The remaining fraction of copies might be retained and might survive long after whole genome duplications following different fates (Blanc and Wolfe 2004). Multiple models have been proposed to explain the fate of duplicated genes pairs (i.e. paralogs) including: neofunctionalization, i.e. a new function evolves; subfunctionalization, i.e. each daughter copy retains a part of the functionality or both copies evolve new functionalities; and dosage balance (Zhang 2003; McGrath and Lynch 2012). Even so, the evolutionary mechanisms underlying the maintenance of gene duplicates remains elusive, and

specific selection pressures acting on diverging duplicate pairs or on one duplicate, depending on the time elapsed since whole genome duplication, were predicted (Innan and Kondrashov 2010). It is often assumed that duplication *per se* is selectively neutral, because of the functional redundancy of gene copies; one copy is freed from purifying selection (Lynch and Conery 2000). This relaxed selection can cause an accumulation of new mutations potentially leading to the evolution of a new function. It was demonstrated that a major part of recently diverged paralogs were under similar purifying selection across the tree of life (e.g. *Achaea, Saccharomyces cerevisiae, Arabidopsis thaliana, Drosophila Melanogaster*, (Kondrashov 2012). The ability of accelerated evolution after whole genome duplications is thought to be a primary source of evolutionary novelties through neo- or subfunctionalization (Lynch and Conery 2000). However, the forces underlying gene retention, to which extent gene duplications underlie diversification and adaptive evolution, as well as the timing between whole genome duplication and radiations remain to be addressed (McGrath and Lynch 2012; Schranz et al. 2012).

Different molecular signatures can be used as diagnostic to reveal early stages of functional divergence between paralogs. The ratio of nonsynonymous to synonymous substitution rates (i.e. K_a/K_s) between two sequences allows deducing selection pressure and a putative molecular adaptation (Yang and Bielawski 2000). Thus, a K_a/K_s-ratio of one describes a neutral amino acid change becoming fixed at the same rate as a synonymous mutation. If the amino acid change is deleterious, purifying selection ($K_a<K_s$) will reduce its fixation. If this change offers a selective advantage it will be fixed at higher rate $K_a>K_s$), indicating positive selection.

The investigation of molecular adaptations in the context of whole genome

duplications is facilitated due to the increasing availability of genetic sequences. Under the assumption that synonymous substitutions are accumulated at a stochastic rate proportional to time, K_s measures the relative age of gene pairs (Lynch and Conery 2000; Vanneste et al. 2013). During a WGD (i.e. polyploidization events) all genes are duplicated, therefore a WGD can be detected as a simultaneous enrichment of pairs in a K_s range representing the whole genome duplication. These enrichment peaks are expected to be approximately Gaussian (Blanc and Wolfe 2004). For non-model organisms lacking complete genomic sequences, expressed sequence tags (EST) may be used to infer paleopolyploidy (Zhang and Hewitt 2003; Blanc and Wolfe 2004; Barker et al. 2009).

It was demonstrated that a burst of polyploidy events was associated with changing environments during the Cretaceous-Tertiary extinction (Fawcett et al. 2009; Vanneste et al. 2014). During angiosperm evolution, with its multiple polyploidy events, groups of genes involved in reproduction and flower development were preferentially retained (Jiao et al. 2011). Thus, assessing genes preferentially retained after whole genome duplications and their link to ecology and radiations is of major interest.

Our study group, *Biscutella*, arose after the At-α WGD during the radiation of core Brassicaceae (Franzke et al. 2011). The genus *Biscutella* can be divided in two clades, the first mainly consisting of annual species with a basic chromosome number of n = 8 (Manton 1932; Olowokudejo and Heywood 1984; Olowokudejo 1986). The second clade (series *laevigatae*) is composed of perennial species with a basic chromosome number of n = 9 and includes diploids as well a polyploids (Manton 1932; Schönfelder 1968). *Biscutella laevigata* is a highly variable species containing three diploid and one tetraploid evolutionary group (Tremetsberger et al. 2002). *Biscutella laevigata* is

thought to have survived the Pleistocene as a diploid and recolonized the Alps as an autotetraploid after the end of the ice ages through polytopic autopolyploidy, with recolonization by independent lineages (Manton 1937; Tremetsberger et al. 2002; Parisod and Besnard 2007). Therefore, it was speculated that polyploidy could play a role in the success of tetraploids over diploids in alpine habitats (Tremetsberger et al. 2002). Therefore, *Biscutella laevigata* represents an interesting species to assess the putative implication of whole genome duplications in diversification and colonization of the Alps.

To tackle these questions, I first assembled a *de novo* transcriptome for the polyploid *Biscutella laevigata* to infer paleopolyploidy events shaping the evolution of the *Biscutella* genus. The comparison between the At-α WGD, shared with *Arabidopsis thaliana*, and the more recent *Biscutella* specific B-WGD highlighted differences in proportions of retained duplicates under purifying vs. positive selection. Further, a functional bias of retained genes under specific selection pressures in both WGDs was assessed in more details to assess what gene categories underwent a period of positive or purifying selection.

Material and Methods

Transcriptome assembly and annotation

For non-model organisms ESTs (expressed sequence tags) represent an interesting starting point, to address whole-genome questions in evolutionary biology i.e. a snapshot from the genome (Bouck and Vision 2007). Transcriptome assemblies in model organisms with an available reference genome are usually performed by technologies producing massively short reads (e.g. SOLiD, Sequencing by Oligonucleotide Ligation and Detection) with high throughput and high accuracy. However, in the absence of a reference, longer reads (e.g. Roche 454) are often preferred since long reads are more amenable to *de novo* assembly (Parchman et al. 2010). Here, the advantage of high throughput SOLiD and longer 454 reads are combined in an assembly with MIRA in the hybrid mode (Chevreux et al. 2004). The hybrid mode allows assembling long and short reads in the very same assembly step. The alternative would be to assemble first long, then short reads and combine both at the end. The hybrid mode can improve the assembly because it allows to control for the coverage of reads needed to build a contig (overlapping reads that represents a consensus region) depending on the sequencing technology (Chevreux et al. 2004).

Leaf and root tissues were collected in liquid nitrogen from four individuals of *B. laevigata* from a well-described population (Parisod and Christin 2008). Plants were previously grown for one year in a common garden (Jardin Botanique, Neuchâtel, Switzerland). Total RNA was extracted using PureLink RNA Mini Kit (Life Technologies, Zug, Switzerland), following the manufacturer's instructions, and its quality was checked on an Agilent Bioanalyzer. RNAs from leaf and root tissues from one individual were

prepared. A 454 normalized library was prepared on pooled root and leaf RNAs from one individual, shotgunned and single-end sequenced on a quarter of a plate using Roche GS FLX Titanium Series (service provided by Microsynth AG, Balgach, Switzerland, following manufacturer's instructions). Six quantitative single-end cDNA libraries (four individuals; two with pooled root and leaf; two with root and leaf separated) were prepared and barcoded for SOLiD on 6 lanes of SOLiD 5500 xl (service provided by Microsynth, Balgach, Switzerland, in collaboration with Vertis Biotechnology, Freising, Germany, following manufacturer's instructions).

Libraries were trimmed for vectors, quality and GC bias, low complexity and size using PRINSEQ (Schmieder and Edwards 2011) as well as FASTX-Toolkit (http://hannonlan.cshl.edu/fastx_toolkit). After trimming, perfect duplicates, i.e. perfectly identical sequences, were removed and the transcriptome assembly was performed using the hybrid mode of MIRA 3.4 (Chevreux et al. 2004). Due to the high memory requirements of this algorithm, computational challenges were overcome by performing the assembly through six iterative steps. In the first step long reads (454) were assembled with a first library of short reads (SOLiD) to obtain a first set of contigs and singletons. A contig is a consensus sequence of overlapping reads originated from set of long (454) and short (SOLiD) reads. This definition is an oversimplification of the modern assembly algorithm, e.g. MIRA includes over 100 options to form a contig (Chevreux et al. 2004). A MIRA-contig is assembled according to complex algorithms combining information about e.g. coverage, drawbacks of each sequencing technology, quality information, of each library and read. These options can be found in the mira manual (http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html). The log-files (more than 10Gb of text files for each assembly step) cannot be shown here

due to their length and size. Accordingly, a singleton is a read that could not be included in a contig because it did not fulfill the assembly criteria of MIRA. Contigs assembled in this first step were combined with the 454 library to represent a new set of long reads to be assembled with the second short reads library. Thereby, previously assembled contigs represent long reads and were used in the next assembly steps to be elongated. After the second step (and in all following steps) singletons were not only unused single short reads but also large contigs that were not elongated further in the specific assembly step. Therefore in step three to six, these "large singletons" were included in the long read input dataset to optimize the procedure. The long reed datasets were further composed by: 454 long reads, contigs newly assembled in the previous step and contigs assembled in all previous steps (not elongated in the previous step therefore filtered out as singleton by the algorithm). This iterative procedure was repeated until all SOLiD short reads libraries were included. In the last step assembled contigs and singletons were combined, contigs shorter that 150bp were discarded as well as perfect duplicates to produce the final set of transcripts (i.e. transcriptome).

Annotation of the transcriptome

The annotation of the transcriptome was performed with BLASTx against TAIR10 peptide database (Altschul et al. 1990). By removing redundant annotations (i.e. assembled sequences with the same *Arabidopsis* gene ID) the number of unigenes, relative to known *Arabidopsis thaliana* genes, was assessed. To ensure genome-wide coverage by the transcriptome, the unigenes were assigned to Brassicaceae building blocks. In Brassicaceae genomes, including the model organism *Arabidopsis thaliana,* 24 genomic blocks (called A-X) are known to be conserved across species (Schranz et al.

2006). This subdivision allowed assigning each transcript to one of the blocks according to its gene ID from *A. thaliana* (block coordinates according to Schranz et al. 2006) with a custom script (supplementary material S 3). The proportion of transcripts in *Biscutella laevigata* relative to all known genes in *Arabidopsis thaliana* was reported for each building block separately. To assess homogeneity of coverage inside each building block, the proportion of transcripts in *Biscutella laevigata* relative to all known genes in *Arabidopsis thaliana* was analyzed within sliding windows of 40 genes (sliding by 10) using the zoo package (Zeileis 2014), along the genes coordinates accounting for synteny with *Arabidopsis thaliana*.

Inference of whole genome duplication

To obtain a valid set of EST (expressed sequence tags), allowing testing for past whole genome duplications, reading frames were identified and excised through comparison with the TAIR10 database using Exonerate (Slater and Birney 2005), custom scripts by Nils Arrigo. Mitochondrial and plastid genes as well as transposable element containing genes were previously filtered out. Framed contigs showing higher than 30% sequence similarity over at least 450bp were clustered in gene families by single linkage clustering (custom script by Nils Arrigo). Transcripts from clusters with more than two members were aligned using the codon-aware MACSE algorithm (Ranwez et al. 2011). Corresponding phylogenetic trees were inferred with FastTree (Price et al. 2009) to infer the pairwise synonymous and nonsynonymous substitution rate (K_s and K_a, respectively) in PAML using the F3X4 model (Yang 1997). The custom pipeline was elaborated in collaboration with Stefan Zoller and automatized over families as a service by the GDC (Genetic Diversity Center, ETH Zürich). Identical pairs of contigs (i.e. $K_s = 0$)

were removed; these pairs represent identical sequences that were not removed by the clustering algorithm. These pairs were detected at the end of the analysis, but the alignment was done including these redundant sequences and consequently they are included in multiple pairs. Therefore, every pair containing the shortest of these identical sequences was removed by hand in the whole dataset to avoid a bias due to redundancy.

Polyploidy events yield an enrichment of pairs in the K_s-range, which allows estimating the relative age of gene pairs, corresponding to a given whole genome duplication because all genes are duplicated simultaneously (Blanc and Wolfe 2004). This peaks are expected to be approximately Gaussian (Blanc and Wolfe 2004). The number and position of peaks in the distribution of K_s histogram of pairs including sequences larger than 450 bp, was inferred with Gaussian mixture models using the R cran (Benaglia et al. 2009; Development Core Team R 2013; Team 2013). Only pairs of contigs with $K_s < 2$ were analyzed since mixture modeling inference of WGDs proved less reliable for larger synonymous distances (Vanneste et al. 2013). A parametric bootstrap analysis (100 bootstraps) tested the most likely number of peaks, the presence of one to eight peaks was tested (Benaglia et al. 2009). In this test 100 bootstrap realizations of the likelihood ratio statistics are produced to test the null hypothesis that k-components vs. (k+1)-components (here from one to eight components tested). Thereby, the best model, i.e. the best number of Gaussian peaks is retained.

Functionally biased retention

To tackle the forces underlying gene duplications, duplicated pairs (i.e. pairs in a gene

family) putatively originating in the α- and B-WGD were compared according to their K_a/K_s ratio. Pairs were assigned to a whole genome duplication according to their K_s values. For this dataset, the nonsynonymous rate (K_a) and synonymous substitution (K_s) rate between paralogs was extracted from the PAML results and the K_a/K_s ratio was calculated (Yang 1997). A first analysis focused on the distribution of the K_a/K_s ratio in the α- and B-WGD. A two-sided non-parametric proportion test was used to assess significant differences in proportion between both events.

To tackle the functional categories differentially retained according to their K_a/K_s ratio, corresponding gene lists were split in seven categories for the α- and B-WGD respectively: pairs showing signs of strong purifying selection (i.e. K_a/K_s of 0-0.3, 0.3-0.6 or 0.6-0.9), neutral divergence (i.e. K_a/K_s of 0.9-1.1) or positive selection (i.e. K_a/K_s of 1.1-2, 2-4 or 4-12). For this issue, a gene ID list (according to *Arabidopsis* annotations) for framed contigs larger than 150 bp was used. The annotation with *Arabidopsis* gene IDs generated plenty of redundancy, i.e. several different pairs annotated with a single *Arabidopsis* gene ID. Due to uncertainties linked to the reason underlying this redundancy, e.g. multiple retention, multicopy family, sequencing or assembly errors; the analysis of functional annotations focused on unigene lists. Even when several IDs are used as input, only one will be used in the comparison. Therefore, for all functional analyses, the number of different gene IDs is reported as it represents the values used in the statistical analysis, not the number of pairs. To increase the chances that a retained pair originated in a given duplication the quantiles (5% and 95%) were calculated. Accordingly pairs were assumed to have originated in the α-WGD when K_s values ranged from 0.41 to 0.9 and from the B-WGD when K_s values ranged from 0.085 to 0.37. Each of these 14 gene lists was analyzed separately, with the transcriptome as

background dataset to test for overrepresented functional categories with FatiGo (Al-Shahrour et al. 2004). Significantly over-represented GO categories among these lists were assessed with FatiGo using adjusted p-values from Fisher's exact test, after correcting for multiple testing with the false discovery rate procedure (Benjamini and Hochberg 1995; Al-Shahrour et al. 2004). To summarize the data, overrepresented categories were summarized by their GO-Slim classification (ftp://ftp.arabidopsis.org) and presented in heatmaps.

Results

Transcriptome assembly and annotation

A normalized cDNA library of 270'469 long 454 reads of *B. laevigata* was iteratively assembled with over six SOLiD libraries containing 13 millions of short SOLiD (Table 1) using an hybrid approach in MIRA (Chevreux et al. 2004). The hybrid assembly method allowed increasing the number of assembled contigs, from 404'105 after the first assembly step until over one million in the final dataset (Table 2).

Table 1: The different libraries used in the hybrid assembly. In the first column the used sequencing technology. The second column describes sequenced organs. In the third column the name of individuals sampled according to a previous study (Parisod and Christin 2008). The last column shows the number of reads used as input during the hybrid assembly.

NGS libraries	Organs	Individuals[#]	Number of reads
454	Leaves & roots	Ab	270469
SOLiD1	Leaves & roots	Ca	17123098
SOLiD2	Leaves & roots	Bj	20662107
SOLiD3	Leaves	Ab	16621814
SOLiD4	Leaves	Ab	14131769
SOLiD5	Roots	Bd	10182227
SOLiD6	Roots	Bd	35266604

[#]See(Parisod and Christin 2008)

The increase in contigs in the last library is due to the increased sequencing effort on this individual as well as pooling with singletons in this final step, this redundancy being removed in the final dataset. Also, the size of contigs was improved, the largest contig being 3692 bp long after the first step until 5282 for the largest contig in the final

dataset (Table 2). The mean coverage in long 454 reads remained constant in the different assembly steps (3.71-4.96 sequences for large contigs) whereas the coverage in short reads decreased from 24.64 to 15.83 in the final dataset (Table 2). After the last assembly step when all libraries were included in the assembly, perfect duplicates were removed as well as sequences smaller than 150bp. The final transcriptome resulted in 102'998 contigs (mean length: 413.50bp, N50: 573, for a total of 42'589'854 bp).

Table 2: Iterative hybrid assembly of libraries. The rows represent the assembly steps. In the second column the number of contigs produced in each step followed by the number of large contigs (>1000bp), the size of the largest contig and mean coverage by long and short reads.

Assembly step	Number of contigs	Number of large contigs[a]	Largest contig (bp)	Mean coverage long reads	Mean coverage short reads
1	404105	15772	3692	4.75	24.64
2	402172	20036	5402	4.31	25.29
3	333770	16874	4118	4.96	10.45
4	347677	22214	6236	4.06	18.03
5	313668	22711	4305	4.06	12.18
6	1134879	24462	5282	3.71	15.83

[a]Large contigs > 1000 bp

Annotation of the transcriptome

Annotation of 102'998 against TAIR10 peptide database revealed the high quality of this inferred transcriptome, as 72'273 (70.1%) and 64'645 (62.7%) contigs were annotated at e-values of e^{-5} and e^{-10}, respectively. These annotated genes represented 16,389 unigenes (49% of known unigenes in *Arabidopsis thaliana*). The whole-genome coverage

was warranted as all building blocks were covered (Table 3). Even so, differences could be observed between building blocks, the coverage ranging from 12.5% to 62.9% (Table 3). Whereas some building blocks were homogeneously covered (e.g. Block A Figure 1), differences could be observed inside some others, e.g. a decrease of coverage at the end of block O or Q (Figure 2).

Building Block	Genes in *A.t.*	*B.l.* transcriptome
blockA	2030	60.39%
blockB	1845	47.20%
blockC	1269	47.12%
blockD	739	39.92%
blockE	1719	60.15%
blockF	2872	59.47%
blockG	395	12.15%
blockH	652	49.85%
blockI	914	51.09%
blockJ	1937	58.85%
blockK	276	53.62%
blockL	566	40.81%
blockM	778	42.67%
blockN	1354	57.31%
blockO	692	43.21%
blockP	435	39.77%
blockQ	782	43.73%
blockR	2180	62.94%
blockS	1145	28.30%
blockT	468	47.01%
blockU	2782	60.03%
blockV	675	54.37%
blockW	1262	58.08%
blockX	775	62.06%
Mean	-	49.17%

Table 3: Frequencies of transcripts of *Biscutella laevigata* (B.l.) as compared to the known genes in *Arabidopsis thaliana* for each building block. Rows represent the Brassicaceae building blocks, the second column represents the number of known genes in *Arabidopsis thaliana* and the last column the proportion of these genes represented in the transcriptome of *Biscutella laevigata*.

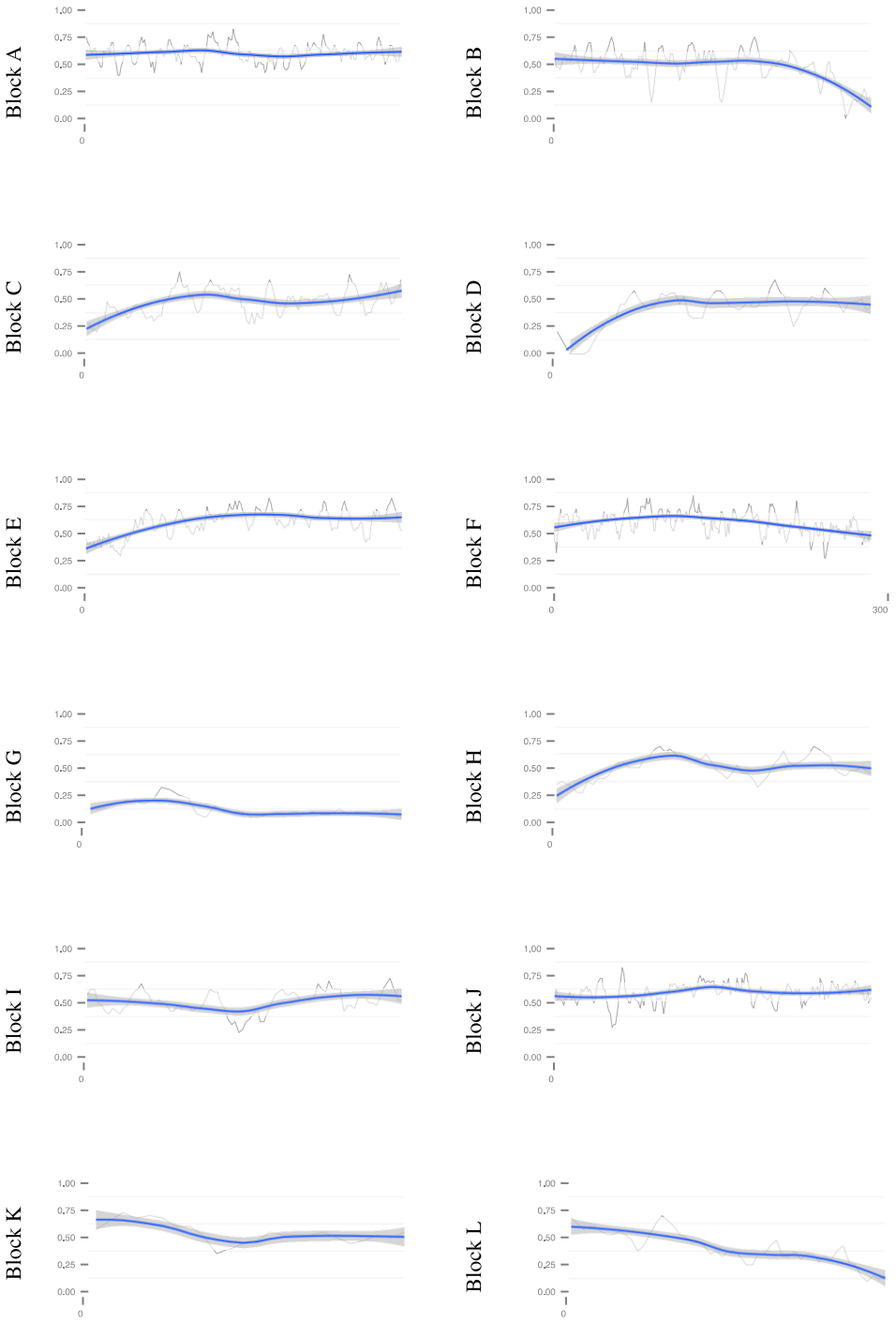


Figure 1: Unigene coverage by the transcriptome for Brassicaceae building blocks A-L. Proportion of coverage of *Biscutella* unigenes compared to knowm genes in *Arabidopsis*, in the order of their gene IDs. Proportion of assembled genes was computed in a sliding window of 40 genes.

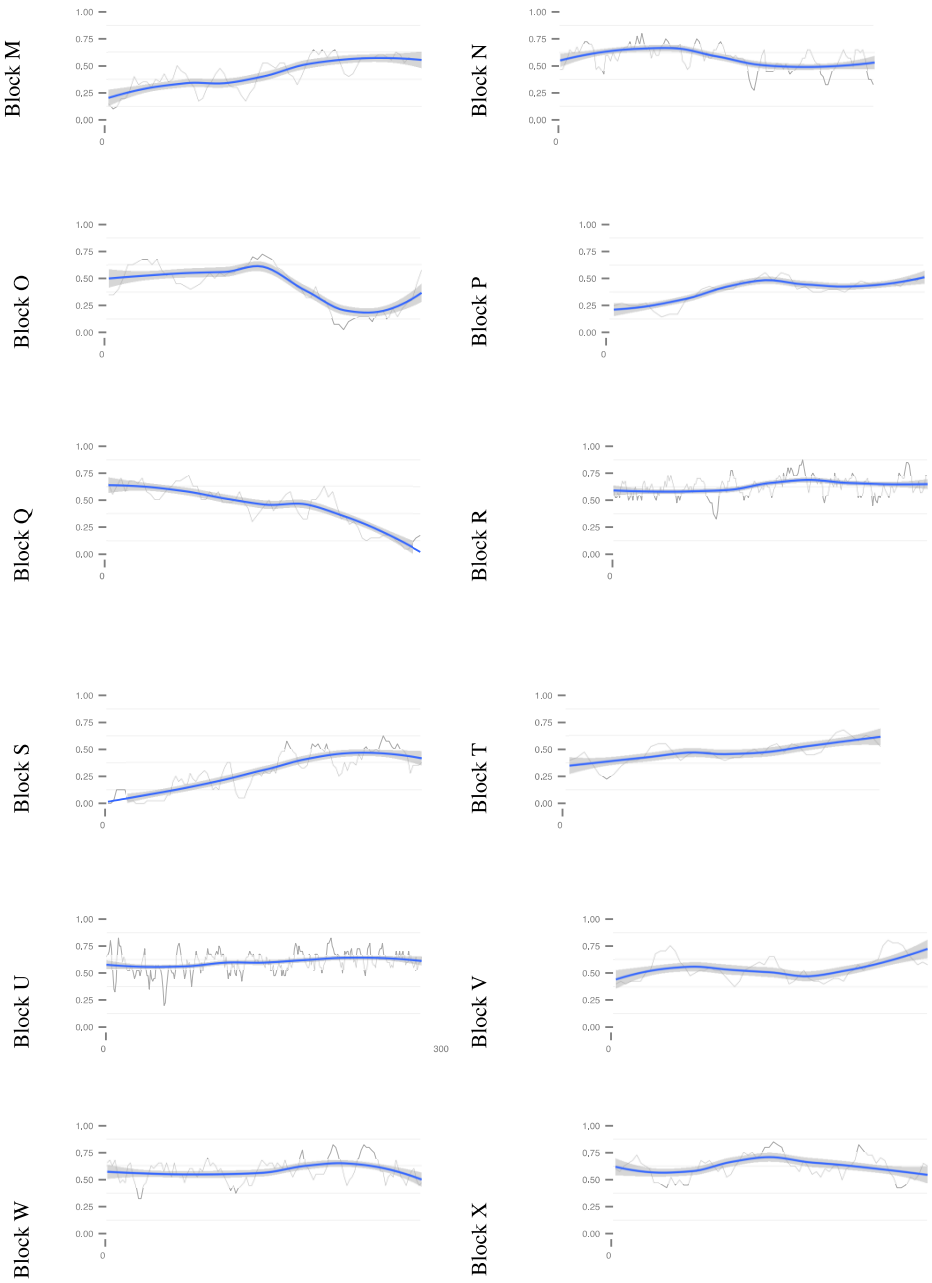


Figure 2: Unigene coverage by the transcriptome for Brassicaceae building blocks M-X. Proportion of coverage of *Biscutella* unigenes compared to known genes in *Arabidopsis*, in the order of their gene IDs. Proportion of assembled genes was computed in a sliding window of 40 genes.

Inference of whole genome duplication

To obtain a valid set of EST (expressed sequence tags), out of the 102,998 assembled contigs, 61'278 (70.2%) were translated into a reading frame. Removal of mitochondrial, plastid and transposable element containing genes allowed obtaining a dataset of 57'783 sequences for clustering. Clusters (n=3'141), i.e. gene families, were built with 26'815 contigs with at least 30% homology with other sequences, representing gene families with two to 169 members.

Enrichment peaks in the histogram of gene duplications (K_s), i.e. large scale duplications, were fitted and their significance was assessed with mixture models on a total of 20'890 paralogous gene pairs with $K_s<2$. The "expectation-maximization" algorithm converged at four peaks, whereas stringent bootstrapping supported three peaks. The oldest significant peak at mean $K_s = 1.67$ (±0.006; Table 4, Figure 3) corresponds most probably to the At-β (Table 4, Figure 3). An additional, intermediate peak was fitted at a mean $K_s = 0.89$ (±0.01; Table 4, Figure 3) corresponding to the At-α, called α-WGD here, corresponds to the event at the basis of the core Brassicaceae (Bowers et al. 2003), even so detected by the mixture model this peak was not supported by stringent bootstrap. The two youngest peaks, B-WGD with a mean $K_s = 0.23$ (±0.005; Table 4, Figure 3) and a L-WGD at a mean $K_s = 0.03$ (±0.002; Table 4, Figure 3) were seemingly characteristic to the lineage analyzed in this study.

Table 4: Whole genome duplication events. Means Ks and standard errors of inferred whole genome duplication peaks. Standard errors were performed on 100 bootstraps. "*" represents peaks supported by parametric bootstrap hypothesis testing.

WGD event	L	B	α	β
Mean	0.03* ± 0.002	0.23* ± 0.005	0.89 ± 0.01	1.67* ± 0.006
Standard deviation	0.02 ± 0.002	0.12 ± 0.004	0.36 ± 0.01	0.20 ± 0.004

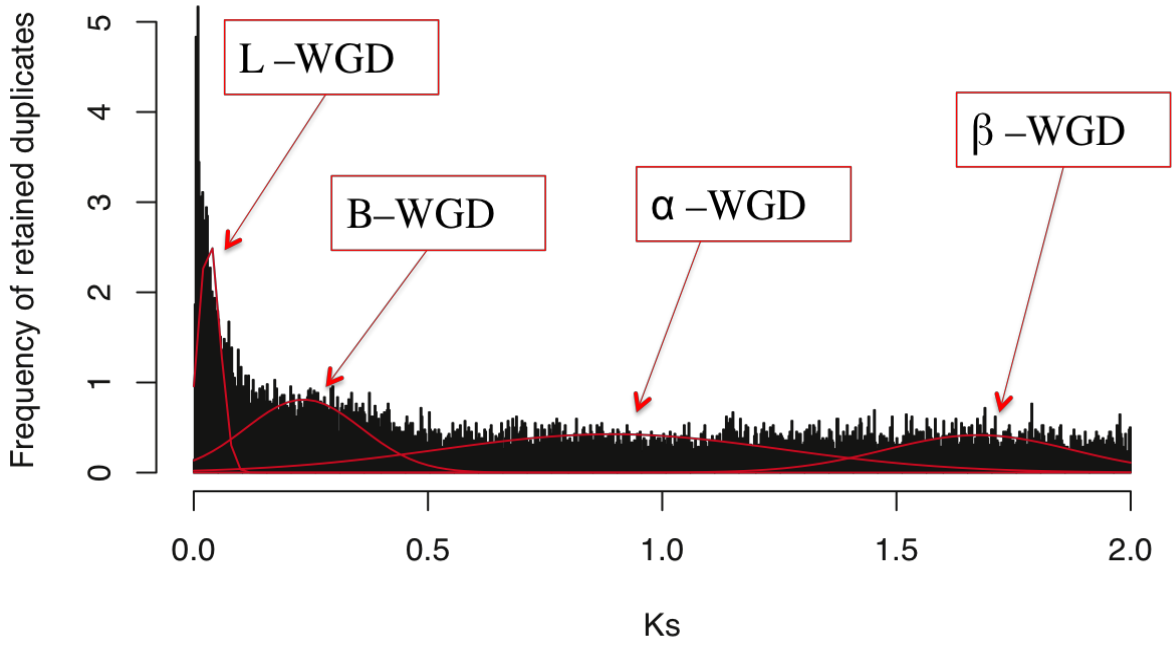


Figure 3: Histogram of Ks. The histogram corresponds to the number of pairs in a given K_s-range. The red distributions represent mixture model fits of inferred whole genome duplications. From left to right, the most recent lineage specific L-WGD, the lineage specific B-WGD studied in the following part, α-WGD at the origin of core Brassicaceae and the oldest β-WGD corresponding most probably to the At- β event.

Functionally biased retention

A major and similar proportion of pairs of duplicates originated during the α-WGD and B-WGD (n=9059 pairs, respectively 5126 and 3933 pairs) were under purifying selection (Figure 4). A significantly lower proportion (p-value < 0.01) of pairs experienced putatively positive selection after B-WGD compared to those in the α-WGD (respectively 150 and 592 pairs).

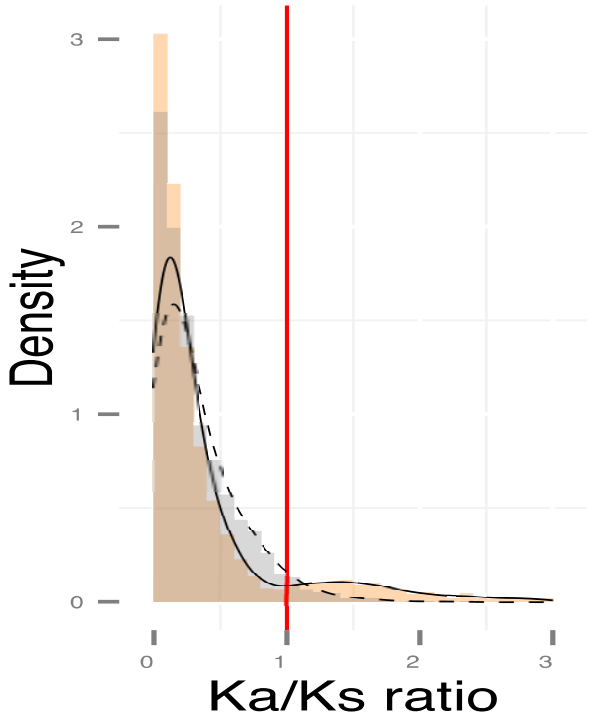


Figure 4: Histograms of Ka/Ks ratios between paralogous pairs originated during the B-WGD (grey) and α-WGD (orange). The red line represents a similar proportion of synonymous and nonsynonymous substitutions, i.e. neutral selection.

The analysis of functional categories of genes retained during the α-WGD and B-WGD in seven categories showed important difference in the number of GO IDs used as input for each category ranging from four to 1188 (Table 5).

Table 5: Number of IDs used in the comparison along the magnitude of selection. The first column represents the analysed event. The second column represents the Ka/Ks ratio ranges analysed followed by the number of IDs entered as background and the number of IDs contained in each category.

Event	Ka/Ks ratio range	GO IDs background	GO IDs specific category
α	4-12	16442	10
α	2-4	16442	21
α	1.1-2	16442	54
α	0.6-0.9	16442	119
α	0.3-0.6	16442	426
α	0-0.3	16442	1188
B	4-12	16442	4
B	2-4	16442	9
B	1.1-2	16442	52
B	0.6-0.9	16442	210
B	0.3-0.6	16442	297
B	0-0.3	16442	456

A much higher proportion of GO categories retained in the α-WGD, compared to the more recent B-WGD, are under positive selection (11 out of 67 categories, bottom of Figure 5 and 6) and related to cytoskeleton organization or development. In addition, these categories were highly overrepresented, as shown by their high odd-ratio value (orange in Figure 5 and 6). The only over-represented category, under positive selection retained in the B-WGD event is cytoskeleton organization (bottom of Figure 5). Following the α-WGD, categories putatively under neutral and purifying selection represent annotations linked to e.g. response to hormones, cell death or development (Figure 5). Various categories including response to biotic and abiotic stress are under

strong purifying selection after α-WGD. A highly overrepresented second group in B-WGD, under neutral or nearly neutral selection (K_a/K_s, 0.9-1.1), represents annotations linked to photosynthesis. Responses to biotic and abiotic stimuli are under strong purifying selection following both WGD (Figure 6).

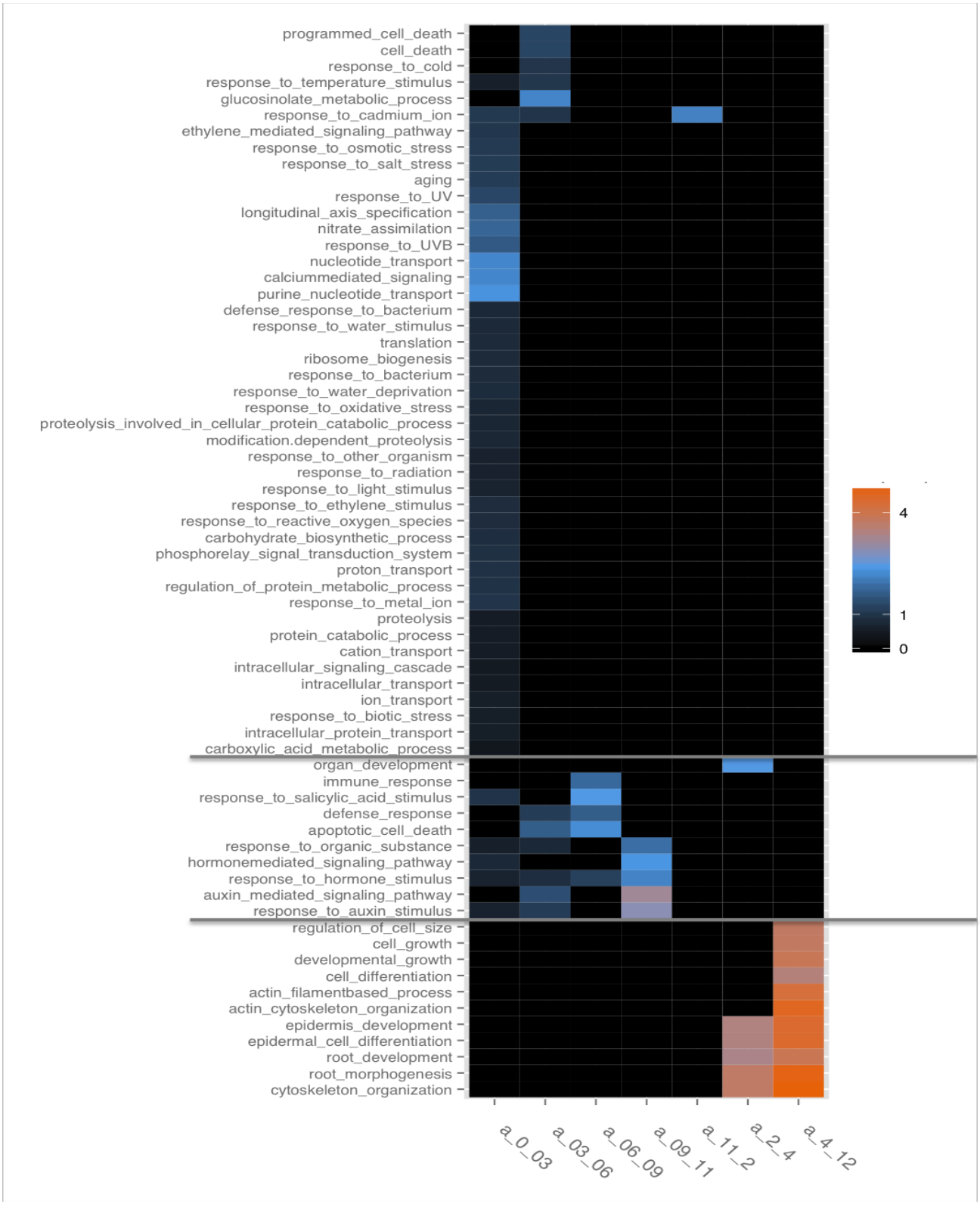


Figure 5: Over-represented GO-slim categories from retained gene duplicates having originated during the α-WGD. Categories are split according to the magnitude of selection, from left to right, Ka/Ks ratio of 0-0.3, 0.3-0.6, 0.6-0.9, 0.9-1.1, 1.1-2, 2-4, 4-12. The color range represents the odd-ratios of significantly (0.05) overrepresented GO categories. In black, categories that are not significantly over-represented.

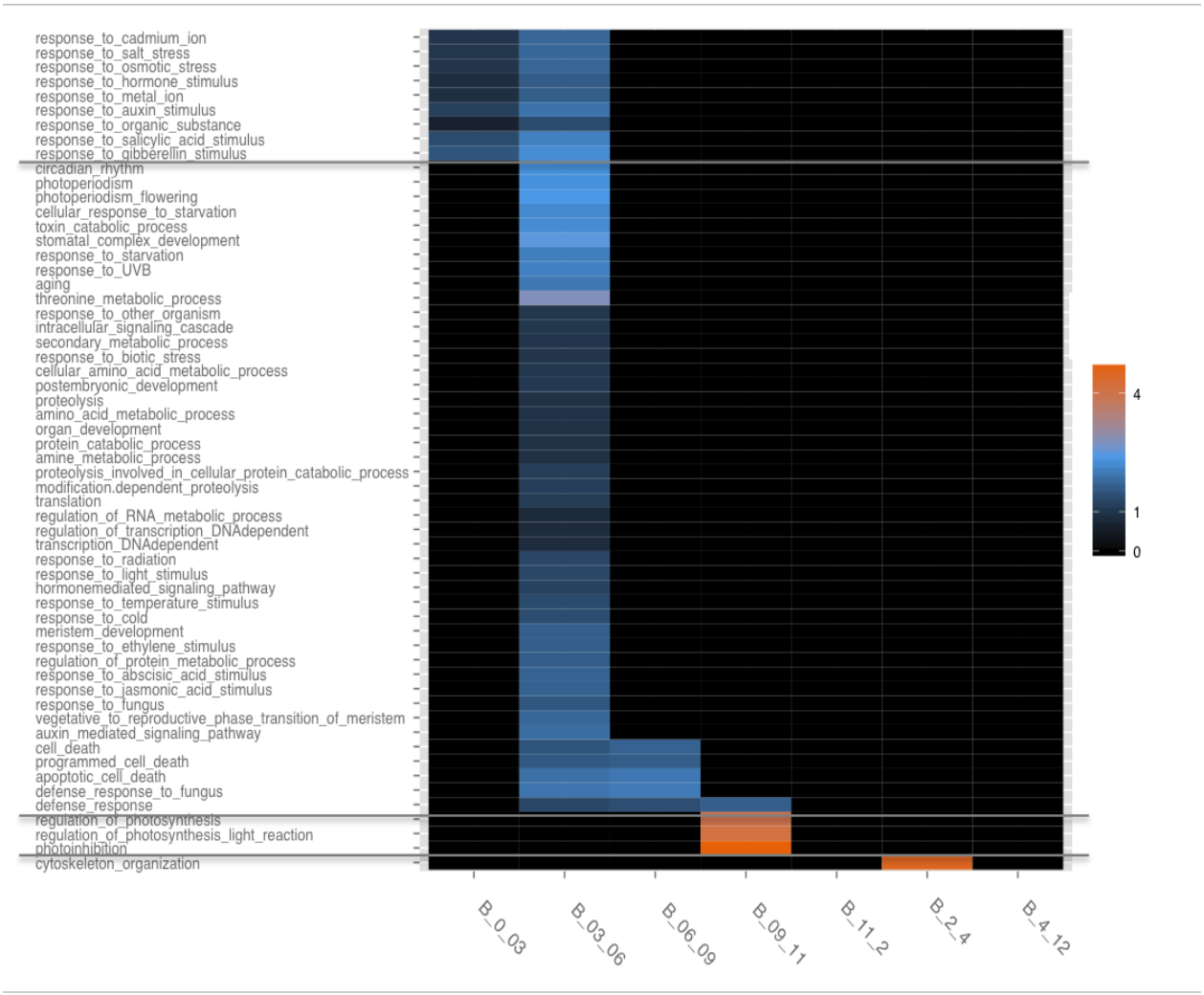


Figure 6: Over-represented GO slim categories from retained gene duplicates having originated during the B-WGD. Categories are split according to the magnitude of selection, from left to right, Ka/Ks ratio of 0-0.3, 0.3-0.6, 0.6-0.9, 0.9-1.1, 1.1-2, 2-4, 4-12. The colour range represents the odd-ratios of significantly (0.05) overrepresented GO categories. In black, categories that are not significantly over-represented.

Discussion

The *de novo* transcriptome of a non-model species, proved to be a useful snapshot of the genome to reveal past ploidy events that shaped the genome of *Biscutella laevigata*. In addition, comparative analyses of the forces of selection underlying gene duplicates between the α-WGD, shared with *Arabidopsis thaliana*, and a lineage specific B-event allowed to get first insights in the forces underlying gene retention and target ecologically relevant categories.

The transcriptome assembly combining the advantages of short and long reads allowed producing a representative set of EST of sufficient length and high quality to infer whole genome duplication from the non-model polyploid *B. laevigata.* Even so highly memory consuming, the hybrid mode in MIRA was powerful in producing long and accurate contigs. This was revealed by the quality of annotation, thus this assembly method allowed overcoming the drawbacks of single-end sequences for *de novo* assemblies.

The high accuracy annotation and framing of contigs with the *Arabidopsis thaliana* database allowed i) to use the most accurate database in plants to further retrieve high quality GO information, and ii) to use known reading frames to correct potential chimaeras and misassemblies for further analyses. By localizing the genes in the building blocks, it could be ensured that all building blocks (i.e. the whole genome) were covered by the transcriptome, even so to different extents. The lack of coverage of some building blocks or parts of blocks could be due to i) that not all organs were sequenced in the transcriptome (in particular flowers were not sequenced), ii) plants were sampled at a single life stage, iii) reorganization and gene loss specific to *Biscutella*.

Mixture models allowed identifying four peaks putatively corresponding to

paleopolyploidy events. The two oldest peaks (β-, α-) correspond to well-known paleopolyploidy events shared with A*rabidopsis thaliana*. The α-event, shared by the core-Brassicaceae was assessed and correspond to the K_s-range of different dating in literature. Therefore it is considered as the last shared whole genome duplication between *Arabidopsis thaliana* and *Biscutella laevigata*. Two significant peaks, the B- and the L-WGD, are more recent lineage specific peaks. The most recent L-WGD, probably highlighting a very recent whole genome duplication will need confirmation and was not further assessed here. Here, I focused on the lineage specific B-WGD possibly explaining the radiation of tetraploid lineages inside the *Biscutella* genus and colonization of the Alps. In addition, as the *Biscutella laevigata* contains diploids and polyploids, it could allow gaining first insights in genes retained that are implicated in this diversification.

The distributions of the K_a/K_s ratio of pairs originated in the α- and B-WGD allowed to highlight that the majority of pairs are under purifying selection. This confirms earlier studies showing that paralog pairs evolve under purifying selection (Lynch and Conery 2000; Kondrashov et al. 2002). Even so, the observation that $K_a << K_s$ for a pair is not sufficient to conclude that both paralogs evolve under purifying selection. The very same result would have been obtained if one paralog was freed from selective constraints (neutral evolution), whereas the other maintained by purifying selection. In the future, this issue should be addressed for target pairs with known orthologs, in a comparative framework.

Here it was shown that more paralogs originating in the α- WGD than in the more recent B-WGD experienced positive selection. Positive selection is a diagnostic for pair of genes that acquires novel beneficial mutations, putatively promoting novelties or/and diversification (neofunctionalization or subfunctionalization). One hypothesis

underlying the prevalence of genes under positive selection in the α-WGD would support the lag-time model. In this model, it was proposed that an amount of time is needed between a whole genome duplication and the associated radiation (Schranz et al. 2012). Therefore, this species represents an outstanding framework to assess the timing not only of gene loss but also for new beneficial mutations to arise and be subject to natural selection.

Assessing a functional bias in categories that were retained (pairs included in a gene family) highlighted that most functional categories were under (strong) purifying selection after α-WGD and B-WGD. Pairs under purifying selection after α-WGD represent various categories putatively underlying the radiation of core Brassicaceae. On the contrary, pairs under strong purifying selection after B-WGD represent specific categories of abiotic and biotic stress response. These could represent interesting categories underlying the radiation of *Biscutella* and colonization of the Alps. Several categories related to cytoskeleton organization were under positive selection after the α-WGD and one after the B-WGD. These categories are of major interest, since they represent genes that potentially evolved novelties. Two hypotheses could underlie the evolution of new functions related to cytoskeleton. First, changes in cytoskeleton proteins in roots could be associated with drought or salt stress, thereby representing a major adaptive advantage (Ghosh and Xu 2014). Second, polyploids undergoe important physiological changes in cellular architecture with increasing cell volume (Comai 2005). In this case modification in cytoskeleton proteins could represent an adaptation to polyploidy. These categories represent promising candidates for future studies, since the evolution of novel traits are the key in defining major radiations across the tree of life.

In this context, recurrent paleopolyploidy events revealed in this study represent a promising framework to refine the overall timing in the retention of genes duplicates. In addition, revealing duplicated genes in ecologically relevant categories allowed defining targets for in debt study of forces and mechanisms underlying this retention. Thereby, tackling the importance of retentions and novelties by whole genome duplications and their role in radiation and adaptation of polyploids.

Acknowledgments

I thank Stephan Zoller for his help and enthusiasm from first trimming of the data until the last steps of the pipeline. I also thank Nils Arrigo for the overall help in bioinformatics and R, from transcriptome assembly until the final functional analysis. I also want to thank Amélie Bardil for the help during the cleaning by hand of this huge dataset and her expertise in functional annotations. I thank Yvonne Willi and Felix Kessler for comments on the manuscript. The project was financed by the Swiss National Science Foundation and the Velux Stiftung.

Literature Cited

Al-Shahrour, F., R. DÌaz-Uriarte, and J. Dopazo. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20:578-580.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignement search tool Journal of Molecular Biology 215:403-410.

Barker, M. S., H. Vogel, and M. E. Schranz. 2009. Paleopolyploidy in the *Brassicales*: analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other *Brassicales*. Genome Biol Evol 1:391-399.

Benaglia, T., D. Chauveau, D. R. Hunter, and D. Young. 2009. mixtools: An R Package for Analyzing Finite Mixture Models. . Journal of Statistical Software 32:1-29.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate - A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological 57:289-300.

Blanc, G. and K. H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667-1678.

Bouck, A. and T. Vision. 2007. The molecular ecologist's guide to expressed sequence tags. Molecular Ecology 16:907-924.

Bowers, J. E., B. A. Chapman, J. K. Rong, and A. H. Paterson. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433-438.

Chevreux, B., T. Pfisterer, B. Drescher, A. J. Driesel, W. E. G. Muller, T. Wetter, and S. Suhai. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Research 14:1147-1159.

Comai, L. 2005. The advantages and disadvantages of being polyploid. Nature Review Genetics 6:836-846.

Development Core Team R. 2013. R: a language and environment for statistical computing. , Vienna Austria.

Fawcett, J. A., S. Maere, and Y. Van de Peer. 2009. Plants with double genomes might

have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A 106:5737-5742.

Franzke, A., M. A. Lysak, I. A. Al-Shehbaz, M. A. Koch, and K. Mummenhoff. 2011. Cabbage family affairs: the evolutionary history of Brassicaceae. Trends Plant Sci 16:108-116.

Ghosh, D. and J. Xu. 2014. Abiotic stress responses in plant roots: a proteomics perspective. Frontiers in Plant Science 5.

Innan, H. and F. Kondrashov. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97-108.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, Y. Hu, H. Liang, P. S. Soltis, D. E. Soltis, S. W. Clifton, S. E. Schlarbaum, S. C. Schuster, H. Ma, J. Leebens-Mack, and C. W. dePamphilis. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97-U113.

Kondrashov, F. A. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proceedings of the Royal Society B-Biological Sciences 279:5048-5057.

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Selection in the evolution of gene duplications. Genome Biology 3.

Leitch, A. R. and I. J. Leitch. 2008. Perspective - Genomic plasticity and the diversity of polyploid plants. Science 320:481-483.

Lynch, M. and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Manton, I. 1932. Introduction to the general cytology of the Cruciferae. Annals of Botany 46:509-556.

Manton, I. 1937. The problem of *Biscutella laevigata* L. II. The evidence from meiosis. Annals of Botany 1:439-U414.

McGrath, C. L. and M. Lynch. 2012. Evolutionary significance of whole-genome duplication *in* Springer, ed. Polyploidy and Genome Evolution.

Olowokudejo, J. D. 1986. The infrageneric classification of *Biscutella* (Cruciferae). Brittonia 38:86-88.

Olowokudejo, J. D. and V. H. Heywood. 1984. Cyto-taxonomy and breeding system of the genus Biscutella (Cruciferae). Plant Systematics and Evolution 145:291-309.

Otto, S. P. and J. Whitton. 2000. Polyploid incidence and evolution. Annu Rev Genet 34:401-437.

Parchman, T. L., K. S. Geist, J. A. Grahnen, C. W. Benkman, and C. A. Buerkle. 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. Bmc Genomics 11.

Parisod, C. and G. Besnard. 2007. Glacial in situ survival in the Western Alps and polytopic autopolyploidy in *Biscutella laevigata* L. (Brassicaceae). Molecular Ecology 16:2755-2767.

Parisod, C. and P. A. Christin. 2008. Genome-wide association to fine-scale ecological heterogeneity within a continuous population of *Biscutella laevigata* (Brassicaceae). New Phytologist 178:436-447.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution 26:1641-1650.

Ranwez, V., S. Harispe, F. Delsuc, and E. J. P. Douzery. 2011. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. PLoS One 6.

Schmieder, R. and R. Edwards. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863-864.

Schönfelder, P. 1968. Chromozomenzahlen einiger Arten der Gattung *Biscutella* L. Österreiches Botanikal Zeitung 115:363-371.

Schranz, M. E., M. A. Lysak, and T. Mitchell-Olds. 2006. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends in Plant Science 11:535-542.

Schranz, M. E., S. Mohammadin, and P. P. Edger. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Current Opinion in Plant Biology 15:147-153.

Simillion, C., K. Vandepoele, M. C. Van Montagu, M. Zabeau, and Y. Van de Peer. 2002. The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci U S A 99:13627-13632.

Slater, G. S. and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Team, R. D. C. 2013. R: a language and environment for statistical computing.

Tremetsberger, K., C. Konig, R. Samuel, W. Pinsker, and T. F. Stuessy. 2002. Infraspecific genetic variation in *Biscutella laevigata* (Brassicaceae): new focus on Irene Manton's hypothesis. Plant Systematics and Evolution 233:163-181.

Vanneste, K., G. Baele, S. Maere, and Y. Van de Peer. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Research.

Vanneste, K., Y. Van de Peer, and S. Maere. 2013. Inference of Genome Duplications from Age Distributions Revisited. Molecular Biology and Evolution 30:177-190.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. Proceedings of the National Academy of Sciences of the United States of America 106:13875-13879.

Yang, Z. and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496-503.

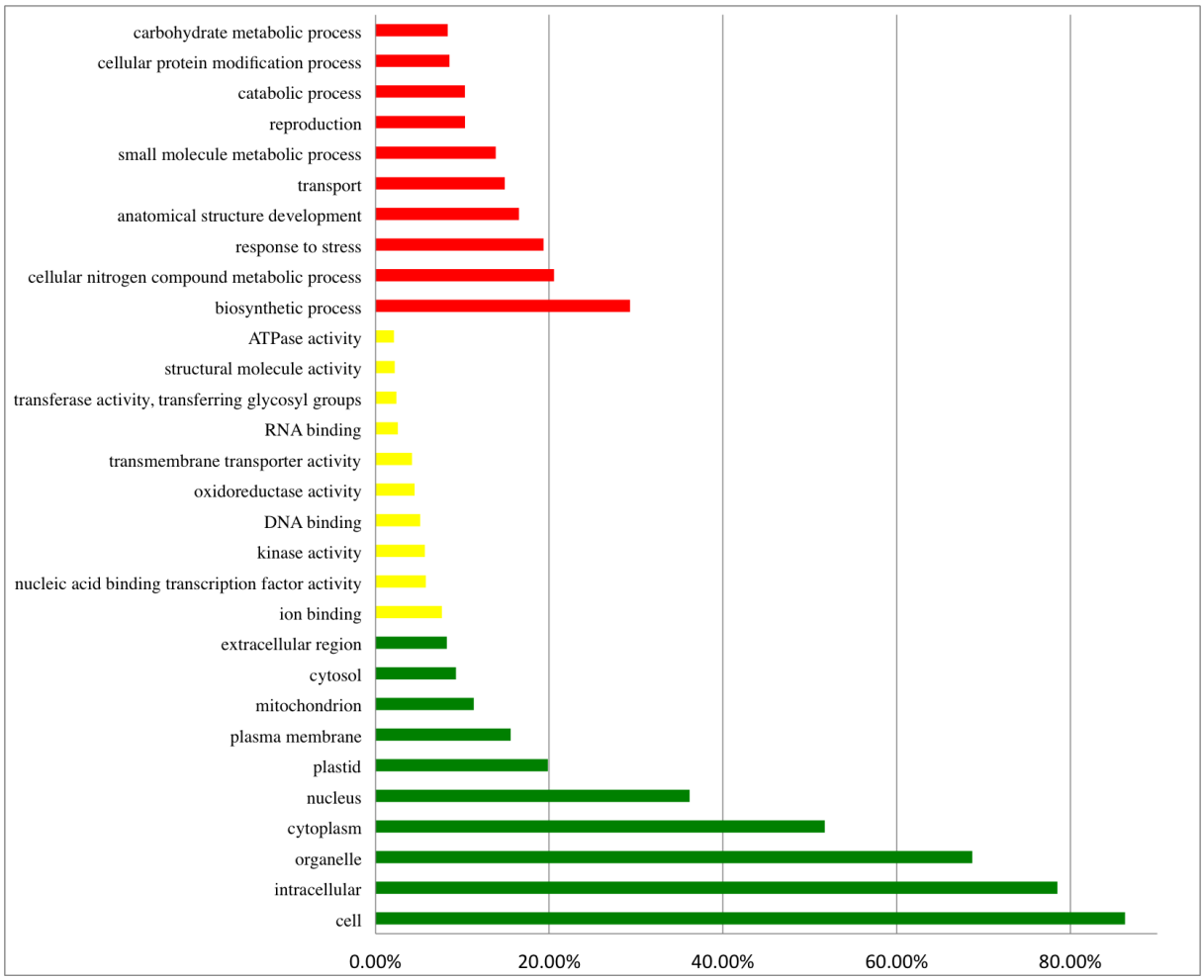
Yang, Z. H. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556.

Zeileis, A. 2014. zoo package: S3 Infrastructure for Regular and Irregular Time Series.

Zhang, D. X. and G. M. Hewitt. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol Ecol 12:563-584.

Zhang, J. Z. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292-298.

Supplementary Materials



S 1: Functional classification of the de novo transcriptome. The ten most representative GO slim categories in the ontologies: biological process (red), molecular function (yellow) and cellular component (green).

S 2: Number of GO categories over-represented in the Biscutella transcriptome vs. Arabidopsis reference. The first column corresponds to the ontology. The second column corresponds to GO terms with significant differences in each ontology. The third column represents terms over-represented in Arabidopsis and the last columns terms over-represented in Biscutella. For the GO analysis the procedure described in material and methods was used.

Category	Number of significant terms	Over-represented in *A. t.*	Over-represented in *B. l.*
Biological process	228	4	224
Cellular component	55	3	52
Molecular function	82	2	80

S 3: Custom script to allocate the annotation IDs to a given building block.

```
A<-gsub(pattern ='^(At1g+0*(1(5[6-9][0-9]|[6-9][0-9]{2}|[0-8][0-9]{3}|9([0-2][0-9]
{2}|3([0-2][0-9]|30)))|[2-9][0-9]{3}))$',replacement = "blockA",x=inp_read$Annot)
B<-gsub(pattern ='^(At1g+0*(198[5-9][0-9]|199[0-9]{2}|2[0-9]{4}|3[0-5][0-9]{3}|
36[01][0-9]{2}|362[0-3][0-9]|36240))$',replacement = "blockB",x=A)
C<-gsub(pattern ='^(At1g+0*(4359[0-9]|43[6-9][0-9]{2}|4[4-9][0-9]{3}|5[0-5][0-9]
{3}|560[0-9]{2}|561[0-3][0-9]|5614[0-5]))$',replacement = "blockC",x=B)
D<-gsub(pattern ='^(At1g+0*(565[2-9][0-9]|56[6-9][0-9]{2}|5[7-9][0-9]{3}|6[0-2]
[0-9]{3}|63[0-6][0-9]{2}|637[0-6][0-9]|63770))$',replacement = "blockD",x=C)
E<-gsub(pattern ='^(At1g+0*(650[4-9][0-9]|65[1-9][0-9]{2}|6[6-9][0-9]{3}|7[0-9]
{4}|80[0-3][0-9]{2}|804[01][0-9]|80420))$',replacement = "blockE",x=D)
F<-gsub(pattern ='^(At3g+0*(10[4-9][0-9]|1[1-9][0-9]{2}|[2-9][0-9]{3}|1[0-9]{4}|
2[0-4][0-9]{3}|25[0-4][0-9]{2}|255[01][0-9]|25520))$',replacement = "blockF",x=E)
G<-gsub(pattern ='^(At2g+0*(51[7-9][0-9]|5[2-9][0-9]{2}|6[0-9]{3}|7[0-6][0-9]{2}|
77[0-2][0-9]|773[0-3]))$',replacement = "blockG",x=Fdd)
H<-gsub(pattern ='^(At2g+0*(156[7-9][0-9]|15[7-9][0-9]{2}|1[6-9][0-9]{3}|20[0-9]
{3}|210[0-9]{2}|211[0-3][0-9]|21140))$',replacement = "blockH",x=G)
I<-gsub(pattern ='^(At2g+0*2(1(1[6-9][0-9]|[2-9][0-9]{2})|[2-7][0-9]{3}|8([0-8]
[0-9]{2}|9(0[0-9]|10))))$',replacement = "blockI",x=H)
J<-gsub(pattern ='^(At2g+0*(310[4-9][0-9]|31[1-9][0-9]{2}|3[2-9][0-9]{3}|4[0-6]
[0-9]{3}|47[0-6][0-9]{2}|477[0-2][0-9]|47730))$',replacement = "blockJ",x=I)
K<-gsub(pattern ='^(At2g+0*(12[5-9][0-9]|1[3-9][0-9]{2}|2[0-9]{3}|3[0-6][0-9]{2}|
37[0-4][0-9]|3750))$',replacement = "blockK",x=J)
L<-gsub(pattern ='^(At3g+0*2(5(8(5[5-9]|[6-9][0-9])|9[0-9]{2})|[6-8][0-9]{3}|
9([0-6][0-9]{2}|7([0-6][0-9]|7[0-2]))))$',replacement = "blockL",x=K)
M<-gsub(pattern ='^(At3g+0*4(3(7[4-9][0-9]|[89][0-9]{2})|[4-8][0-9]{3}|9([0-8]
[0-9]{2}|9([0-6][0-9]|70))))$',replacement = "blockM",x=L)
N<-gsub(pattern ='^(At3g+0*(509[5-9][0-9]|5[1-9][0-9]{3}|6[01][0-9]{3}|62[0-6]
[0-9]{2}|627[0-8][0-9]|62790))$',replacement = "blockN",x=M)
O<-gsub(pattern ='^(At4g+0*([3-9][0-9]|[1-9][0-9]{2}|[1-3][0-9]{3}|4[0-8][0-9]{2}|
49[0-4][0-9]|495[0-5]))$',replacement = "blockO",x=N)
P<-gsub(pattern ='^(At4g+0*(869[0-9]|8[7-9][0-9]{2}|9[0-9]{3}|1[01][0-9]{3}|
120[0-6][0-9]|12070))$',replacement = "blockP",x=O)
Q<-gsub(pattern ='^(At5g+0*2(2[89][0-9]{2}|[3-7][0-9]{3}|8([0-7][0-9]{2}|8([0-8]
[0-9]|9[0-7]))))$',replacement = "blockQ",x=P)
R<-gsub(pattern ='^(At5g+0*(12[4-9][0-9]|1[3-9][0-9]{2}|[2-9][0-9]{3}|1[0-9]{4}|
2[01][0-9]{3}|220[0-2][0-9]|22030))$',replacement = "blockR",x=Q)
S<-gsub(pattern ='^(At5g+0*(3262[1-9]|326[3-9][0-9]|32[7-9][0-9]{2}|3[3-9][0-9]
{3}|40[0-9]{3}|41[0-8][0-9]{2}|41900))$',replacement = "blockS",x=R)
T<-gsub(pattern ='^(At4g+0*1(2(7[5-9][0-9]|[89][0-9]{2})|[3-5][0-9]{3}|6(0[0-9]
{2}|1([0-3][0-9]|4[0-3]))))$',replacement = "blockT",x=S)
U<-gsub(pattern ='^(At4g+0*(1(6(2[5-9][0-9]|[3-9][0-9]{2})|[7-9][0-9]{3})|[2-9]
[0-9]{4}|[12][0-9]{5}|3([0-7][0-9]{4}|8([0-6][0-9]{3}|7([0-6][0-9]{2}|7([0-6][0-9]|
70))))))$',replacement = "blockU",x=T)
V<-gsub(pattern ='^(At5g+0*4(29[7-9][0-9]|[3-7][0-9]{3}|8([0-4][0-9]{2}|5([01]
[0-9]|20))))$',replacement = "blockV",x=U)
W<-gsub(pattern ='^(At5g+0*(49(4[3-9][0-9]|[5-9][0-9]{2})|5[0-9]{4}|60([0-2][0-9]
{2}|3([0-8][0-9]|90))))$',replacement = "blockW",x=V)
X<-gsub(pattern ='^(At5g+0*6(0(5[5-9][0-9]|[6-9][0-9]{2})|[1-6][0-9]{3}|7([0-2]
[0-9]{2}|3([0-7][0-9]|8[0-5]))))$',replacement = "blockX",x=W)
```

Chapter 2: Reproductive isolation in a hybrid zone of *Biscutella laevigata*

Abstract

Studies about the reproductive isolation in natural populations can greatly contribute to refine our knowledge about speciation. Whereas multifarious pre- and postzygotic barriers underlying reproductive isolation are well known, their interplay and relative importance remains to be solved. Here, I first characterized the genetic and spatial structure of the hybrid zone. Further, I estimated multiple components underlying reproductive isolation in a natural hybrid zone of the polyploid *Biscutella laevigata* in the Swiss Alps. This study tackles two prezygotic barriers (flowering time and asynchrony and prezygotic immigrant inviability) as well as three postzygotic barriers (hybrid unfitness, hybrid inviability and postzygotic immigrant inviability). These barriers were assessed by an extensive field survey with crosses used in a reciprocal transplant experiment to test for the intrinsic or extrinsic nature of these barriers. I found that low spatial co-flowering represents a main barrier contributing to the maintenance of the genetic boundary between the genetic groups. Indirect evidence of an extrinsic postzygotic barrier could be assessed in the field as the genetic groups inhabited differentiated ecological niches and the habitat of each group was described. Even so, the experimental approach failed to confirm this evidence, possible reasons for this discrepancy are discussed.

Introduction

A major goal in evolutionary biology is to understand the origin of species (Darwin 1859; Coyne and Orr 2004). For genetically divergent but interbreeding populations to become distinct species, reproductive isolation must counteract homogenization caused by gene flow, that turns the odds against divergence (Felsenstein 1981; Smajda 2011). Often strong reproductive isolation involves several types of barriers acting together, eventually leading to a complete cessation of gene flow and to speciation (Lowry et al. 2008; Barton and de Cara 2009; Widmer et al. 2009). So far, our understanding of the relative importance of different types of barriers and their interaction is still limited. In this study, I assessed the role of pre- and postzygotic mechanisms theoretically known to drive the evolution of reproductive isolation in a population of *Biscutella laevigata* consisting of two genetic groups with a relatively well defined hybrid zone.

To trigger and maintain genetic divergence several barriers can act together and are classified depending on the life stage upon which they act, (i) prezygotic barriers occur before fertilization (e.g. geographic separation, flowering time, gametic incompatibilities, prezygotic immigrant inviability) and (ii) postzygotic barriers operate after zygote formation and include all mechanisms that lead to fitness aberration in hybrids: necrosis, weakness, sterility and lethality, but also immigrant inviability. Postzygotic barriers can be due to intrinsic selection, i.e. selection acts against hybrids regardless of the environment, or extrinsic, i.e. selection depending on environmental factors (Burke and Arnold 2001). For the latter, local adaptation can be a major contributor to the evolution of reproductive isolation (Rundle and Nosil 2005; Nosil 2012) and trigger barriers to gene flow through various mechanisms (Schluter 2001; Nosil 2012). First, when locally adapted gene pools exchange migrants they should show

higher mortality in the habitat they are not adapted to. Second, a reduction of hybrid fitness should be observed and this difference can be intrinsic when the fitness decrease is independent of the environment or extrinsic when habitat dependent (Nosil 2012). Immigrant inviability can represent a postzygotic extrinsic barrier, when it results in the increased mortality of maladapted immigrants (Nosil et al. 2005). Also, when immigrants do not survive, they lower the chances of ecotype interbreeding, as they are present at a very low density. This represents a prezygotic barrier *sensu* Nosil et al. (Nosil et al. 2005).

Hybrid zones, where secondary contact brings previously separated taxa together, represent an interesting situation to study the build-up of reproductive isolation in the face of gene flow (Barton and Hewitt 1985; Barton and de Cara 2009; Payseur 2010). Most hybrid zones are maintained due to a balance between dispersal and natural selection (Barton and Hewitt 1985). Stable hybrid zones can be maintained by postzygotic intrinsic selection, i.e. fitness interactions among genes and/or extrinsic selection, i.e. local adaptation. But hybrid zones can also be transient and face different types of outcomes. First, the hybrids can become reproductively isolated from the parents leading to homoploid hybrid speciation, with the hybrids forming a new species (Rieseberg and Willis 2007). Whereas polyploid speciation, i.e. implying a change in chromosome number is well known, homoploid speciation, i.e. speciation in taxa with the same polyploid level are rare (Rieseberg and Willis 2007). A second outcome would be the evolution of increasing reproductive isolation, due to reinforcement toward speciation.

Prezygotic isolation in plants can occur pre-mating, e.g. flowering time or through gametic incompatibilities (Rieseberg and Willis 2007). For example, prezygotic isolation

factors were assessed to be the primary isolating barrier between *Mimulus lewisii* and *M. cardinalis* (Ramsey et al. 2003). An additional study showed that prezygotic isolation was the major contributor to reproductive isolation between *M. guttatus* and *M. nasutus*, likely as a direct or indirect result of drought avoidance via phenological acceleration (Martin and Willis 2007). Generally, prezygotic barriers affects more reproductive isolation than postzygotic (reviewed in Lowry et al. 2008). Even so, studies showed that postzygotic extrinsic barriers can represent the main trigger in divergence, e.g. the temperature gradient in the hybrid zone on Mt Etna (Ross et al. 2012). Even so, plenty studies demonstrated interplay between pre- and postzygotic barriers (e.g. Scopece et al. 2013).

This study tests for the presence of multiple forms of reproductive isolation in a natural, tetraploid population of the plant species *Biscutella laevigata* in the Alps. In a previous study, genetic divergence in this population was assessed with anonymous genome-wide markers (Parisod and Christin 2008). At this scale, gene flow was expected to homogenize the genetic variability. In contrast, the authors showed a high mosaic-like fine-scale genetic structure. In the mosaic hybrid zone model a strong correlation between different genotypes and a patchy habitat are expected (Buerkle et al. 2003). The authors assessed "the factors responsible for the repartition of the genetic variance and revealed a composite effect of isolation by distance, phenological divergence and local adaptation to habitats characterized by different solar radiation regimes" (Parisod and Christin 2008). In addition, a genome scan analysis on the same dataset revealed that 7% of loci were putatively under divergent selection and significantly correlated with environmental variables, mainly solar radiation (Parisod and Joost 2010). Accordingly, strong evolutionary forces seem to have maintained

divergence and local environment could sustain this divergence.

Therefore, the tetraploid *Biscutella laevigata* in this alpine population represents an ideal non-model species to investigate barriers underlying diversification between homoploid lineages and the implication of natural selection. The goal of the current study was to spatially characterize the population in regard to habitat distribution, occurrence of genetic groups, and localization of contact zones. Furthermore, I tested for the existence of pre- and postzygotic barriers between the two groups. In order to find potential prezygotic barriers, differences in flowering phenology were studied in the field. Furthermore, hybrids between both genetic groups were produced to test for early-acting barriers, i.e. hybrid dysfunction affecting seed weight. The distinction between postzygotic intrinsic and extrinsic reproductive barriers as well as testing for immigrant inviability is only possible if survival and fitness data of multiple environments can be compared (Kawecki and Ebert 2004; Nosil et al. 2005). Therefore, I performed a reciprocal transplant experiment in nature. The reciprocal transplant of hybrids informs about the forces underlying postzygotic barriers whether they are intrinsic or extrinsic as well as F1 hybrid inviability. The reciprocal transplant of pure group A and B individuals allows to test for postzygotic extrinsic isolation, i.e. local adaptation (Nosil et al. 2005).

Material and Methods

Sampling of plants and environmental data

Biscutella laevigata is a long-lived perennial autotetraploid species with a sporophytic self-incompatibility system (Olowokudejo and Heywood 1984). The species is thought to have survived the Pleistocene as a diploid (2n=2x=18) and to have recolonized the Alps as an autotetraploid (2n=4x=36) after the ice ages (Manton 1937). Our study site is located in the western Alps of Switzerland, in the Rochers-de-Naye, 1864-2043 meters above sea level. Here the species occurs in patches along a mountain ridge of 1.2 km (Parisod and Christin 2008).

During summer 2011, random cluster sampling was conducted in the area of general occurrence of *B. laevigata*. 348 individuals were sampled as follows: randomly selected areas of 4x4m were subdivided in four 2x2m plots, and separated at distances between 0-25m for transects. Plots were sampled when at least five plants were observed and most areas were thus sampled on less than four plots. When less than five plants were reported in any of the four plots, a new area at a random distance (over 25m) was selected for sampling, i.e. a new transect. Transects represent groups of areas separated by less than 25m from each other, thus representing the real distribution of this species in the field. Leaves of each of the five individuals were collected and immediately dried in silica gel. In addition, pictures of the surrounding habitat (20x20cm) were taken *in situ* to derive habitat coverage (stone, soil/humus, vegetation). In each plot, four plants were labeled for *in situ* observation over three years (2011-2013) and one plant was transplanted in the common garden (Jardin Botanique de Neuchâtel, Switzerland). All plants where georeferenced precisely by GPS with 2cm

accuracy.

In 2012 and 2013, 212 labeled individuals were surveyed during the whole vegetation season. Flowering time and seed set were recorded four times during the vegetation period, from flowering to maturation of seeds. A fifth survey was made every year to ensure that no plants had flowers any more. This data was to compare flowering phenology across the hybrid zone. In addition, the size and number of stems were recorded once a year in 2011, 2012 and 2013. Originally, over 285 plants were marked on the field survey. Plants were lost during survey due to several issues i) labels were removed, probably human induced (e.g. in one plot of transect TC labels were arranged in a circle near the cliff) ii) grazing of stems, i.e. no reliable flower count (probably due to cows and chamois) iii) important loss during the winter iv) small landslips.

For a subset of these 193 individuals, soil humidity was reliably measured with « FieldScout TDR 300 Soil Moisture » (Spectrum Technologies, 3600 Thayer Court, Aurora, IL 60504, USA). The unit of measurement is volumetric water content (VWC) and represents an electrical signal transformed into percentage soil moisture content using an equation valid over a wide range of mineral soils. And, to further characterize the environment at a fine scale, a very high-resolution digital elevation model (DEM, with 0.5m resolution) was acquired for the study area. A set of variables (e.g. altitude, slope, aspect, curvature, downslope distance gradient, vector ruggedness measure, wetness index) at different spatial resolution was derived from this DEM.

Habitat analysis

A first step toward the understanding of diversification patterns in this population relied on the characterization of heterogeneity of the habitat. For this analysis seven

environmental variables were used at the individual level after being log transformed. Here, I used three uncorrelated DEM-derived variables (Spearman correlations above ± 0.5) at the highest resolution (0.5m): ruggedness, slope, downslope distance. In addition the three types of coverage (vegetation, soil, stone) as well as the direct measurement of soil humidity were included. Ruggedness is a measure for the "unevenness" of terrain, e.g. broken, rocky vs. flat terrain (Sappington et al. 2007). Plenty of ruggedness measures were developed that were unable to incorporate variability in aspect and gradient components, i.e. making the difference between a steep broken area and a steep flat area (Sappington et al. 2007). On the alpine ridge studied here, most individuals are located in areas that are not flat. Therefore, the vector ruggedness index (here with a radius=2pixel) was a valuable measure to distinguish steep and broken areas vs. steep flat areas. To complement this variable by quantifying the steepness of the area, the slope (radians) was used. Finally I used the downslope distance gradient as proxy for the local drainage patterns, i.e. the removal of surface water from an area. The downslope distance gradient (radians) calculated here was shown to represent drainage issues better than the commonly used slope measure (Hjerdt et al. 2004). Soil humidity was used as a second proxy for water availability at the individual level measured directly in the field. Finally, the three types of coverage (soil, stone and vegetation) were used as descriptors of the abiotic (stone coverage) but also biotic (presence/absence of other plants, i.e. soil and vegetation cover) environment of plants.

To obtain a general pattern of habitat variance in the dataset, the habitat along the ridge was summarized at the individual level with a principal component analysis (PCA) on a matrix of 193 individuals and the seven environmental variables (DEM-derived: ruggedness, slope, downslope distance; cover: vegetation, soil, stone; direct

measurement: soil humidity). Unless stated otherwise all statistical analyses were achieved with the R-software (Development Core Team R 2013). Six individuals where removed because they represented outliers. These errors were probably due to ambiguous calculations of topographic variables when the points were in highly unstable soil situation. The magnitude and statistical significance of a shift in habitats between individuals in transects was assessed with a between-class PCA (BPCA). The statistical significance of the BCA was tested using a Monte-Carlo test based on 999-permutation (Romesburg 1985).

Genotyping and genetic structure analysis

The population (n= 348) was genotyped at 266 polymorphic loci with six reliable AFLP (amplified fragment length polymorphism) primer combinations (MCAG/EATC, EAGG/MCGG, MCAG/EAAT, EACT/MCAC, MCGA/EATA and MCGG/EATA; following Parisod and Christin 2008). A set of 38 primer combinations was tested previously on a subset of eleven individuals; the six best selective primer combinations regarding polymorphism and reproducibility were used. As advised 15% of the individuals were replicated and an error rate of 2.93% was calculated (Bonin et al. 2007).

The dominant nature of AFLP markers does not allow gaining information about allele dosage and heterozygozity, i.e. assessing Hardy-Weinberg equilibrium. Therefore, the use of traditional genetic population methods developed for diploids are inadequate here and multivariate analyses, k-means and c-means were used (Dufresne et al. 2014). The population genetic structure was assessed using clustering methods based on the multidimensional partitioning of the AFLP profiles of 348 individuals (266 AFLP markers).

In a first analysis the number of genetic clusters was determined to assess whether I faced reproductive isolation between two diverging groups or, if the presence of a third group could indicate homoploid hybrid speciation. The population was divided in two to ten clusters by K-means and the optimal number of genetic groups was estimated using Calinski-Harabasz criterion (Calinski and Harabasz 1974). K-means hard clustering (i.e. each individual belong to one cluster) was chosen because it offers a robust statistical framework to test for the optimal number of clusters (Calinski and Harabasz 1974). Once the optimal number of clusters was selected, the absolute membership (1 or 0) was obtained for each individual. Other algorithms, e.g. c-means (i.e. soft clustering), do allow to infer not only to which group an individual strictly belongs, but also the degree to which individuals showed deviation from a strict assignment. These deviations from a strict assignment can be interpreted as proportion of genetic admixture (Gompert et al. 2010). The individuals were assigned to clusters by fuzzy c-means clustering (n=348, using the R package « e1071 » with a fuzzification exponent at 1.02, 100'000 iterations, (Meyer et al. 2014). The c-means algorithm was bootstrapped over loci (266 times, custom scripts) the bootstrapped median was chosen as hybrid-index. Further, hybrid-indexes were ordered, the visual identification of the inflexion points of the resulting curve was used to choose a threshold between pure individuals and hybrids. Individuals were considered to belong to « group A» with a hybrid-index from 0.86-1, to « group B » with a hybrid index from 0-0.14 and as hybrids when their admixture level was higher than 15% (Hybrid-index: 0.15-0.85). To test for the reliability of assignments of K-means and c-means methods, their results were compared. The c-means memberships were rounded and compared to the absolute K-means assignment and the number of individuals with same assignment by both

methods was reported.

To assess the spatial structure of the hybrid zone, the hybrid index was plotted against the linearized spatial coordinates from West to East. To test for difference between numbers of individuals of group A vs. B inhabiting each side of the ridge a non-parametric proportion test was performed (two-sided with continuity correction).

Production of crosses

For the reciprocal transplant experiment artificial crosses were performed. During summer 2011, 110 plants were collected in the field (at least one in each plot) and grown over winter in the common garden (Jardin Botanique, Neuchâtel, Switzerland). Plants on the field were collected within a plastic bag; the block of soil around plants was conserved and enclosed in wet paper sheets to preserve the plant. Each block of soil contained several plants; these plants were all transplanted into separate pots in the botanical garden and genotyped separately. All individuals in a soil block were clones (tested with AFLPs); therefore only 64 genetically different plants could be used for the crosses. Further, only 36 plants produced flowers during the crossing period.

The cross design was performed to obtain three balanced categories of crosses: intra-genetic group crosses, thus representing two pure individuals from the genetic group A and B (genetic distance < 0.26 between parental hybrid-indexes) and hybrids between A and B (genetic distance > 0.64 between parental hybrid-indexes).

Prezygotic reproductive isolation: flowering synchrony and immigrant inviability

Prezygotic reproductive isolation was assessed with three analyses of flowering time and synchrony. First, flowering time and synchrony was assessed between years and survey periods, second flowering synchrony was assessed over the whole ridge in years separately taking into account the genetic distance between individuals. Third, the flowering synchrony was assessed in space, as individuals nearby i.e. in transects have more chances to interbreed.

First, evidence for a prezygotic temporal barrier, i.e. flowering synchrony, was assessed comparing differences in peaks of maximum open flowers between both genetic groups separately in time. Due to important meteorological variations between the two survey years (in particular a longer snow cover in 2013) the first comparison focused on differences in flowering time between both genetic groups to control for yearly variations. For each of the five survey periods (each year separately), the number of open flowers from a genetic group A vs. B was fitted (five in each year, day 0 being the 24th of June until day 59, i.e. 22nd of August). This analysis should help to identify major shifts in flowering between both genetic groups and disentangle shifts in flowering time due to yearly meteorological, thus represent a prezygotic barrier.

Second, the flowering synchrony was assessed over the whole ridge to test if flowering synchrony, i.e. producing flowers at the same time allowed interbreeding, was fixed in both genetic groups. To assess the overall flowering synchrony that is fixed genetically I assessed whether genetically close plants do more often flower at the same time than more distant plants based on differences in their hybrid index. For this issue a co-flowering matrix was computed between individuals for each year separately. A value of one was allocated if two individuals had at least one open flower at the same

time in any given survey period, i.e. co-flowering. A matrix of Euclidean genetic distance (distance between hybrid-indexes) was built to assess the genetic difference between individuals in a pair. To assess a relation between the co-flowering matrix and genetic distances I performed a Mantel test with 10'000 resampling (Bjornstad 2013).This was done separately for 2012 and 2013. This test was to assess if genetically distant individuals, i.e. belonging to the different genetic groups, flower less often together, thereby triggering reproductive isolation.

The third analysis of flowering synchrony incorporated space by fitting the numbers of flowers for each transect separately. The chances for both genetic groups to interbreed are higher for nearby individuals, represented here in transects. In the sampling, a transect represents a group of individuals (in plots and areas) separated by less than 25 meters from another. For this issue the numbers of flowers in one transect in each survey time were fitted in barplots, i.e. co-flowering in space.

In a hybrid zone with locally adapted ecotypes, migrants can physically settle in the "other" habitat. Here, assessing the number of immigrants in habitats mostly inhabited by native individuals of the other genetic group could give first indications about the presence of such a barrier. First the number of migrants of each genetic group was assessed. Further, a non-parametric proportion test allowed testing for asymmetric immigrant inviability, i.e. if immigrants of one group survived better in on one side of the hybrid zone.

Postzygotic reproductive isolation – indirect evidence in nature

I first assessed indirect evidence for postzygotic barriers in the field. A proxy for hybrid fitness was stem production; this was measured over the three years in nature. This variable was chosen, because I could ascertain to survey it completely. For other variables, e.g. flowers or seeds, the survey was representative, not complete. As a matter of fact, even so stems were eaten or broken (herbivory or human induced) they could be counted, in contrary to flowers or seeds. In addition stems represent an interesting variable, as it represents also the reproductive organs. The fitness of hybrids was estimated comparing proportions of hybrids and pure individuals (split in group A and B) that had at least one stem every year (during three consecutive years) with a non-parametric proportion test.

Furthermore, to gain some indirect evidence for local adaptation, i.e. pure individuals of group A and B inhabiting different habitats, I characterized habitat differences between these genetic groups. The first analysis focused on whether A and B individuals occupy different niches, but also if these niches differ from that of hybrids. In case of local adaptation, the niches of "adapted" pure individuals should be more specific than those of hybrids that should inhabit less specific habitats (Nosil 2012). Therefore, differences in the seven environmental variables (see habitat analyses for details about the variables) between genetic group A, genetic group B and hybrids were performed with multiple Kruskal-Wallis tests. Here, multiple refers to the genetic groups A, B and hybrids tested (Giraudoux 2013).

Once niche differences were assessed, I focused on specific niche differences characterizing the habitat of A and B individuals (i.e. without hybrids). To do so I performed GLM (generalized linearized model) with the rounded hybrid index being the

binary response, i.e. representing the genetic groups, with a binomial error distribution. In the first GLM, seven environmental variables (ruggedness, slope, downslope distance gradient, soil humidity, soil coverage, vegetation coverage, stone coverage) were the fixed effects. A stepwise-selection of these variables allowed extracting the environmental variables explaining best the groups based on the Aikaike Information criterion (AIC). A second GLM was fitted with these selected variables (ruggedness, slope, downslope distance gradient, soil humidity) and the overall explained variance was assessed with pseudo-R^2 computed according to Nagelkerke (Nagelkerke 1991). To control for the absence of strong collinearity the variance inflation factor (VIF) was computed. To control for the absence of strong spatial autocorellation, i.e. degree of dependency among environmental values for individuals in the geographic space, a spline correlogram was fitted for the GLM. This analysis was important because of the sampling scheme in plots that could cause a strong autocorrelation of environmental variables at small distances for nearby individuals biasing the model. The correlogram was fitted for individuals distant of a maximum of 500m according to the spatial coordinates X and Y and the Pearson residuals of the GLM. The significance of each variable was assessed by Wald Chi-square tests. To quantify the effect of each variable, the probability to belong to each group according to the values of the environmental variables assessed in the GLM was fitted. To completely remove the effect of spatial autocorrelation a GLMM (generalized linearized mixed model) was fitted with "Area" as random effect (Bolker et al. 2009; Bates et al. 2013). Again, significance for each variable was assessed and the pseudo-R^2, explained variance was calculated (Nakagawa and Schielzeth 2010).

Postzygotic reproductive isolation – experimental evidence

Direct evidence of postzygotic barriers could be assessed with the crosses and their reciprocal transplant experiment. Putative hybrid dysfunction was obtained by assessing the seed weight of hybrid vs. pure crosses. Seed weight (mass [mg]) was measured with an analytical balance (Mettler Toledo XA 205 DualRange). Differences in seed weight between hybrid and pure crosses were compared with a Kruskal-Wallis test. These seeds were grown for one month in a culture room (16h: 8h, light: dark cycle) in small pots filled with a mix of standard potting mix with 10% sand. Germination and growth i.e. the number of leaves was monitored every second day. Plants were grown one month in the culture room, i.e. the minimal amount of time to obtain at least two leaves (plants had between two and five leaves) and to avoid acclimation to laboratory condition. One leaf was sampled for each individual and immediately dried in silica-gel before transplant in the field.

Seedlings were transplanted in the field in four transplant blocks representing the spatial extent of the population along the ridge. Two blocks were chosen in transects inhabited by a majority of genetic group A (TC and TE, transplant block BC and BE) and two inhabited by a majority of group B individuals (TG and TL, transplant blocks BG and BL). Individuals from each cross were evenly allocated between all four blocks and transplanted randomly inside each block (August 2012). The transplant blocks (~1m20x1m) were protected as far as possible from animals (herbivores) and human influences with anti bird nets.

During the two months after transplant and during the main vegetation period the following year the survival and performance of plants were assessed. An individual was considered as surviving when at least one green leaf could be observed. For each

plant, vegetative traits, such as the number of leaves, length of the longest leaf and number of inflorescences were measured when present. The number of leaves was counted, i.e. every green leaf. In addition the longest leaf was measured on the field with a ruler (cm), i.e. leaf size. When all leaves were dried out, i.e. period of snow or heat in the summer, the survival was recorded as zero, i.e. died, as it was not possible to assess whether a plant would recover or not. When seedlings originating from the surrounding population settled nearby transplanted individuals and the original transplanted individual couldn't clearly be identified, it was discarded from the dataset. Regular gardening around the transplanted individuals in the field was done to remove individuals from the same species in order to allow the recognition and survey of the transplanted individual. Other species that settled and grow nearby were considered as being part of the tested environment and were not removed. To assess differential performance in the transplanted habitats, measures taken at three different measurement times, i.e. at different life stages were selected. First, the number of leaves one day before the transplant experiment, i.e. performance in mild conditions. Second, the size of leaves (i.e. size of the largest leave in cm) in the first transplant year was measured, one and a half month after transplant into nature. This date was chosen as representing the size of individuals after short-term acclimation to field conditions. Third, the survival, size and the largest leaf one year after the transplant experiment was used as representing the "long-term" performance.

Direct proof of local adaptation, i.e. the presence of postzygotic extrinsic barriers (immigrant inviability, hybrids inviability and unfitness) was investigated by the analysis of crosses and the reciprocal transplant according to two main hypotheses. The first hypothesis tested was: each genetic group (A and B) inhabits a specific niche to

which it is adapted, i.e. western versus eastern part of the hybrid zone. This was the hypothesis used to prepare the reciprocal transplant experiment. Here the knowledge about habitat variation between transects gathered during the habitat analysis, performed after reciprocal transplant, were ignored. For this hypothesis the transplanted individuals where separated in three cross types: “intra A”, “intra B” and hybrids. A given cross was considered as “home” in a transplant block on the western or eastern side of the hybrid zone where its genotype was represented in majority. Therefore transplant blocks BC and BE represented the habitat of the western part of the ridge, i.e. genetic group A; whereas transplant blocks BG and BL represented the habitat of the eastern part of the ridge, i.e. genetic group B. This hypothesis was called the home/away hypothesis.

A second hypothesis was based on the habitat analysis (see Figure 1) that revealed heterogeneity between transects inhabited by a specific genetic group, i.e. heterogeneity inside the western and eastern part of the hybrid zone. Here, I wanted to test if individuals are locally adapted to a specific habitat (not the whole region inhabited majortiarly by group A or group B). Habitats were defined as being “clusters” of transects in a quadrant of BPCA space (3 quadrants, i.e. 3 habitats were represented in transplant blocks, see orange arrows in Figure 1). Here, the F1 individuals were split in four groups for the analysis: crosses with both parents native i) from region of the transect TC defined by the second quadrant of BPCA (intra RC crosses), ii) from the region of transect TE and TG defined by the third quadrant of BPCA (intra REG crosses), iii) from the region of transect TL defined by the fourth quadrant of BPCA (intra RL crosses) and finally, iv) with mixed origin that cannot be tested with the transplant blocks in this analysis (intra none crosses). This hypothesis will further be referred as

the native/ non-native hypothesis. This analysis contained the same crosses, individuals and transplant blocks as the first home/away hypothesis. The difference is that the categories, in which a cross is categorized, depend on more precise habitat knowledge as in the home/away hypothesis.

To test these hypotheses, differences in long-term survival between home vs. away, respectively native/ non-native F1s were tested one year after transplant with a non-parametric proportion test, alternative hypothesis greater, without continuity correction. The proportion of surviving individuals was tested for each hypothesis separately. I tested whether home, respectively native individuals had a better survival than away, respectively non-native individuals to test for immigrant inviability. This was to test for exogenous, i.e. environmentally induced, immigrant inviability. Second, I tested if home/native individuals experienced a better survival than hybrids in each transplant block separately. This was to test for postzygotic exogenous, i.e. environmentally induced hybrid inviability. To test whether a putative decreased hybrid survival was imputable to an extrinsic or intrinsic barrier I tested also differences in hybrid survival across transplant blocks. To control for different survival inside a cross type (genetic intra A, B or intra RC, REG, RL) I also tested differences in survival between transplant blocks.

The same test design was used to test for performance differences with Kruskal-Wallis tests. I tested whether home, respectively native individuals had a better performance than away, respectively non-native individuals. This was to test for exogenous, i.e. environmentally induced, immigrant unfitness, eventually leading to inviability in the future. Second, I tested if home/native individuals experienced a better performance than hybrids in each transplant block separately. This was to test for

postzygotic exogenous, i.e. environmentally induced hybrid unfitness.

Three performance variables were tested for individuals in the field. First, differences in number of leaves were tested, i.e. the number of leaves one year after transplant minus number of leaves at day zero of transplant. This difference should represent the growth of individuals taking into account differences in performances in the common environment. Second a difference in growth, i.e. leaf growth was tested as the largest leave one year after transplant minus largest leaves one and a half month after transplant. This variable was tested to assess a potential growth in field conditions, i.e. from short-term acclimation to the field (one and a half month after transplant) until "long-term" acclimation (one year after transplant). Growth variables were tested on individuals being present, i.e. alive, the first year after the transplant experiment. This was done because individuals were very small, even after one year of transplant, and could have no leaves recorded at a given time, but recover soon after. A third variable was chosen to represent all life stages from growth in a common environment, including mortality and growth in nature. Therefore, the number of leaves one year after the transplant experiment was tested for each cross type. Individuals that had no leaves at that time were analyzed as having zero leaves and were not removed in contrary to the previous analyses. Even so, the species has deciduous leaves, i.e. falling off at maturity, the abscission process should coincide with the end of the vegetation period. Observation of plants in the field and especially in the botanical garden, showed that plants grow in size between years, the exact period of abscission is unknown but starts after mid-October in the field for plants that did not produce flowers and even later in the botanical garden.

Results

Habitat analysis

After removal of outliers, the habitat analysis was computed on 14 transects containing between four and 41 individuals (Table 1). The transects form an East to West gradient (TA-TM) along this nearly linear population.

Transect	N	Coordinate X	Coordinate Y
TA	8	564498.8	142490.3
TB	5	564555.4	142520.5
TC	22	564681.7	142650.7
TD	13	564747.3	142740.8
TE	8	564785.5	142784.5
TF	11	564811.2	142806.2
TG	22	564936.6	142956.9
TH	11	565002.3	142995.4
TI	41	565115	143043.3
TJ	15	565203.8	143041.7
TK	4	565258.2	143049.9
TL	18	565289.3	143051.5
TM	8	565371.1	143052.8
TN	7	565420.5	143064.4
Tot	193	-	-

Table 1: Transect locations. The habitat analysis, BPCA, was computed between 14 transects, first column. Each transect contained between four and 41 individuals, see second column. The mean coordinate of transects (CH1903, Swiss coordinate system) are shown in columns three and four.

The two first axes of the between-class (i.e. transects) principal component analysis (BPCA) represented over 75% of total inertia and were strongly significant by the Monte-Carlo test (n = 999 permutations, $p < 0.001$, Figure 1). This showed a significant between-transect variance. Transects were scattered in several groups representing a mosaic habitat and no clear West to East gradient was observed (Figure 1). The first quadrant contained two main clusters (transects TN and TK, and TA, TD, TJ and TM, Figure 1) characterized mainly by low ruggedness, slope and downslope

distance gradient. The second quadrant (transects TC, TB and TH, Figure 1) is characterized by high soil coverage and low soil humidity. The third quadrant (transects TE, TF and TG, Figure 1) is mainly characterized by a steep habitat with high ruggedness and downslope gradient, whereas the fourth quadrant corresponds to a habitat with important vegetation coverage and high soil humidity (transects TI and TL, Figure 1).

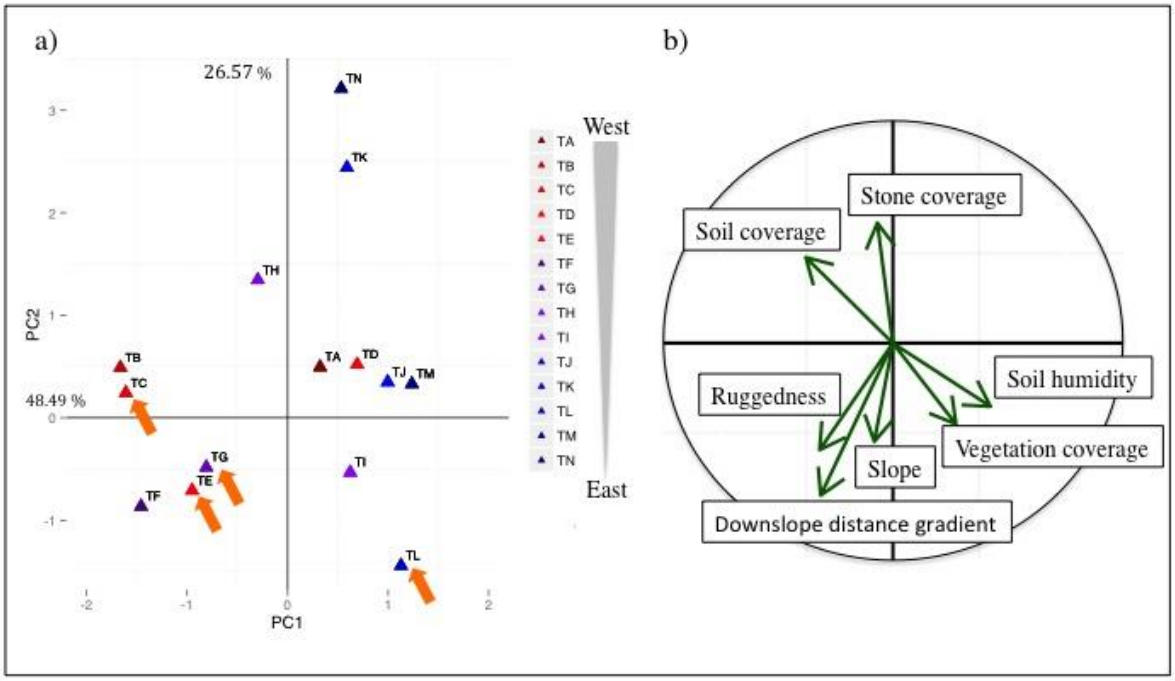


Figure 1: Between-class analysis. On the panel a) the projection of transects (i.e. classes) on the two principal axes of the BPCA (with the proportion of explained inertia for each axis, i.e 48.49% and 26.57%). The orange arrows represent the location of a transplant block for further analyses in this chapter. The color gradient of transects represents the spatial position of transects (from the most western in red to the most eastern in blue). On panel b) the correlation circle for axes one and two with the environmental variables included in the analysis: ruggedness, downslope distance gradient, slope, vegetation coverage, soil coverage, stone coverage as well as soil humidity.

The variance in the first axis was mainly explained by coverage of soil and vegetation as well as soil humidity, whereas the second axis was mainly explained by stone coverage and DEM-derived, topographic variables (ruggedness, downslope

distance gradient and slope) and in a lesser extent by vegetation coverage and soil humidity. The missing transect TY and lower sample size across transects are due to i) some direct measurements were unreliable, ii) some individuals were lost (i.e. the label was removed) and iii) six strong outliers in the DEM derived variables were removed. Soil humidity measures on highly rocky areas as well as cemented areas, where the sensor could not be correctly inserted in the soil were considered as unreliable and removed.

Population genetic structure

K-means clustering provided evidence for two genetic groups (by highest Calinski-Harabasz criterion, Figure 2), the first containing 134 (39%) and the other 214 (61%) individuals (n=348).

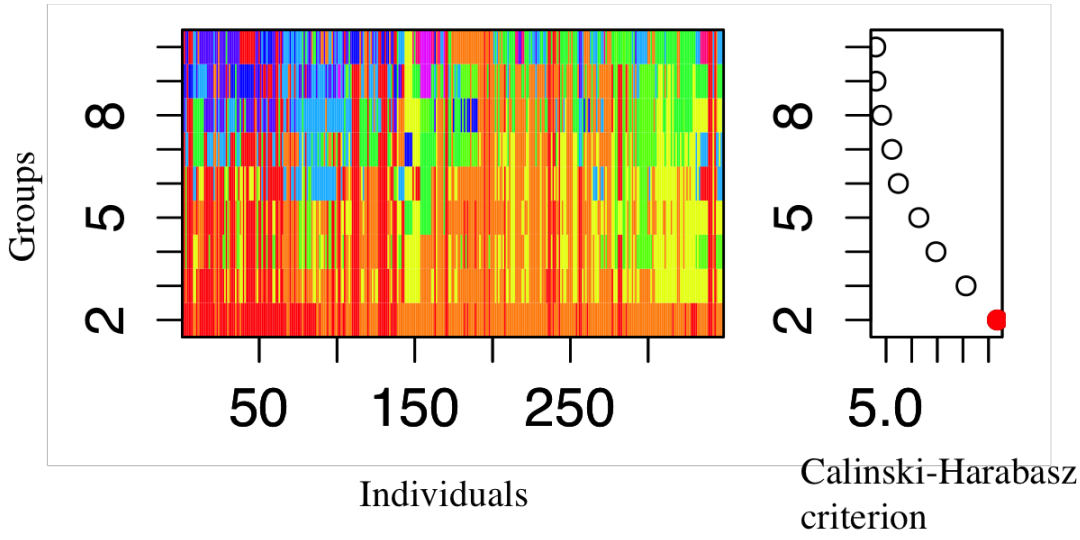


Figure 2: Optimal number of genetic groups. On the left panel the different clustering solutions on the y-axis (up to ten groups tested) for the 348 individuals (on the x-axis). Different colors in each line represent the assignement of each individual to a group according to the different clustering solutions. On the right panel the Calinski-Harabasz criterion on the x-axis for each clustering solution (y-axis). The best clustering solution by Calinski-Harabasz criterion (K=2 groups) is highlighted in red.

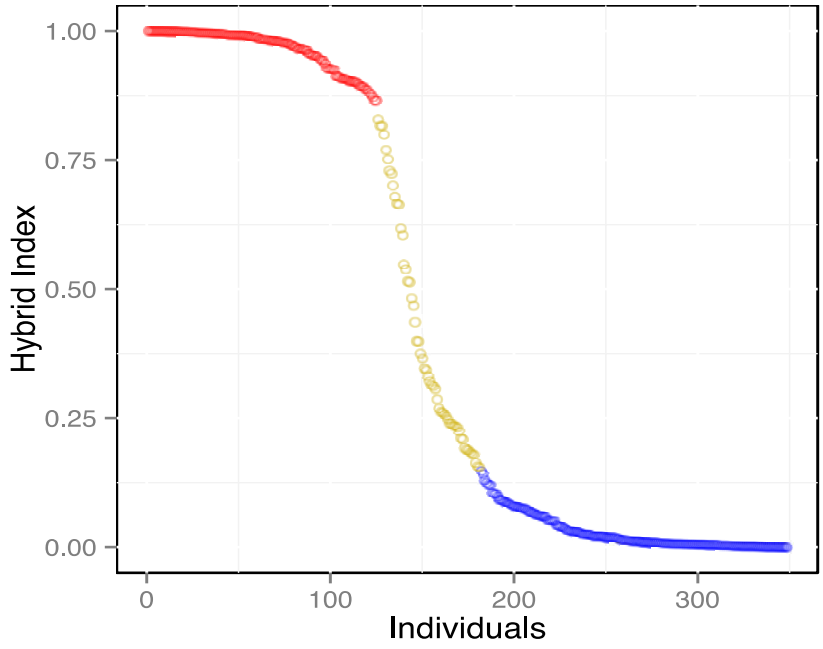


Figure 3: Genomic cline among individuals. On the x-axis the 348 individuals ordered by their hybrid-index (y-axis) as calculated by c-means. The color of the points represents the individuals in each category: genetic group A in red, B in blue and hybrids in yellow.

The ranking of hybrid-indexes produced by the c-means clustering highlighted a genomic cline typical of a hybrid zone between two gene pools, called A and B (A in red, 36% of individuals; B in blue, 48% of individuals, Table 2) with few hybrids (16%, yellow, Figure 3, Table 2). The comparison between K-means and c-means showed that 99% individuals were assigned to the equivalent cluster and confirmed the reliability of the procedure.

Less individuals were sampled in the western Part group A (n=138, Table 2) than the eastern Part (n=210, Table2). A low and similar proportion of hybrids was sampled on each side of the hybrid zone (respectively 25 and 31, Table 2). In the western part, inhabited by a majority of A individuals, 14 individuals were sampled that belong to the genetic B group, i.e. immigrants (Table 2). In the eastern part 26 immigrants were sampled.

Table 2: Number of individuals on each side of the hybrid zone. The first column represents the category of individuals, whether they belong to group A, B, hybrids and the total in the western or eastern part of the hybrid zone. In red, immigrant individuals, that inhabit the part of the ridge where their genotype is not in majority.

	Western Part	Eastern Part
A	99	26
B	14	153
Hybrids	25	31
Tot	138	210

Plotting genotypes (i.e. the individual hybrid index) along their linearized coordinates (i.e. spatial extent of this nearly linear population) revealed a clear geographic cline (Figure 4). A strong spatial component was revealed, where “A” and “B” groups inhabited significantly (p-value < 0.01, Table 2) the western or the eastern parts of the ridge. Transects, i.e. groups of sampled areas separated by less than 25m, reflected the distribution of the species. TY, TA, TB, TC, TE and TF were mostly inhabited by individuals of group A, while transects TG, TH, TI, TJ, TK, TL, TM and TN were mostly inhabited by individuals of group B (Figure 4).

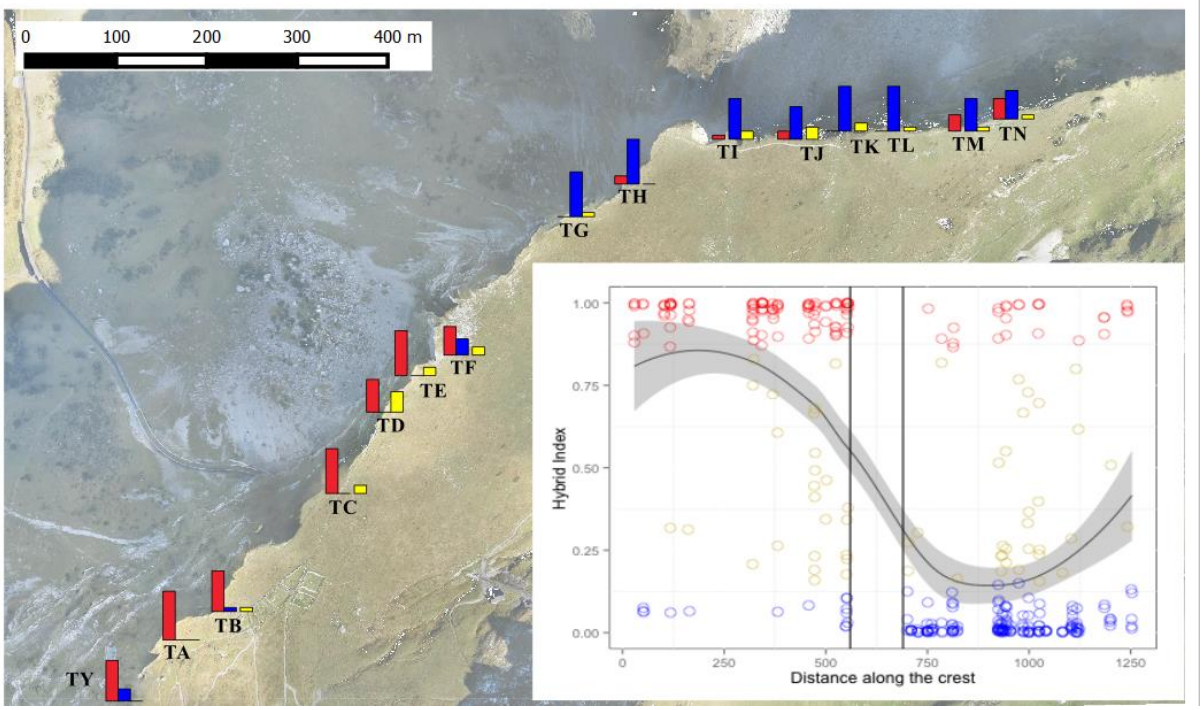


Figure 4: Spatial structure of the hybrid zone. The map in the background represents the ridge, the barplots represent the proportion of individuals in group A (red), B (green) and hybrids (yellow) in each transect according to their hybrid index (n=348). On the graph, bottom right, the hybrid index of individuals (y-axis, assessed by c-means) along the ridge (x-axis; linearized coordinates along the crest line, from zero the most western individuals to 1250 the most eastern individuals). The space between the two vertical lines represents a region of very low density (<5 individuals/ m2) coinciding with the inflexion point of the hybrid zone.

Production of crosses

From March to July 2012, 336 F1 were produced in the greenhouse by hand pollination. Flowers were bagged to avoid non-controlled pollinations. A balanced number of crosses in each category (n=15, Table 3) could be produced: five hybrid crosses (genetic difference between 0.64 and 0.99), four intra A group crosses (genetic difference between 0.0005 and 0.1013) and five intra B crosses (genetic difference between 0.0084 and 0.1420). The maximal number of crosses was produced (15 parental pairs with

hybrid indexes matching the cross design) with parental plants that flowered together in the greenhouse. All crossed parental plants produced seeds.

Table 3: Crosses. In the first column the name of the cross, followed by the number of individuals in each cross. The third column represents the genetic difference between parents, low differences for pure crosses, high differences for hybrids. The mean hybrid index between parents is showed in the fourth column, low values represent crosses from group B, high values from group A and intermediate the hybrids. The two last columns represent the cross category used in the analysis of the reciprocal transplant experiment. *Hypothesis 1 referring to the home/away hypothesis, hypothesis two to the native/non-native hypothesis.

Cross Name	Number of individuals	Genetic difference between parents	Mean hybrid index between parents	Genetic category	Category Hypothesis 1*	Category Hypothesis 2*
CaGd	9	0.9929	0.5025	hybrid	hybrid	hybrid
DaHb	20	0.9850	0.4954	hybrid	hybrid	hybrid
CbHb	12	0.9969	0.5014	hybrid	hybrid	hybrid
LbJa	43	0.9214	0.5348	hybrid	hybrid	hybrid
CbIb	15	0.9076	0.5459	hybrid	hybrid	hybrid
NaCa	6	0.6441	0.6432	hybrid	hybrid	hybrid
CaZ	63	0.1013	0.9484	intra_A	intra_A	Intra_RC
Zcap	24	0.0677	0.9315	intra_A	intra_A	Intra_RC
DaCb	3	0.0115	0.9937	intra_A	intra_A	none
CbEc	17	0.0005	0.9991	intra_A	intra_A	none
LaIf	48	0.0084	0.0057	intra_B	intra_B	intra_RL
GeGe	30	0.0073	0.0082	intra_B	intra_B	intra_REG
HaHb	8	0.0035	0.0047	intra_B	intra_B	none
MbLa	11	0.0224	0.0212	intra_B	intra_B	none
HbIc	27	0.1420	0.0739	intra_B	intra_B	none
Total	336	-	-	-	-	-

Prezygotic reproductive isolation: flowering synchrony and immigrant inviability

During the first sampling year (Figure 5, dashed lines) the maximum of open flowers peaked in June for both genetic groups (day 0: 24th of June). Afterwards the number of flowers decreased until the last sampling survey with zero flowers (day 59, 22nd of August). The beginning of flowering in 2012 was not surveyed, but very few plants were observed flowering end of May. In 2013, the flowering peak was reached 20 days later than in 2012 (Figure 5). In ten days, from day 20 to day 30 the number of flowers decreased quickly (Figure 5, solid lines). In 2012 the same decrease was reached after more than 30 days (Figure 5, dashed lines). Flowering peaks are highly depending on yearly variations as showed by the different patterns in 2012 and 2013. At this scale the two genetic groups do not seem to have a strong shift in flowering time.

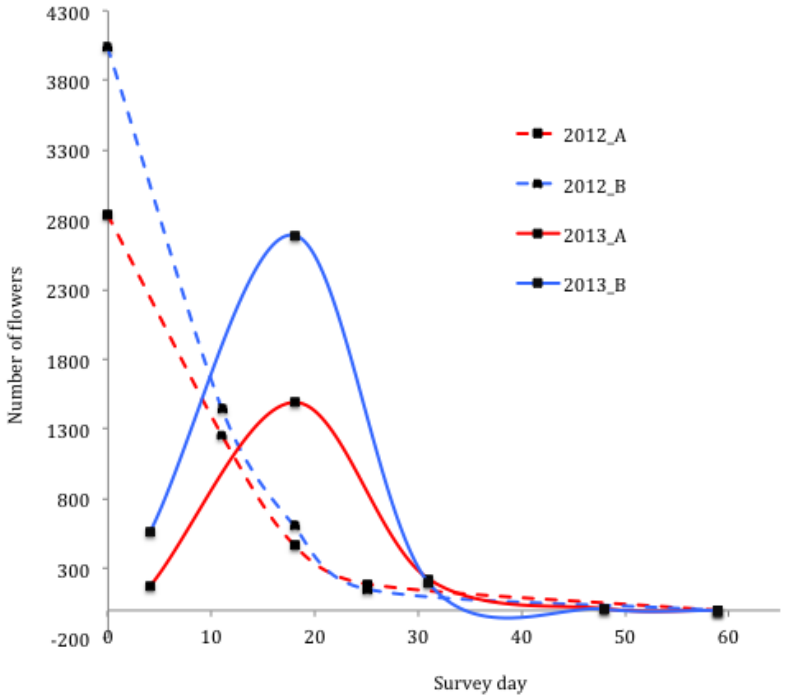


Figure 5: Flowering peak in a yearly comparison. The two dashed lines represent the number of flowers opened at a given sampling day in the year 2012, the solid lines in 2013. On the y-axis the number of opened flowers are represented, on the x-axis the sampling day. Day zero corresponds to the 24th of June until day 59, the 22nd of August, no flowers were found in any sampling year.

The Mantel test between the co-flowering matrix and genetic distances allowed to infer a negative, low correlation in 2012 (correlation: -0.05, p-value= 0.02). This correlation was not confirmed in 2013 (i.e. not significant). Therefore, the smaller the genetic distance between two plants, the more plants produced flowers at the same time.

The total number of flowers spatially co-occurring (i.e. in the same transect) during a given survey period (one to four in chronological order, for each year separately) were fitted in Figure 6. In panel a) the individuals that inhabit a transect according to their genetic group were fitted (Figure 6). In panel b) bars represent open flowers from each genetic group (red for A and blue for B) at a given survey day (Figure

6). Therefore red/blue bars represent transects where both groups had open flowers at the same time, i.e. the possibility for interbreeding. In a majority of transects at a given survey period flowers did not co-occur, or very few co-occurrences, in 2012 and 2013 (Figure 6). In 2012, transect TI had the highest number of co-occurring flowers of both groups at the same time, i.e. during survey period two, three and four (Figure 6, top part of panel b). In 2013, transects TD, TF and TL had co-flowering individuals (Figure 6, lower part of panel b). An interesting situation is shown in transect TF, corresponding to the contact zone, this transect is inhabited by a nearly equal number of individuals that belong to group A and B (Figure 6 panel a). Even so, in 2012 only individuals belonging to group B flowered. In 2013, a small and incomplete shift was observed, where individuals that belong to the group B flowered in majority during the first survey period followed by a flowering of individuals in the group A (few of group B) during the second survey period (Figure 6).

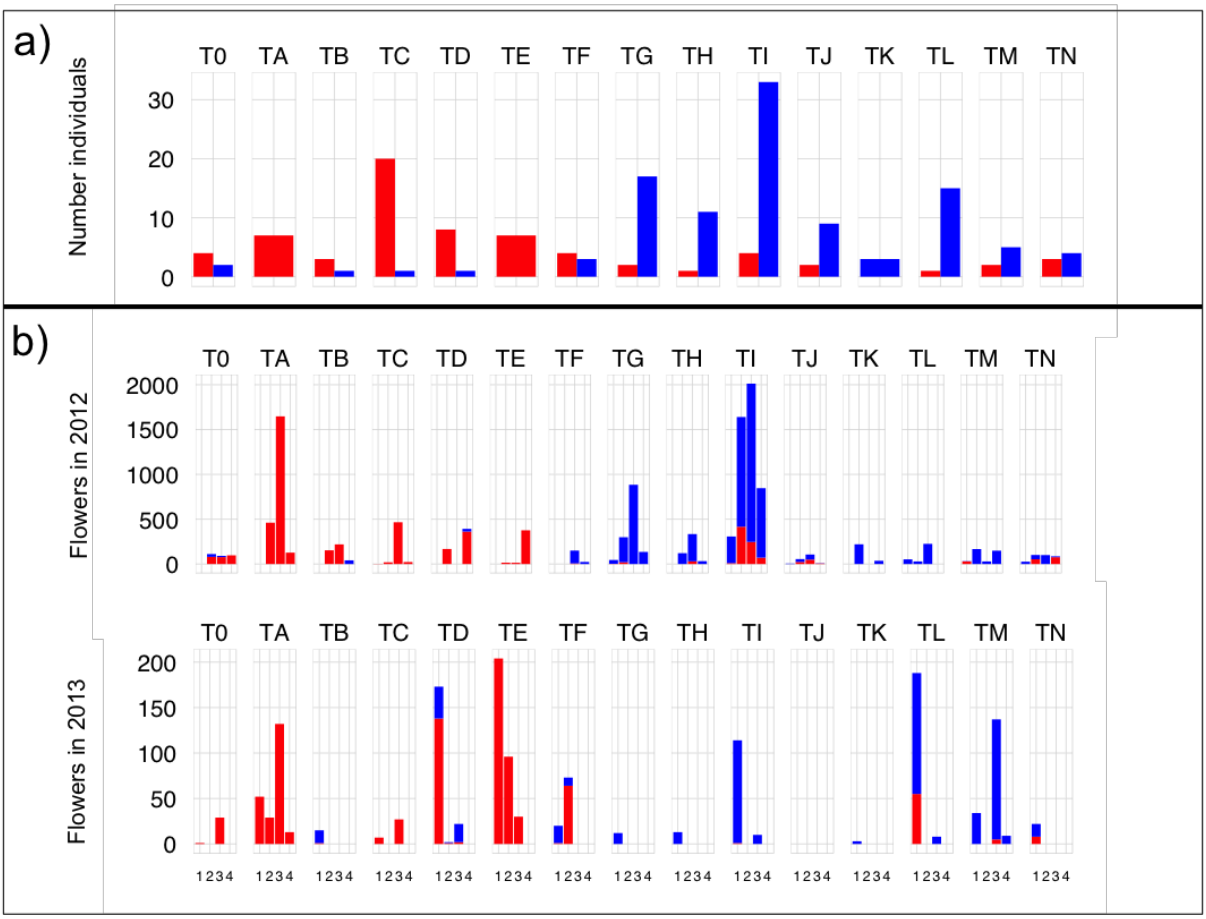


Figure 6: Co-flowering. In panel a) the number of individuals (n=212) from each genetic group (A in red, B in blue) in each transect. In panel b) a barplot of the number of open flowers during the two survey years (2012 top graph, 2013 bottom graph). In each graph from panel b) each transect is subdivided in survey periods (from one to four in chronological order, corresponding to survey periods in Figure 5). Colors represent open flowers of group A in red and of group B in blue. A bar with two colors represents co-flowering of individuals of both genetic groups at a given time in a given transect. The name of TY has been replaced by T0 for computational issues.

The number of immigrants was assessed, i.e. the number of individuals belonging to a genetic group (not hybrids) that were found on the « other » side of the hybrids zone in the other genetic background (in red in Table 2). Few immigrants were observed (n=14 immigrants of group B in “habitat” A and n=26 immigrant of group A in “habitat” B). The proportion of immigrants, on each side of the hybrid zone was not significantly different, i.e. no asymmetry was assessed.

Postzygotic reproductive isolation – indirect evidence

The indirect measure of hybrid unfitness in the natural population showed that a lower fraction, 21% of hybrids (29% A, 30% B; non significant Table 4) had at least one stem every year, this represents fitness but also the possibility to produce flowers and reproduce.

Table 4: Hybrid fitness. In the first line, the number of individuals that had at least one stem every year (out of the dataset of surveyed plants, n=212). In the second line the percentage over the total in each group with the 95% confidence interval. The bottom line shows the number of individuals in each group (A, B, hybrids and total).

	A	B	Hybrids	Total
Individuals with at least one stem every year	20	31	8	59
Percentage	29%	30%	21%	28%
95% CI	19.29-41.87	21.22-39.34	9.87-36.94	22.01-34.46
N	68	105	39	212

The analysis of habitat revealed four environmental variables that showed significant differences between A and B individuals (ruggedness, soil humidity, soil coverage and vegetation coverage, Table 5). No variable was discriminating B and hybrids, and only soil humidity showed significant differences between A individuals and hybrids (Table 5).

Table 5: Environmental difference between group A, B and hybrids. It was tested whether the seven environmental variables in the first column were different for individuals of genetic group A vs. B; of genetic groups A vs. hybrids and B vs. hybrids. The difference is TRUE when significant with Kruskal Wallis, otherwise, FALSE (n=187).

Variables	A vs. B	A vs. Hybrids	B vs. Hybrids
Ruggedness	TRUE	FALSE	FALSE
Slope	FALSE	FALSE	FALSE
Downslope distance gradient	FALSE	FALSE	FALSE
Soil humidity	TRUE	TRUE	FALSE
Soil coverage	TRUE	FALSE	FALSE
Vegetation coverage	TRUE	FALSE	FALSE
Stone coverage	FALSE	FALSE	FALSE

Stepwise variable selection in the GLM showed that four variables out of the seven tested (same as in Table 5) explained significantly the probability to belong to a genetic group (n=155, dataset without the 32 hybrids): ruggedness, slope, downslope distance gradient and soil humidity. This model, i.e. these four environmental variables explained 32.81% (R^2=0.3281) of variation to belong to one genetic groups or the other. All variation inflation factors were between 1.13 and 1.65, thus ascertaining a low collinearity that is not biasing GLM. The confidence interval around the spatial correlation curve was including or very close to zero, i.e. no or negligible spatial autocorrelation (Figure 7).

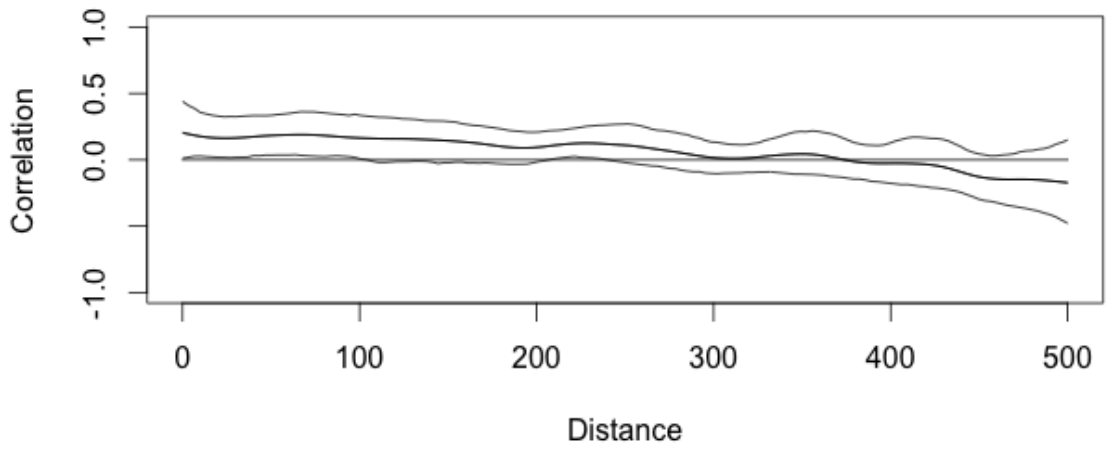


Figure 7: Spline correlogram. Correlogram of Pearson residuals of the GLM (on the y-axis), against distance between individuals on the y-axis (until 500m). The confidence interval around the spline demonstrates a very low degree of spatial autocorrelation, i.e. zero means no autocorrelation.

The four environmental variables included in the GLM were all significant, with ruggedness, slope and soil humidity being highly significant (pvalue <0.01, Table 6).

Table 6: GLM results. Results for the GLM after stepwise variable selection (n=155) including ruggedness, slope, downslope distance gradient and soil humidity. The second and third columns are the estimate and standard error from the binomial GLM. The two last columns represent the test of significance for the variables, i.e. the result of Chi-square test and its significance.

	Estimate	SE	Chisq	Pr(>Chisq)
Intercept	7.098	1.843	-	-
Ruggedness	4.758	1.167	21.830	< 0.01
Slope	-7.311	2.277	12.563	< 0.01
Downslope distance gradient	-5.477	2.438	5.261	0.02
Soil humidity	-4.585	1.280	14.632	< 0.01

By assessing the predicted probability provided by GLM to belong to a genetic group for each variable separately it was shown that high values of ruggedness implied high probability of being from genetic group A, whereas a high downslope distance gradient, slope and soil humidity increased the probability to belong to group B (Figure 8). For the downslope gradient and slope important confidence intervals were observed.

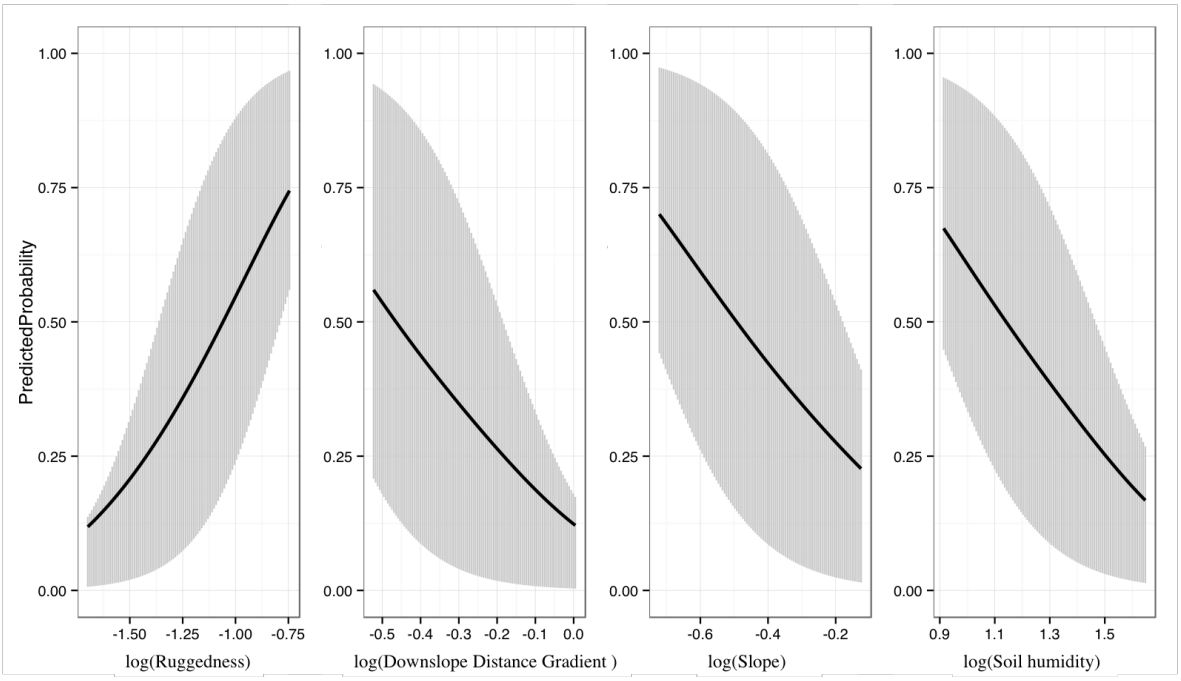


Figure 8: Predicted probabilities to belong to group A or B (n=155). Predicted probabilities, i.e. results from GLM to belong to one genetic group given (from left to right): ruggedness, downslope distance gradient, slope and soil humidity values. The grey area around the curve represents the 95% confidence interval.

In the GLMM with area used as random effect, removing any trace of spatial autocorellation due to sampling in clusters, slope and downslope distance gradient were only marginally significant whereas ruggedness and soil humidity remained highly significant (Table 7).

Table 7: GLMM results. Results for the binomial GLMM, variables are the same as selected after stepwise variable selection in the GLM before (n=155; ruggedness, slope, downslope distance gradient and soil humidity). The random effect is area. The second and third columns are the estimate and standard error from the binomial GLMM. The two last columns represent the test of significance for the variables, i.e. the result of Chi-square test and its significance.

	Estimate	SE	Chisq	Pr (>Chisq)
Intercept	8.22	3.45	1.5	-
Ruggedness	5.488	2.15	1.03	0.01
Slope	-6.641	3.69	1.7	0.07
Downslope distance gradient	-8.266	4.426	2.1	0.06
Soil humidity	-5.532	2.238	1.07	0.01

Postzygotic reproductive isolation – experimental evidence

Direct evidence of postzygotic reproductive isolation was obtained comparing seed weight between pure and hybrids F1 as well as with the reciprocal transplant experiment. A significantly lower seed weight of hybrid crosses was assessed (p-value<0.01).

The different hypotheses used to test for local adaptation with the reciprocal transplant experiment are shown in Figure 9. On the top panel, hypothesis one, i.e. home/away suggests two habitats, represented by transplant block BC, BE and BG, BL respectively (Figure 9). The underlying hypothesis is that each side of the hybrid zone (western, eastern) is inhabited in majority by a specific genetic (group, A or B). Here, the intra A cross type (Table 3) is considered home in the red transplant blocks (BC and BE in the top panel of Figure 9), where the natural population is mainly represented by the genetic group A (red bars in Figure 9). The intra B crosses (Table 3) are considered

home in the blue transplant blocks (BG and BL in the top panel of Figure 9), this is the part of the crest on the eastern side of the hybrid zone, inhabited by a majority of individuals that belong to group B (blue bars in Figure 9).

On the lower panel of Figure 9, corresponding to hypothesis two (i.e. native/ non-native) three habitats were differentiated as revealed by the habitat analysis (Figure 1). In the brown transplant block (BC, bottom panel of Figure 9) individuals were considered home, when both their parents originated in the habitat represented by quadrant II of the BPCA (Figure 1), i.e. intra_RC crosses (Table 3). In the orange transplant blocks (BE and BG, bottom panel of Figure 9) individuals were considered native, when both their parents originated in the habitat represented by quadrant III of the BPCA (Figure 1), i.e. intra_REG cross (Table 3). In the green transplant block (BL, bottom panel of Figure 9) individuals were considered native, when both their parents originated in the habitat represented by quadrant IV of the BPCA (Figure 1), i.e. intra_RL cross (Table 3).

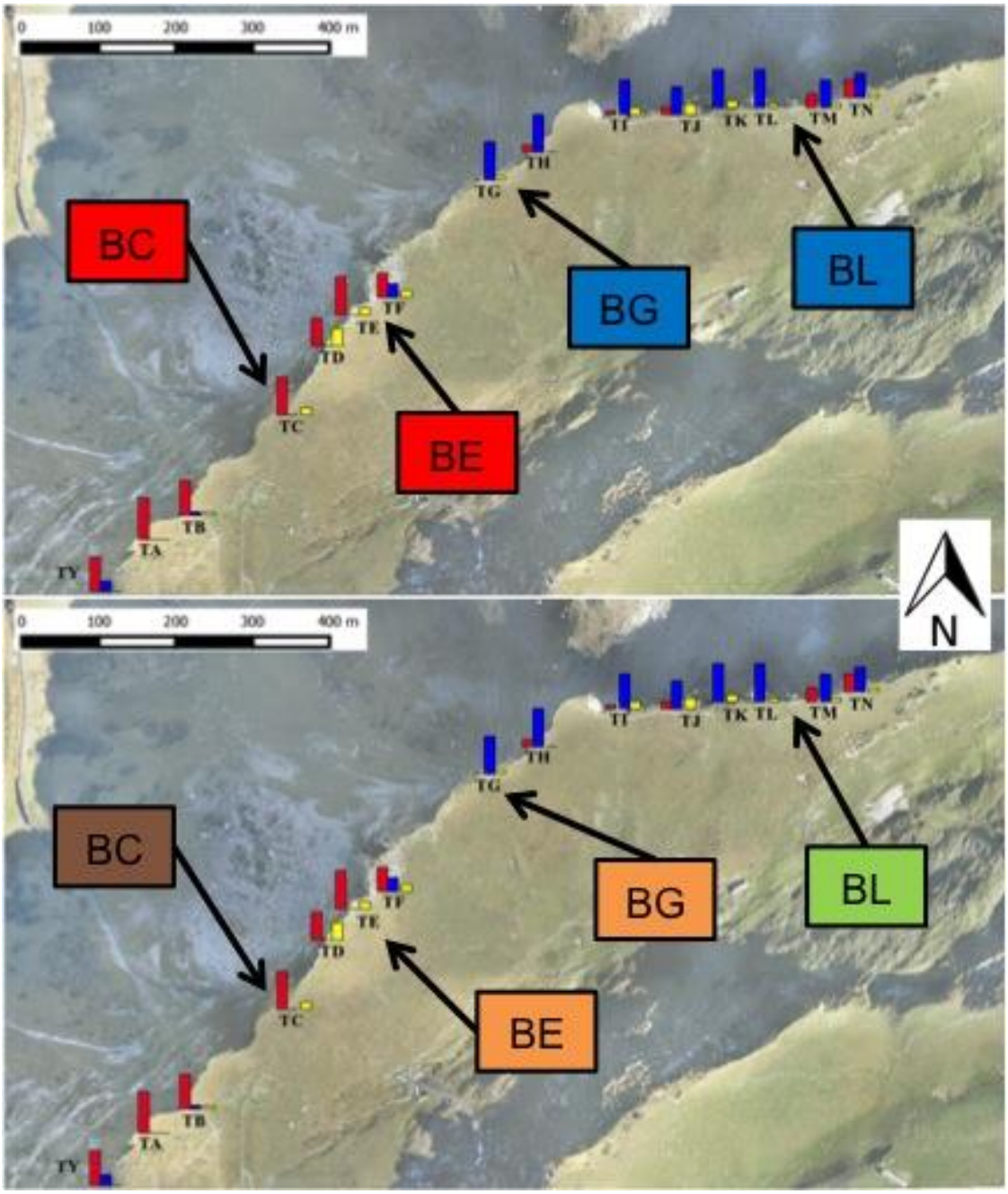


Figure 9: The four transplant blocks. The background map represents the studied population with transects and the proportion of individuals that belongs to group A (red), group B (blue) and hybrids (yellow) in the natural population. On the top panel the four transplant blocks (BC, BE, BG, BL) with colors corresponding to the hypothesis one, i.e. home/ away, with red blocks representing home habitat of intra A group and blue blocks home habitat of intra B group. The colors in the bottom panel correspond to the hypothesis two, native/ non-native, with the brown blocks (BC) representing habitat as described by the second quadrant of BPCA space, orange blocks (BE and BG) by the third quadrant and the green block (BL) by the third quadrant.

Table 8 summarizes the individuals surveyed in each block and those that survived or were considered dead one year after the transplant experiment. A total of 84 individuals in block BC, 82 in block BE, 80 in block BG and 87 in block BL was surveyed until 2013. The number of parental pairs pooled in each cross type can be found in Table 3. For the home/away hypothesis individual numbers in each cross type were balanced as it corresponded to to targeted cross design, i.e. 107 individuals in intra A, 122 in intra B and 104 in hybrids. For the native non-native hypothesis the number of individuals ranged between 29 and 104 in a cross type. Individuals that survived the first transplant year ranged between 80 and 87 depending of the transplant block (Table 8).

Table 8: Transplanted individuals surveyed one year after the transplant experiment. The first column defines the hypothesis that was tested, i.e. home/away (habitat differs on each side of the hybrid zone) vs. native/non-native (habitat differs also between transects). The second column describes the cross types according to the tested hypothesis, numbers in bold represent a home or native cross-type in a given transplant block. The third column represents the number of individuals in each cross type. The right side is divided in transplant blocks (BC, BE, BG, CL), with total number of individuals, number of individuals considered surviving and dead.

			Block BC			**Block BE**			**Block BG**			**Block BL**		
Hypotheses	Cross type	Ind	N tot	N surv	N dead	N tot	N surv	N dead	N tot	N surv	N dead	N tot	N surv	N dead
Home/ Away	Genetic intra A	107	**29**	**14**	**15**	**28**	**11**	**17**	26	13	13	24	14	10
	Genetic intra B	122	29	9	20	29	10	19	**29**	**14**	**15**	**35**	**19**	**16**
	Hybrids	104	26	11	15	25	11	14	25	8	17	28	17	11
Native/ Non-native	intra RC	87	**24**	**10**	**14**	22	10	12	20	11	9	21	11	10
	intra REG	29	8	1	7	**7**	**3**	**4**	**6**	**2**	**4**	8	3	5
	intra RL	47	11	4	7	11	2	9	11	4	7	**14**	**6**	**8**
	none	66	15	8	7	17	6	11	18	10	8	16	13	3
	Hybrids	104	26	11	15	25	11	14	25	8	17	28	17	11
Tot		333	84	34	50	82	32	50	80	35	45	87	50	37

Comparing all crosses together in each transplant block separately between 39.02-57.47% of individuals survived one year after the transplant (bottom line of Table 9). The highest survival was observed in block BL (bottom line inTable 9). Testing the first hypothesis, i.e. home/away, neither significant survival differences were found between home and away cross types, nor between home and hybrids. Even so, a trend showed a higher survival of home individuals in transplant blocks BC and BE (respectively 48.28%, vs. 31.03 % away individuals and 39.28%, vs. 34.48%, not significant (top of Table 9). In Blocks BG and BL differences were very low between home and away individuals. In blocks BC and BG hybrids showed a lower survival than home crosses (not significant), but they showed a higher survival in BE and BL (not significant, see top of Table 9).

The second hypothesis, i.e. native/ non- native, focused on individuals for which both parents originated in a specific habitat tested by the transplant experiment. Here, the only marginally significant difference (p-value=0.06) in survival was between native individuals of Block TC and individuals originating from transect TG (intra REG cross, p-value=0.06, 41.66% vs. 12.5%, see lower part of Table 9). Since only one individual survived in the away cross, results of the statistical test are probably biased. Confidence intervals are generally very large, due to the small differences between survival proportions and the low sample size.

Table 9: Relative survival one year after the transplant experiment. The first column defines the hypothesis that was tested, i.e. home/away (habitat differs on each side of the hybrid zone) vs. native/non-native (habitat differs also between transects). The second column describes the cross types according to the tested hypothesis, numbers in bold represent a home or native cross-type in a given transplant block. In each subdivision, representing transplant blocks (BC, BE, BG, CL), the first column represents the relative survival proportion with the 95% confidence interval in brackets, the category in bold represents the home, respectively native cross type.

Hypotheses	Description	**Survival in Block BC**	**Survival in Block BE**	**Survival in Block BG**	**Survival in Block BL**
Home/ Away	Genetic intra A	**48,28%** (29,89-67,10)	**39,29%** (22,13-59,27)	50% (32,06-67,94)	58,33% (36,94-77,20)
	Genetic intra B	31,03% (15,98-50,95)	34,48% (18,60-54,34)	**48,28%** (29,89-67,10)	**54,29%** (36,87-70,78)
	Hybrids	42,31% (23,97-62,81)	44% (25,02-64,73)	32% (15,73-53,55)	60,71% (40,73-77,87)
Native/ Non-native	intra RC	**41,67%** (22,80-63,06)	45,45% (25,07-67,33)	55% (32,05-76,17)	52,38% (30,34-73,61)
	intra REG	12.5% (0,66-53,32)	**42,86%** (11,81-79,76)	**33,33%** (6,00-75,89)	37,5% (10,24-74,11)
	intra RL	36,36% (12,37-68,39)	18,18% (3,21-52,24)	36,36% (12,37-68,39)	**42,86%** (18,81-70,35)
	none	53,33% (27,42-77,72)	35,29% (15,26-61,38)	55,56% (31,35-77,60)	81,25% (53,69-95,03)
	Hybrids	42,31% (23,97-62,81)	44% (25,02-64,73)	32% (15,73-53,55)	60,71% (40,73-77,87)
Tot	Mean survival	40.05%	39.02%	44.75%	57.47%

Three performance variables were tested for individuals in the field. First, differences in number of leaves were tested, i.e. the number of leaves one year after transplant minus number of leaves at day zero of transplant, thereby taking into account differential growth in a common environment. This overall differences yielded increase but also decrease in leaf numbers. Small differences were revealed between blocks with block BE having the lowest increase (0.78) and BL the highest (1.94, bottom line Table 12). No significant difference in home (respectively native) vs. away (respectively non-native) or hybrids was assessed. No trend could be observed.

Table 10: Difference in number of leaves from transplant until one year after transplant experiment for individuals that survived the first year. The first column defines the hypothesis that was tested, i.e. home/away, native/non-native. The second column describes the cross type. The right side is divided in transplant blocks (BC, BE, BG, BL). In each subdivision, the first column represents the number of individuals that survived one year after transplant, i.e. for which performance could be calculated and tested. The second column represents mean leaf difference followed by its standard deviation. The bottom row represents mean growth in the transplant block.

Hypotheses	Description	Block BC			Block BE			Block BG			Block BL		
		N	mean	sd	N	mean	sd	N	mean	sd	N	mean	sd
Home/ Away	Genetic intra A	14	**0.55**	2.02	11	**0.36**	0.92	13	1.38	2.53	14	2.00	1.18
	Genetic intra B	9	1.44	2.07	10	1.50	0.97	14	**1.14**	2.03	19	**1.42**	1.90
	Hybrids	11	2.27	3.66	11	0.55	1.64	8	1.25	1.83	17	2.47	2.04
Native/ Non- native	intra RC	10	**0.80**	2.10	10	0.40	0.97	11	1.55	2.73	11	1.91	1.30
	intra REG	1	-1.00	NA	3	**0.67**	1.15	2	**1.50**	0.71	3	-0.33	0.58
	intra RL	4	2.00	2.94	2	2.50	0.71	4	1.50	3.00	6	**0.83**	1.84
	none	8	0.70	1.56	6	1.33	0.82	10	0.80	1.69	13	2.31	1.55
	Hybrids	11	2.27	3.66	11	0.55	1.64	8	1.25	1.83	17	2.47	2.04
Tot	Mean		1.34			0.78			1.26			1.94	

The second measured fitness variable was the difference in leaf size until one year after transplant for plants that survived the first transplant year. Here again, I didn't observe an overall growth and differences yielded growth but also decrease in leaf size.

Here, mean differences by blocks showed that in block BG plants had the tendency to decrease the leaf size over time (mean= -0.42, Table 11), whereas in block BL a small increase was observed (mean=0.4). No significant difference in home (respectively native) vs. away (respectively non-native) or hybrids was assessed.

Table 11: Leaf size differences after the first transplant year. The first column defines the hypothesis that was tested, i.e. home/away, native/non-native. The second column describes the cross type. The right side is divided in transplant blocks (BC, BE, BG, BL). In each subdivision, the first column represents the number of individuals that survived one year after transplant, i.e. for which performance could be calculated and tested. The second column represents mean leaf size difference followed by its standard deviation. The bottom row represents mean leaf size difference in the transplant block.

Hypotheses	Description	Block BC			Block BE			Block BG			Block BL		
		N	mean	sd	N	mean	sd	N	mean	sd	N	mean	sd
Home/ Away	Genetic intra A	14	**-0.21**	1.44	11	**-0.33**	0.75	13	-0.83	1.00	14	0.38	0.77
	Genetic intra B	9	1.11	1.99	10	0.57	0.81	14	**-0.02**	0.96	19	**0.71**	1.05
	Hybrids	11	-0.29	1.60	11	-0.13	0.87	8	-0.46	0.83	17	0.07	1.03
Native/ Non-native	intra RC	10	**-0.61**	1.10	10	-0.29	0.78	11	-0.89	1.04	11	0.26	0.81
	intra REG	1	0.40	NA	3	**0.00**	0.87	2	**-0.05**	0.92	3	0.07	0.84
	intra RL	4	-0.23	2.01	2	1.00	0.14	4	0.30	0.91	6	**0.55**	0.69
	none	8	1.71	1.68	6	0.50	0.96	10	-0.24	1.01	13	0.95	1.09
	Hybrids	11	-0.29	1.60	11	-0.13	0.87	8	-0.46	0.83	17	0.07	1.03
Tot	Mean survival		0.12			0.02			-0.42			0.40	

The third performance variable was chosen to represent all life stages from growth in a common environment, including mortality and growth in nature. Therefore, the number of leaves one year after the transplant experiment was tested for each cross type, including individuals considered dead (i.e. zero leaves). The mean leaf size was the highest in Block BL (2.52cm) blocks BC, BE and BG had similar leaf sizes (respectively 1.54, 1.32, 1.61cm). No significant difference in home (respectively native) vs. away (respectively non-native) or hybrids was assessed. No trend could be observed. Only a single inflorescence was recorded and therefore is not further discussed here.

Table 12: Leaf size. Leaf size one year after transplant, dead individuals had zero leaves. The first column defines the hypothesis that was tested, i.e. home/away, native/non-native. The second column describes the cross type. The right side is divided in transplant blocks (BC, BE, BG, BL). In each subdivision, the first column represents the total number of individuals surveyed one year after transplant. The second column represents mean leaf size in a cross type followed by the standard deviation. The bottom row represents mean leaf size in a given transplant block.

Hypotheses	Description	Block BC			Block BE			Block BG			Block BL		
		N	mean	sd	N	mean	sd	N	mean	sd	N	mean	sd
Home/ Away	Genetic intra A	**29**	**1.5**	**2.01**	**28**	**1.07**	**1.46**	26	2	2.83	24	2.67	2.55
	Genetic intra B	29	1.21	2.19	29	1.41	2.13	**29**	**1.76**	**2.25**	**35**	**2.09**	**2.37**
	Hybrids	26	1.96	3.34	25	1.48	1.94	25	1.04	1.84	28	2.93	2.92
Native/ Non-native	intra RC	**24**	**1.29**	**1.99**	22	1.23	1.51	20	2.3	3.06	21	2.29	2.49
	intra REG	8	0.25	0.71	**7**	**1.57**	**2.07**	**6**	**1.33**	**2.07**	8	0.88	1.25
	intra RL	11	1.73	2.94	11	1	2.32	11	1.55	2.62	**14**	**1.71**	**2.3**
	none	15	1.77	1.97	17	1.29	1.9	18	1.78	1.99	16	3.63	2.47
	Hybrids	26	1.96	3.34	25	1.48	1.94	25	1.04	1.84	28	2.93	2.92
Tot	Mean survival		1.54			1.32			1.61			2.52	

Discussion

Anonymous genome-wide markers highlighted a hybrid zone between two diverging genetic groups of the polyploid *Biscutella laevigata*. To identify reproductive barriers triggering this divergence I combined indirect survey methods in the field with an experimental transplant experiment in nature. Four main conclusions emerged from my analyses. First, two diverging groups were clearly demonstrated, excluding the possibility of homoploid hybrid speciation. Second, a strong spatial component was revealed, where "A" and "B" groups inhabited significantly the western or the eastern parts of the ridge. Third, a prezygotic barrier was revealed in space, suggesting that this is probably a major force shaping the genetic structure in this population as it is often the case. Fourth, postzygotic extrinsic barriers, i.e. depending from the environment could also play a role in maintaining this divergence.

The genetic structure clearly demonstrated two diverging groups. The contact zone between these groups was coinciding with an area with very few plants in the center of the population. Genetic group A inhabited in majority the western part of the contact zone and group B the eastern part.

Over the whole crest, flowering time assessed by peaks of open flowers was not shifted in time between both groups, theoretically allowing interbreeding. Comparing the number of flowers spatially, it becomes clear that in a given transect (geographically close individuals) very few flowers of both groups co-occur at the same period or flowering was shifted between groups in time. This decreases dramatically the chances to interbreed and produce hybrids. In most transects, even so inhabited by individuals from both genetic groups, few co-flowering between groups was assessed. A low, but

significant correlation showed that genetically distant individuals had a lower tendency to produce flowers at the same time. During the second year, the flowering-genetic association could not be confirmed and the overall genetic basis for flowering time remains puzzling. This raises the question about reinforcement in this population, i.e. direct natural selection that strengthen prezygotic isolation (Widmer et al. 2009). A long-term experimental study (started 1856) of the grass *Anthoxanthum odoratum*, showed that flowering time has shifted at boundaries, i.e. in the contact zone and provided compelling evidence for reinforcement (Silvertown et al. 2005). Reinforcement in this population could explain the low overall genetic-flowering synchrony association because a strong correlation would only be present at the boundaries between the groups. A complementary evidence for the genetic basis of flowering time, is the flowering differences observed during the cross experiment. In the common garden, important shifts were observed, not quantified, and some plants were manipulated (stems cut) to force some crosses. The genetic basis of flowering time should be assessed in the future by comparing flowering time of individuals in the contact zone in a common environment.

The influence of yearly variations on divergence remains to be addressed. One hypothesis to explain the lower genetic-flowering synchrony in the second year could rely on different flowering patterns between survey years, e.g. a short flowering season in 2013. Different patterns could be caused by a long winter in 2013, for certain areas near the main contact zone of the hybrid zone, the snow cover lasted very long and delayed the formation of the first leaves and flowers. For two plots in this area, no flowers were produced (in transect TF). To what extent a yearly varying long snow cover at the interface of both groups can promote divergence remains to be addressed.

Immigrant inviability, i.e. a prezygotic barrier *sensu* (Nosil et al. 2005) was assessed. Few immigrants are found in their foreign habitat; in addition each genetic group inhabits a specific region of the crest, lowering chances for interbreeding. Therefore, the low number of immigrants, the geographical low-density area in the contact zone and the low spatial co-flowering between individuals of both groups suggests an important role for prezygotic barriers in triggering the maintenance in divergence between both groups. This is in accordance with literature that shows that prezygotic isolation is often the major contributor shaping reproductive isolation (Lowry et al. 2008; Widmer et al. 2009).

Postzygotic reproductive isolation was assessed by field survey, i.e. indirect evidence and a reciprocal transplant experiment, i.e. direct evidence. In the field, environments were better discriminated for pure individuals than for hybrids, indicating that hybrids could inhabit intermediate or unspecific habitats (Nosil 2012). In addition the low proportion of hybrids in nature indicates hybrid inviability. In addition, a first evidence of hybrid unfitness was suggested by the lower number of stems in hybrids, even so not significantly lower than for pure individuals.

For pure individuals it was demonstrated that the specific habitat of individuals that belong to a given group is best discriminated by four environmental variables out of the seven tested, i.e. ruggedness, slope, downslope distance gradient and soil humidity. This anaylsis revealed that the habitat of individuals of group A was characterized by a high ruggedness and low slope, i.e. uneven rather flat terrain as well as a low downslope distance gradient and soil humidity. The downslope distance gradient represents a proxy for drainage, i.e. short-term water displacement, whereas soil humidity probably represents the long-term capacity of the soil to retain water. The habitat analysis

demonstrated a negative correlation between soil humidity and soil cover. Therefore the habitat of individuals of group A is an arid uneven terrain with very high soil cover, i.e. low vegetation cover. On the contrary, the habitat of B individuals is a steep, humid habitat with plenty vegetation. Coverage (e.g. percentage of soil or vegetation) is a variable that can easily be observed in the field, in addition soil humidity being strongly correlated with vegetation coverage this variable could be derived. The DEM-derived variables describe important topographic measures putatively representing important abiotic factors, i.e. drainage and unevenness. These factors calculated with a precision of 50cm are nearly impossible to observe in nature. The availability of these measures in combination with direct measures and observations strongly increased the understanding of the habitat and allowed a good description from the different niches of both genetic groups. Indirect survey provide evidence for a postzygotic extrinsic reproductive isolation triggering divergence between both genetic groups.

Experimental evidence for postzygotic reproductive isolation was first assessed by a lower seed weight of hybrids. This is a first evidence for a possible hybrid dysfunction, i.e. postzygotic barrier, but there was no influence of seed weight on later performance of the hybrids. Therefore, this barrier is probably negligible. The experimental hybrid crosses could not reveal any early-acting barrier since all cross types could be performed and produced seeds, but it would be necessary to quantify this.

The experimental transplant of hybrids in the field could not reveal any significant postzygotic barrier, neither extrinsic nor intrinsic. The analysis of postzygotic immigrant inviability was assessed by the survival and performance of away, i.e. non-native individuals. In the most western transplant block a trend was observed indicating

a higher survival for home, i.e. native individuals suggesting a mild and asymmetric role for natural selection. Even so, it could not be demonstrated that crosses with parents originating from this region did perform less good in the other transplant blocks. Taking into account the different performance variables, it was shown that the most eastern transplant block, i.e. BL, showed the highest overall survival, highest positive difference in leaf number and growth of leaves. A high vegetation cover and important soil humidity characterizes this block. This evidence could indicate a confounding effect of habitat, i.e. whether one cross type outperforms the other due to local adaptation or simply because one habitat is generally "better" (e.g. more nutrient, less competition) remains unclear and local adaptation cannot be confirmed (Kawecki and Ebert 2004).

One explanation for the failure of the reciprocal transplant experiment to reveal clear pattern of local adaptation could be linked to polyploidy. Whereas most studies focus on the advantages of increased polyploidy, or advantages in hybrid speciation forming allopolyploids very few tackle the outcome of hybridization inside autopolyploids (Parisod et al. 2010). Lower response to selection as well as genetic redundancy is predicted for autopolyploids with tetrasomic inheritance (Ronfort et al. 1998; Parisod et al. 2010). In what extent genetic redundancy is even increased in the hybrids allowing heterosis, similarly as in hybrid polyploid speciation remains to be addressed. This could not strictly be addressed here as the reasons for the failure of the experiment could also be due to low sample size. Even so, the occurrence of few hybrids in the field and specialized groups tends to prefer the hypothesis that the response of autopolyploids could be generally low, be it in hybrids or pure crosses, possibly delaying their response to selection. *Biscutella laevigata* is a perennial plant, plenty individuals in the field experienced selective forces for several years, whereas only few transplanted

individuals had one year to respond to the similar forces. The implications of the lower response to selection remain to be addressed. Another explanation comes from the experiment *per se*, a low sample size and a high nearly symmetric (in the different cross categories) mortality in young plants until one year after the transplant lowered the chances for statistically meaningful inferences.

"Hybrid zones can be viewed as natural arenas in which combinations of mutations that have accumulated between species can be tested" (Payseur 2010). The question remaining is which mutations have differentially accumulated. Determining the loci underlying reproductive isolation and their linkage would ultimately allow confirming evidence gathered in this study.

In this study I could demonstrate a hybrid zone between two diverging groups of *Biscutella laevigata* that are not completely reproductively isolated. Whereas a major force shaping this divergence is a low co-flowering of both groups in space, the role of natural selection underlying this divergence remains to be clarified. Especially, in the autopolyploid context, the analysis of genome reorganization in the hybrids as well as a possibly low response to selection represents a promising framework to address questions about the success of polyploids.

Acknowledgments

I thank Stéphane Joost and Kevin Leempoel for providing the digital elevation model and the DEM-derived variable as well as constructive comment during the regular meetings. I addition I thank Radu Alexandru Slobondeanu for support in statistics, Natacha Senerchia for the collaboration on bootstrapping scripts and Philippa Griffin, Olivier Bachmann, Tobias Richter, Emmanuel Bonjour, Audrey Ueberschlag, Alexandre Tayalé, Natacha Senerchia and Amélie Bardil for support in the field. I thank Yvonne Willi and Felix Kessler for comments on the manuscript. The project was financed by the Swiss National Science Foundation and the Velux Stiftung.

Literature Cited

Barton, N. H. and M. A. R. de Cara. 2009. The evolution of strong reproductive isolation. Evolution 63:1171-1190.

Barton, N. H. and G. M. Hewitt. 1985. Analysis of hybrid zones. Annual Review of Ecology and Systematics 16:113-148.

Bates, D., M. Maechler, B. Bolker, S. Walker, S. Walker, R. H. B. Christensen, and H. Singmann. 2013. lme4 package: Linear mixed-effects model using Eigen and S4.

Bjornstad, O. N. 2013. ncf package: Spatial nonparametric covariance functions.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127-135.

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

Buerkle, C. A., D. E. Wolf, and L. H. Rieseberg. 2003. The origin and extinction of species through hybridization. Population viability in plants: Conservation, management, and modeling of rare plants Volume 165:117-141.

Burke, J. M. and M. L. Arnold. 2001. Genetics and the fitness of hybrids. Annual Review of Genetics 35:31-52.

Calinski, T. and J. Harabasz. 1974. A dendrite method for cluster analysis. Communications in Statistics 3.

Coyne, J. A. and H. A. Orr. 2004. Speciation. Sinauer, Sunderland, MA.

Darwin, C. 1859. On the Origin of Species by Means of Natural Selection or the Preservation of Favored Races in the Struggle for Life, J. Murray, London.

Development Core Team R. 2013. R: a language and environment for statistical computing. Vienna, Austria.

Dufresne, F., M. Stift, R. Vergilino, and B. K. Mable. 2014. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Molecular Ecology 23:40-69.

Felsenstein, J. 1981. Skepticism towards Santa Rosalia, or why are there so few kinds of animals. Evolution 35:124-138.

Giraudoux, P. 2013. pgirmess: data analysis in ecology.

Gompert, Z., L. K. Lucas, J. A. Fordyce, M. L. Forister, and C. C. Nice. 2010. Secondary contact between *Lycaeides idas* and *L-melissa* in the Rocky Mountains: extensive admixture and a patchy hybrid zone. Molecular Ecology 19:3171-3192.

Hjerdt, K. N., J. J. McDonnell, J. Seibert, and A. Rodhe. 2004. A new topographic index to quantify downslope controls on local drainage. Water Resources Research 40.

Kawecki, T. J. and D. Ebert. 2004. Conceptual issues in local adaptation. Ecology Letters 7:1225-1241.

Lowry, D. B., J. L. Modliszewski, K. M. Wright, C. A. Wu, and J. H. Willis. 2008. The strength and genetic basis of reproductive isolating barriers in flowering plants. Philos T R Soc B 363:3009-3021.

Manton, I. 1937. The problem of *Biscutella laevigata* L. II. The evidence from meiosis. Annals of Botany 1:439-U414.

Martin, N. H. and J. H. Willis. 2007. Ecological divergence associated with mating system causes nearly complete reproductive isolation between sympatric *Mimulus* species. Evolution 61:68-82.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. 2014. e1071: Misc Functions.

Nagelkerke, N. J. D. 1991. A Note on a General Definition of the Coefficient of Determination. Biometrika 78:691-692.

Nakagawa, S. and H. Schielzeth. 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biological Reviews 85:935-956.

Nosil, P. 2012. Ecological Speciation.

Nosil, P., T. H. Vines, and D. J. Funk. 2005. Perspective: Reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution 59:705-719.

Olowokudejo, J. D. and V. H. Heywood. 1984. Cyto-taxonomy and breeding system of the genus *Biscutella* (Cruciferae). Plant Systematics and Evolution 145:291-309.

Parisod, C. and P. A. Christin. 2008. Genome-wide association to fine-scale ecological heterogeneity within a continuous population of *Biscutella laevigata* (Brassicaceae). New Phytologist 178:436-447.

Parisod, C., R. Holderegger, and C. Brochmann. 2010. Evolutionary consequences of

autopolyploidy. New Phytologist 186:5-17.

Parisod, C. and S. Joost. 2010. Divergent selection in trailing- versus leading-edge populations of *Biscutella laevigata*. Annals of Botany 105:655-660.

Payseur, B. A. 2010. Using differential introgression in hybrid zones to identify genomic regions involved in speciation. Molecular Ecology Resources 10:806-820.

Ramsey, J., H. D. Bradshaw, and D. W. Schemske. 2003. Components of reproductive isolation between the monkeyflowers Mimulus lewisii and M-cardinalis (Phrymaceae). Evolution 57:1520-1534.

Rieseberg, L. H. and J. H. Willis. 2007. Plant speciation. Science 317:910-914.

Romesburg, H. C. 1985. Exploring, confirming and randomization tests. Computers and Geosciences:19-37.

Ronfort, J. L., E. Jenczewski, T. Bataillon, and F. Rousset. 1998. Analysis of population structure in autotetraploid species. Genetics 150:921-930.

Ross, R. I. C., J. A. Agren, and J. R. Pannell. 2012. Exogenous selection shapes germination behaviour and seedling traits of populations at different altitudes in a Senecio hybrid zone. Annals of Botany 110:1439-1447.

Rundle, H. D. and P. Nosil. 2005. Ecological speciation. Ecology Letters 8:336-352.

Sappington, J. M., K. M. Longshore, and D. B. Thompson. 2007. Quantifying landscape ruggedness for animal habitat analysis: A case study using bighorn sheep in the Mojave Desert. Journal of Wildlife Management 71:1419-1426.

Schluter, D. 2001. Ecology and the origin of species. Trends in Ecology & Evolution 16:372-380.

Scopece, G., A. Croce, C. Lexer, and S. Cozzolino. 2013. Components of reproductive isolation between *orchis mascula* and *orchis pauciflora* Evolution 67:2083-2093.

Silvertown, J., C. Servaes, P. Biss, and D. Macleod. 2005. Reinforcement of reproductive isolation between adjacent populations in the Park Grass Experiment. Heredity 95:198-205.

Smajda, B. 2011. A framework for comparing processes of speciation in the presence of gene flow. Molecular Ecology 20:5123-5140.

Widmer, A., C. Lexer, and S. Cozzolino. 2009. Evolution of reproductive isolation in plants. Heredity 102:31-38.

General Conclusion and Outlook

This thesis focused on two main questions in evolutionary biology, about the role of whole genome duplications in plant speciation and the evolution of reproductive isolation. This study took advantage of modern high throughput sequencing technologies, extensive field surveys and experimental approaches.

Studies in non-model organisms are challenging, not only due to the lack of general knowledge (e.g. evolutionary origin, physiology, morphology, mode of inheritance) but also because of a lack of genomic resources compared to model-organisms or economically relevant species. The transcriptome provided in Chapter 1 allowed tackling the history of genome duplication in an ecologically important species. The common history of duplication and the phylogenetic relatedness with *Arabidopsis thaliana* allowed taking advantage of the important resources (e.g. gene annotation, synteny, functional annotations) produced for this model organism.

The identification of several whole genome duplications revealed in *Biscutella laevigata* paved the way to new insights about functional categories preferentially retained and the role of selection in their retention. Whereas the fact that most pairs are underlying purifying selection confirms earlier studies (Kondrashov et al. 2002), the functional analyses and underlying selection pressures represent advances in this field. One major finding is the functional bias in retained genes under strong purifying selection retained after the α-WGD and the B-WGD. Genes under purifying selection probably represent essential genes protected from deleterious mutations. These categories were analyzed with the whole transcriptome as baseline and it was striking that most genes under purifying selection are associated to stress response. It is debated whether adaptive ecological niche differentiation between polyploids and their diploid progenitors plays an important role in the successful establishment and radiation of

polyploids (Glennon et al. 2014). The findings here could suggest that redundancy after duplication was used to overcome challenges to a changing environment after the last glacial maximum allowing adaptive radiation and/or niche differentiation from the diploid.

Purifying selection indicates that "old" gene functions were probably "protected" of deleterious mutations by selection. Because they are retained as duplicates, they can alter the phenotype of polyploids (Otto and Whitton 2000; Comai 2005). But for lineages to diverge and radiation to be successful the evolution of novel traits can be a key feature. Therefore the categories being putatively under positive selection, indicating the emergence of new advantageous functions were of main interest here. Categories under positive selection after the α- and B-WGD were related to cytoskeleton, some categories specifically involved in root development. It will be crucial in the future to assess whether these categories could be involved in the adaptation to different stresses and supported the colonization of the Alps.

A hypothesis, the lag-time model, was postulated recently (Schranz et al. 2012). It could be shown that whole genome duplications arise before key radiations, e.g. angiosperms (Jiao et al. 2011). Even so, phylogenetic asymmetries with a species-rich main group and species-poor sister clades highlighted the possibility of a substantial amount of time between whole genome duplications and the evolution of key novelties (Schranz et al. 2012). My data could support this hypothesis, since more categories are putatively diverging in the older α- WGD than in the more recent B-WGD.

Main limitations in this study came from the assessment of forces underlying retention inside a transcriptome. First, because it represents only a snapshot from the whole genome, second because I do not have access to the population variation. Even so,

a first improvement would be to include orthologs that are more distant that one paralog from the other to clarify whether both members of a pair are under purifying selection or if different selection forces (e.g. relaxed selection) act on one copy, thereby facilitating the emergence of novelties. In addition, a comparative framework between diploids and polyploids of *Biscutella laevigata* would allow assessing whether potential novelties are really attributable to the evolution of gene duplications rather than by standing variation from the diploid progenitors retained in the polyploid as well as clarifying the timing of whole genome duplications.

In Chapter 2, mechanisms underlying reproductive isolation between polyploids in a natural context were assessed. The use of anonymous, dominant markers allowed characterizing the population genetic structure of the tetraploid *Biscutella laevigata* in the Swiss Alps. Field surveys and experimental approaches allowed the identification of pre- and postzygotic reproductive isolation. Furthermore, the intrinsic and extrinsic nature of these barriers was tackled, thereby investigating the role of natural selection shaping the genetic divergence between two polyploid lineages.

Since more than a century a fundamental question in evolutionary biology is what causes speciation and biodiversity (Darwin 1859; Coyne and Orr 2004). Even so tremendous progress was made toward the deeper understanding of processes leading to speciation, the participation of multifarious mechanisms underlying speciation remains to be fully clarified. Several approaches are used to tackle the evolution of reproductive isolation and differ in their vantage point (Via 2009). Studies focusing on incompatibilities after complete speciation provide a rich source of information, but remain unsatisfactory, as it is difficult to disentangle the effective causes leading to

speciation from the processes accumulated after speciation. In contrast, population-level analyses address partially reproductively isolated "ecotypes" in the face of gene flow before complete speciation. This approach is particularly suitable to address the processes leading to reproductive isolation, but can be questioned, as there is no guarantee that taxa will ever reach a species status. However, the population-based method should allow assessing the initial causes of divergence.

This approach was successfully used in Chapter 2. By addressing the mechanisms underlying reproductive isolation in divergent taxa that are still not completely reproductively isolated I could assess pre- and postzygotic barriers underlying this genetic divergence. Even so, it remains unknown if these barriers will lead to complete speciation between both taxa. I could show that genetic divergence is largely supported by prezygotic barriers confirming plenty studies on the reproductive isolation between diploids (Lowry et al. 2008). Even so, the specificity of habitats assessed in established individuals of the different genetic groups in the field remains puzzling, because local adaptation could not be confirmed experimentally. These results represent important first insights in the study of autopolyploids, as very little is known regarding their evolution and diversification (Parisod et al. 2010). Several hypotheses could explain the discrepancy between field survey and the experimental approach. First, a low response to selection predicted in polyploids or a subordinate role for local adaptation. Also, the high mortality in the first year dramatically decreased the sample sizes in this experiment. Low sample sizes in addition to symmetric mortality lowered the chances to assess statistically relevant differences in survival or fitness with this experiment. *Biscutella laevigata* plants do not produce flowers every year, important differences in flowering time in a common environment were observed and flowers produce only two

seeds per flower, thereby hindering the production of a large amount of crosses. Therefore, increasing the number of F1 would be difficult but necessary to obtain meaningful sample sizes after a high mortality of seedlings. But this could then allow to follow a larger number of individuals over a longer period and assess the timing of a putative local adaptation.

The use of ultra precise environmental data represents a major advance in the field. The first digital model used in this study had a resolution of 2m (swisstopo). It is a very complex task to model the complex topographic situation of a ridge, and the model at 2m was inaccurate for a large number of individuals. Many individuals close to the ridge (as observed in the field), were not on the ridge according to the model but localized in the cliff, thereby biasing the calculation of DEM-derived variables. The new DEM at 50 cm accuracy provided for this study allowed to overcome this issue. By comparison with field observation, this model correctly allocated individuals to the ridge. Even so, all variables used in this study are proxies for the effective factor driving local adaptation the combination of DEM-derived variables with direct measurements and observation allowed a fine description of the habitat of both genetic groups.

A promising environmental factor highlighted in this PhD is water stress or drought. Soil humidity was an important factor differentiating both genetic groups. In addition functional categories linked to water stress were preferentially retained after the B-WGD. Tackling what genes were retained or even diversified, potentially representing a key adaptation represents a promising framework to better understand a putative radiation in the Alps. Toward this goal a better characterization of soil moisture at the individual level and for more individuals will be an important step in the future. But, to really assess the implication of drought stress experimental approaches in

common or controlled environment would be crucial. It could be possible to follow the fitness of the different genetic groups from Chapter 2 in a controlled experiment mimicking drought (e.g. in growth chambers). Assessing the direct factors influencing the survival of individuals studied in Chapter 2 would also allow controlling for habitat effect. Thus, showing whether both groups are really locally adapted, or if survival or fitness depends of habitat quality.

Finally, genetic studies in polyploids yield several issues, particularly because of the difficulty to determine allele dosage, copy number variation and the difficulty of distinguishing orthologs from paralogs. AFLP (amplified fragment length polymorphism) fingerprinting is commonly used in natural populations to provide a large number of dominant and anonymous markers assumed to be scattered across the whole genome (Meudt and Clarke 2007). Even though AFLPs do not allow gaining access to allele dosage, the here-used multivariate approaches were particularly useful at overcoming this issue. But, even genotyping with novel high throughput technologies (e.g. genotyping by sequencing) remains challenging in polyploids. Therefore, the identification of gene families and paralogs, at least partially in transcripts, should allow overcoming some of these issues allowing, e.g., tackling questions relative to the mode of inheritance or targeting loci underlying reproductive isolation with modern sequencing technologies.

Dissecting the genetic mechanisms underlying the genetic divergence, i.e. progression from few loci building up reproductive isolation toward a genome-wide barrier, is central in the theory of speciation. In the future, the genetic mechanisms underlying reproductive isolation, pre- and postzygotic, in this population should be addressed to fully understand the origin of genetic divergence. A first promising step

probably being the identification of loci underlying flowering time and their potential linkage to other genes possibly linked to environmental stress.

Literature Cited

Comai, L. 2005. The advantages and disadvantages of being polyploid. Nature Review Genetics 6:836-846.

Coyne, J. A. and H. A. Orr. 2004. Speciation. Sinauer, Sunderland, MA.

Darwin, C. 1859. On the Origin of Species by Means of Natural Selection or the Preservation of Favored Races in the Struggle for Life, J. Murray, London.

Glennon, K. L., M. E. Ritchie, and K. A. Segraves. 2014. Evidence for shared broad-scale climatic niches of diploid and polyploid plants. Ecology Letters 17:574-582.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, Y. Hu, H. Liang, P. S. Soltis, D. E. Soltis, S. W. Clifton, S. E. Schlarbaum, S. C. Schuster, H. Ma, J. Leebens-Mack, and C. W. dePamphilis. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97-U113.

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Selection in the evolution of gene duplications. Genome Biology 3.

Lowry, D. B., J. L. Modliszewski, K. M. Wright, C. A. Wu, and J. H. Willis. 2008. The strength and genetic basis of reproductive isolating barriers in flowering plants. Philos T R Soc B 363:3009-3021.

Meudt, H. M. and A. C. Clarke. 2007. Almost forgotten or latest practice? AFLP applications, analyses and advances. Trends in Plant Science 12:106-117.

Otto, S. P. and J. Whitton. 2000. Polyploid incidence and evolution. Annu Rev Genet 34:401-437.

Parisod, C., R. Holderegger, and C. Brochmann. 2010. Evolutionary consequences of autopolyploidy. New Phytologist 186:5-17.

Schranz, M. E., S. Mohammadin, and P. P. Edger. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Current Opinion in Plant Biology 15:147-153.

Via, S. 2009. Natural selection in action during speciation. Proceedings of the National Academy of Sciences of the United States of America 106:9939-9946.

Curriculum vitae

CÉLINE GEISER

Date of birth: 28.04.1984

Nationality: Swiss

Languages: French, German, English

ACADEMIC EDUCATION

2011 –2014	**University of Neuchâtel: Ph.D. in Evolutionary Botany** • Thesis: "Genome evolution and mechanisms underlying reproductive isolation in the polyploid *Biscutella laevigata*"
2008 –2011	**University of Geneva & University of Basel: Master in Environmental Sciences** • Master thesis: "Effect of habitat fragmentation on the genetic structure of slow-worm populations."
2010 –2011	**University of Geneva: Certificate in Geomatics** • Specialisation in spatial ecology
2005 – 2008	**University of Zurich: Bachelor of Science in Biology with Minor in Environmental Sciences**

SCHOLAR FORMATION

1999 - 2002	**Grammar school in the "Lycée cantonal de Porrentruy"**
1990 - 1999	**Primary and secondary school in Delémont**

PUBLICATIONS

- Geiser C, Ray N, Lehmann A, Ursenbacher S. (2013). „Unravelling landscape variables with multiple approaches to overcome scarce species knowledge: a landscape genetic study of the slow worm." Conservation Genetics 14:783–794.

- Metzger C, Ferchaud A-L., Geiser C., Ursenbacher S. (2011). "New polymorphic microsatellite markers of the endangered meadow viper (Vipera ursinii) identified by 454 high-throughput sequencing: when innovation meets conservation. Conservation Genetics Resources 3: 589-592

- Schmidt, B. R., C. Geiser, et al. (2009). "Assessing whether disinfectants against the fungus Batrachochytrium dendrobatidis have negative effects on tadpoles and zooplankton." Amphibia-Reptilia 30(3): 313-319.