Genetic diversity and climate adaption in *Arabidopsis lyrata*

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Marco Fracassetti

aus Bergamo, Italia

Basel, 2016

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Yvonne Willi and Dr. Luca Ferretti

Basel, den 18 Oktober 2016

_____

Prof. Dr. Jörg Schibler

TABLE OF CONTENTS

SUMMARY

Applied fields of research such as the one on global climate change has heightened the interest to understand the adaptive evolution process and limits to adaptive evolution. Progress in the field depends on knowing of the traits under selection and their genetic variation. The goal of my PhD thesis was to generally assess genome-wide single nucleotide polymorphism (SNP) diversity across an entire species geographic distribution and to detect SNPs and genes linked to adaptation to climatic variables and substrate type within the herbaceous plant *Arabidopsis lyrata* subsp. *lyrata* (*A. lyrata*). For this work, DNA of 52 populations covering the whole geographic range of *A. lyrata* were analyzed by pooling DNA of multiple individuals of each population, sequencing the pools (Pool-seq) and revealing population SNP frequencies. In the first chapter the wet-lab protocol of Pool-seq and the bioinformatics pipeline were tested. In the second chapter the genetic diversity of different genomic regions was analyzed to trace the history of the populations of *A. lyrata*. In the third chapter, the climatic variables that determine the ecological niche limits of the species distribution were defined. And, in the fourth chapter the SNP frequencies were associated with climatic variables and substrate type to detect the genomic regions involved in adaptation to climate and edaphic conditions, highlighting potentially relevant genes and pathways.

ACKNOWLEDGEMENT

Doing Ph.D in Switzerland was a good experience that allowed me to increase my knowledge in biology and bioinformatics.

I am deeply grateful to my supervisor Yvonne Willi for guidance during the years, help and availability whenever needed.

A thank you to Luca Ferretti who accepted to be co-referee for my Ph.D and for the helpful discussion.

I want to thank the colleagues for the beautiful moments spent together. Those that I met in Neuchatel: Alessio Maccagni, Alfonso Rojas, Antoine Paccard, Benjamin Dauphin, Céline Geiser, Georgi Bonchev, Guillaume Wos, Julie Lee-Yaw, Julien Vieu, Magali Meniri, Olivier Bachmann, Philippa Griffin, Rimjhim Roy Choudhury and Stella Huynh. Those that I met in Basel: Franziska Grob, Georg Armbruster, Jürg Oetiker, Kay Lucek, Lukas Schütz, Markus Funk, Maura Ellenberger, Michael Thieme, Raphael Weber, Silvia Calabrese, Silvia Turco, Tim Hander and Victor Golyaev.

Thanks to the University of Neuchâtel and University of Basel for financial support to conduct this Ph.D.

Thanks to my family who, despite being far away, has given me support during these years.

Finally, my greatest thank to my beloved Veronica, who was the best discovery during these years.

INTRODUCTION

Comparative population genomics can help us getting a better understanding about the genes of adaptation (Hoffmann & Willi 2008). The idea of population comparisons is that differences in the environmental conditions among populations have caused traits and genes underlying them to adaptively diverge. When sequence variation can be linked with a particular environmental state such as a climatic variable across populations, it should be possible to detect the genomic regions under divergent evolution. If sample sizes for the study system are large enough, genes likely involved in e.g., climate adaptation should be detected. This is the principle idea of an association study, to link SNP variation with a trait, which can be the climate where a species can be found. Imperative to such work is to account for differences in relatedness among populations as there may be correlations between relatedness among populations and exposure to climate, which would increase the rate of false positive detections. In my PhD thesis I investigated these topics by analyzing the SNP (Single Nucleotide Polymorphism) frequencies in *Arabidopsis lyrata* using the technique of Pool-seq (Schlötterer *et al.* 2014).

The model organism that I studied is *Arabidopsis lyrata*, which is a member of the family of the *Brassicaceae* and is closely related to the plant model species *Arabidopsis thaliana*. *Arabidopsis lyrata* and *Arabidopsis thaliana* are morphologically, physiologically and genetically similar species. Therefore the experimental designs, molecular tools and software developed for *A. thaliana* can be easily adapted for the study of *A. lyrata*. The size of the *A. lyrata* genome is 206.7 Mbp on 8 chromosomes, while the size of the *A. thaliana* genome is smaller, 125 Mbp on 5 chromosomes (Hu *et al.* 2011). Furthermore, in contrast to *A. thaliana*, *A. lyrata* is a short-lived perennial species, not annual, and it is predominantly outcrossing, not selfing. In fact, the North American subspecies is considered a new model organism to study mating system evolution, because several independent shifts to selfing have happened at the edges of the geographic species

distribution (Willi & Määttänen 2010; Haudry *et al.* 2012; Willi 2013; Griffin & Willi 2014). Other types of studies conducted on *A. lyrata* focused on further adaptive differences: the adaptation to different types of soils (Turner *et al.* 2008, 2010), the interactions with herbivore species (Clauss & Mitchell-Olds 2006; Abel *et al.* 2009; Puentes & Ågren 2012) and differences in flowering time (Sandring *et al.* 2007; Leinonen *et al.* 2013).

Pooling biological samples has been widely used in population genetics analysis for the estimation of SNP frequencies (reviewed in (Sham *et al.* 2002)). The approach consists of pooling the DNA of many samples, in equimolar amounts, and to sequence them together. The quantification of DNA of individual samples is a critical step of this technique. Therefore an accurate quantification based on fluorimetry is strongly recommended. Via the advent of next-generation sequencing (NGS) (Margulies *et al.* 2005; Pandey *et al.* 2008; Bentley *et al.* 2008) techniques of population genetic studies were revolutionized. It has become easier, cheaper and faster to obtain data at a genome-wide level for multiple populations. Whole-genome sequencing of pooled DNA is more recent and known as Pool-seq (Schlötterer *et al.* 2014). Pooling is a cost-effective method because it permits to reduce sequencing cost without reducing the sample size. Hence, Pool-seq has been applied in several field in population genomics: the demographic history (Corander *et al.* 2013), the identification of genomic loci affecting a trait of interest (Bastide *et al.* 2013), the detection of the signature of selection (Kofler *et al.* 2012; Fabian *et al.* 2012) and in genome-wide association studies (GWAS) (Turner *et al.* 2010; Fischer *et al.* 2013).

While the method of Pool-seq has become popular in the last few years, it has also been questioned, particularly in regard to the accuracy of SNP frequency data it produces (Cutler & Jensen 2010; Anderson *et al.* 2014). To address this criticism, I investigated the robustness of Pool-seq in estimating SNP frequencies depending on sample size, sequencing depth and the SNP caller used (**chapter 1**). Particularly, I validated Pool-seq for population genomics by comparing SNP frequencies revealed by pooling and re-sequencing with those revealed by individual-based

Genotyping-By-Sequencing (GBS) (Elshire *et al.* 2011). I analyzed how the pool size and the sequencing depth affect the accuracy of Pool-seq SNP frequency estimates. Furthermore I compared the accuracy of two SNP calling program: VarScan (Koboldt *et al.* 2012) and Snape (Raineri *et al.* 2012).

Through the Pool-seq technique I analyzed 52 populations of *A. lyrata* (25 individuals for each population) across its entire geographic range in North America, which extends from North Carolina and Missouri to upstate New York and Ontario (Schmickl *et al.* 2010; Paccard *et al.* 2016). Particularly, I quantified the relative importance of historic range dynamics compared to current local demographic parameters in explaining genetic diversity in different regions of the genome (**chapter 2**). Classic equilibrium-based theoretical models predict a positive effect of population size, mutation rate, gene flow and outcrossing on genetic diversity and mixed effects of selection depending on their type and strenght (Willi *et al.* 2006). Furthermore, within-population genetic diversity may bear also an important signature of historic demographic processes (Wright & Gaut 2005; Duncan *et al.* 2015). I first reconstructed the phylogeographic history of *A. lyrata* based on nuclear single nucleotide polymorphism (SNP) frequencies and identified possible refugia during the LGM (Last Glacial Maximum). I then compared genomic diversity estimates based on genome-wide SNP frequencies and published microsatellite-based genetic diversity estimates (Griffin & Willi 2014). Lastly, I tested how the phylogeographic history, admixture events, local census size and the mating system affected genome-wide genetic diversity for intergenic regions, introns and coding regions (CDS).

I also investigated how the genetic diversity in *A. lyrata* varied with respect to the distribution of the species (**chapter 3**). Towards the edge of a species distribution, genetic diversity is predicted to decrease (Sagarin & Gaines 2002; Eckert *et al.* 2008; Sexton *et al.* 2009). Furthermore the range limits of the species distribution could coincide with the niche limits (Hargreaves *et al.* 2014; Lee-Yaw *et al.* 2016), where the species experiences increasingly marginal

conditions towards the edge of the range. First, my collaborators and I identified the environmental variables that determine the niche limits of *A. lyrata*. Then we tested if geographic range limits reflect ecological niche limits, and if there is a relationship between environmental suitability and genome-wide patterns of genetic diversity.

Finally, I performed an environmental association analysis (EAA) on the SNPs frequencies of 42 outcrossing populations of *A. lyrata* (**chapter 4**). I tested the association with the environmental variables that determine the niche limits of the species and the substrate type on which the population are located in nature (sandy and rocky sites). I carried out a gene ontology analysis on the associated SNPs and I suggested the top candidate genes linked to these environmental variables.

REFERENCES

Abel C, Clauss M, Schaub A, Gershenzon J, Tholl D (2009) Floral and insect-induced volatile formation in *Arabidopsis lyrata* ssp. *petraea*, a perennial, outcrossing relative of *A. thaliana*. *Planta*, **230**, 1–11.

Anderson EC, Skaug HJ, Barshis DJ (2014) Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology*, **23**, 502–512.

Bastide H, Betancourt AJ, Nolte V *et al.* (2013) A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLoS Genetics*, **9**, e1003534.

Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Clauss MJ, Mitchell-Olds T (2006) Population genetic structure of *Arabidopsis lyrata* in Europe. *Molecular Ecology*, **15**, 2753–66.

Corander J, Majander KK, Cheng L, Merilä J (2013) High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. *Molecular Ecology*, **22**, 2931–2940.

Cutler DJ, Jensen JD (2010) To pool, or not to pool? *Genetics*, **186**, 41–43.

Duncan SI, Crespi EJ, Mattheus NM, Rissler LJ (2015) History matters more when explaining genetic diversity within the context of the core-periphery hypothesis. *Molecular Ecology*, **24**, 4323–4336.

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. *Molecular Ecology*, **17**, 1170–1188.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Fabian DK, Kapun M, Nolte V *et al.* (2012) Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology*, **21**, 4748–4769.

Fischer MC, Rellstab C, Tedder A *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Molecular Ecology*, **22**, 5594–5607.

Griffin PC, Willi Y (2014) Evolutionary shifts to self-fertilisation restricted to geographic range margins in North American *Arabidopsis lyrata*. *Ecology Letters*.

Hargreaves AL, Samis KE, Eckert CG (2014) Are species' range limits simply niche limits writ large? A review of transplant experiments beyond the range. *The American Naturalist*, **183**, 157–173.

Haudry A, Zha HG, Stift M, Mable BK (2012) Disentangling the effects of breakdown of self-incompatibility and transition to selfing in North American *Arabidopsis lyrata*. *Molecular Ecology*, **21**, 1130–42.

Hoffmann AA, Willi Y (2008) Detecting genetic responses to environmental change. *Nature Reviews Genetics*, **9**, 421–32.

Hu TT, Pattyn P, Bakker EG *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, **43**, 476–481.

Koboldt DC, Zhang Q, Larson DE *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**, 568–576.

Kofler R, Betancourt AJ, Schlötterer C (2012) Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1002487.

Lee-Yaw JA, Kharouba HM, Bontrager M *et al.* (2016) A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits (JM Gomez, Ed,). *Ecology Letters*, **19**, 710–722.

Leinonen PH, Remington DL, Leppälä J, Savolainen O (2013) Genetic basis of local adaptation and flowering time variation in *Arabidopsis lyrata*. *Molecular Ecology*, **22**, 709–23.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Paccard A, Van Buskirk J, Willi Y (2016) Quantitative genetic architecture at latitudinal range boundaries: reduced variation but higher trait independence (CG Eckert, JL Bronstein, Eds,). *The American Naturalist*, **187**, 667–677.

Pandey V, Nutter RC, Prediger E (2008) Applied Biosystems SOLiD™ System: ligation-based sequencing. In: *Next Generation Genome Sequencing: Towards Personalized Medicine* (ed Janitz M), pp. 29–42.

Puentes A, Ågren J (2012) Additive and non-additive effects of simulated leaf and inflorescence damage on survival, growth and reproduction of the perennial herb *Arabidopsis lyrata*. *Oecologia*, **169**, 1033–42.

Raineri E, Ferretti L, Esteve-Codina A *et al.* (2012) SNP calling by sequencing pooled samples. *BMC Bioinformatics*, **13**, 239.

Sagarin RD, Gaines SD (2002) The “abundant centre” distribution: to what extent is it a biogeographical rule? *Ecology Letters*, **5**, 137–147.

Sandring S, Riihimäki MA, Savolainen O, Agren J (2007) Selection on flowering time and floral display in an alpine and a lowland population of *Arabidopsis lyrata*. *Journal of Evolutionary Biology*, **20**, 558–567.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**, 749–763.

Schmickl R, Jørgensen MH, Brysting AK, Koch MA (2010) The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evolutionary Biology*, **10**, 1–18.

Sexton JP, McIntyre PJ, Angert AL, Rice KJ (2009) Evolution and ecology of species range limits. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 415–436.

Sham P, Bader JS, Craig I, O’Donovan M, Owen M (2002) DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, **3**, 862–871.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin S V (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics*, **42**, 260–263.

Turner TL, Wettberg EJ Von, Nuzhdin S V (2008) Genomic analysis of differentiation between soil types reveals candidate genes for local adaptation in *Arabidopsis lyrata*. *PLoS ONE*, **3**, e3183.

Willi Y (2013) Mutational meltdown in selfing *Arabidopsis lyrata*. *Evolution*, **67**, 806–15.

Willi Y, Buskirk J Van, Hoffmann AA (2006) Limits to the adaptive potential of small populations. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 433–458.

Willi Y, Määttänen K (2010) Evolutionary dynamics of mating system shifts in *Arabidopsis lyrata*. *Journal of Evolutionary Biology*, **23**, 2123–31.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular biology and evolution*, **22**, 506–19.

CHAPTER 1: Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata*

Marco Fracassetti[1]*, Philippa C. Griffin[1,2], and Yvonne Willi[1]

[1]Institute of Biology, Evolutionary Botany, University of Neuchâtel, 2000 Neuchâtel, Switzerland

[2]School of BioSciences, University of Melbourne, 3010 Parkville, Victoria, Australia

* Corresponding author:

E-mail: marco.fracassetti@unine.ch

PUBLISHED IN PLOSONE (DOI 10.1371/journal.pone.0140462)

Abstract

Sequencing pooled DNA of multiple individuals from a population instead of sequencing individuals separately has become popular due to its cost-effectiveness and simple wet-lab protocol, although some criticism of this approach remains. Here we validated a protocol for pooled whole-genome re-sequencing (Pool-seq) of *Arabidopsis lyrata* libraries prepared with low amounts of DNA (1.6 ng per individual). The validation was based on comparing single nucleotide polymorphism (SNP) frequencies obtained by pooling with those obtained by individual-based Genotyping By Sequencing (GBS). Furthermore, we investigated the effect of sample number, sequencing depth per individual and variant caller on population SNP frequency estimates. For Pool-seq data, we compared frequency estimates from two SNP callers, VarScan and Snape; the former employs a frequentist SNP calling approach while the latter uses a Bayesian approach. Results revealed concordance correlation coefficients well above 0.8, confirming that Pool-seq is a valid method for acquiring population-level SNP frequency data. Higher accuracy was achieved by pooling more samples (25 compared to 14) and working with higher sequencing depth (4.1× per individual compared to 1.4× per individual), which increased the concordance correlation coefficient to 0.955. The Bayesian-based SNP caller produced somewhat higher concordance correlation coefficients, particularly at low sequencing depth. We recommend pooling at least 25 individuals combined with sequencing at a depth of 100× to produce satisfactory frequency estimates for common SNPs (minor allele frequency above 0.05).

Keywords: Brassicaceae; concordance correlation coefficient; coverage; depth of sequencing; genomic library; pooled sequencing; population genomics; SNP detection.

Introduction

The method of pooling biological samples for downstream analysis has been used for more than seventy years [1]. The main advantage of pooling is that more samples can be analyzed in a cost-effective way. Pooling has been widely used in population genetics analysis for the estimation of single-nucleotide polymorphism (SNP) frequencies (reviewed in Sham et al. [2]). More recently, the field of population genetics has been revolutionized by the development of next-generation sequencing (NGS), as it is now possible to study genetic variation at the whole-genome level [3–7]. Whole-genome sequencing of pooled DNA is more recent and known as Pool-seq [8]. While this method has become popular in the last few years, it has also been questioned, particularly in regard to the accuracy of SNP frequency data it produces [9,10]. To address this criticism, we investigated the robustness of Pool-seq in estimating SNP frequencies depending on sample size, sequencing depth and the SNP caller used.

So far, Pool-seq has been used in the study of bacteria [11], yeast [12], flatworm [13], sea urchins [14], plants [15,16], *Drosophila* [17–19], fish [20], birds [21] and mammals [22–25]. The approach has been applied to identify genomic loci affecting a trait of interest [19], to infer the demographic history of populations [20], to detect the signature of selection [17,18,25] and to perform genome-wide association studies (GWAS) [15,16]. In many cases, the pooling of samples is used to reduce costs. But pooling can be obligatory in other cases, such as when separating individuals is problematic [14,26] or when there is insufficient DNA to make individual libraries.

Several weaknesses of the method have been discussed. Low individual numbers, rough DNA quantification, and low sequencing depth can add error to polymorphism frequency estimates [27,28]. While these problems can be resolved and/or the magnitude of impact estimated, there are two more systemic, less easily resolvable limitations. When DNA of individual samples is pooled, information on haplotypes is lost. It is no longer possible to link a polymorphism with the individual to which it belongs [8], which is a problem for studies that require information on

linkage disequilibrium, for example. The other limitation is that sequencing errors cannot easily be distinguished from true rare alleles [9]. Several authors have developed statistical approaches to tackle these two issues [29–33], which have been implemented in software programs to analyse pooled data [34–37]. In line with the intention of such improvements, the goal must be to assess the impact of problems of Pool-seq and to come up with procedures to resolve them, especially as whole-genome re-sequencing of individuals for population genomics is still expensive for species with medium-sized to large genomes.

This study focused on validating Pool-seq for population genomics by comparing SNP frequencies revealed by pooling and re-sequencing with those revealed by individual-based Genotyping By Sequencing (GBS) [38]. Comparisons were based on field-sampled plants of *Arabidopsis lyrata*. Library preparation required very little DNA and was performed with standard laboratory equipment. The three main questions we addressed were: (1) What is the increase in accuracy of Pool-seq SNP frequency estimates when increasing pool size? (2) What is the sequencing depth per individual required to obtain reliable population SNP frequencies with Pool-seq? And, (3) what is the difference in accuracy of SNP calling between a heuristic approach as implemented in the software VarScan [39] and a Bayesian approach as implemented in Snape [35]?

Materials and Methods

The *A. lyrata* plants of population A were collected in Presque Isle State Park (Erie, PA, USA) with a permit granted by the Commonwealth of Pennsylvania. The *A. lyrata* plants of population B were collected in the Clark Reservation State Park (Jamesville, NY, USA) with a permit granted by the New York State Office of Parks, Recreation and Historic Preservation. DNA of field-collected plants was extracted from silica-dried leaves with the DNeasy 96 Plant Kit (Qiagen, Hombrechtikon, Switzerland). Each DNA sample was quantified twice with the DNA quantification kit Quant-IT^{TM} DNA HS (Invitrogen, Paisley, UK), a method based on fluorimetry with a DNA-

specific dye. Samples were only accepted as suitable for the study if the average concentration was at least 0.25 ng/ml and the coefficient of variation between the two rounds of quantification was smaller than 0.1. We sampled 14 individuals from population A and 25 from population B (Fig. 1). The same individuals of these two populations were analysed by pooled (Pool-seq) and individual (GBS) sequencing.

Library preparation: Pool-seq

Libraries for Pool-seq were prepared with the Nextera Kit (Illumina, San Diego, CA, USA) from equimolar-pooled DNA samples for each population. For each library a total of 40 ng of DNA was used, 2.8 ng per individual for population A and 1.6 ng per individual for population B. The protocol was customized to work with strips of 8 PCR tubes. The tagmentation time was increased from the manufacturer's protocol of 5 min to 10 min. The number of PCR cycles was increased to 8 (instead of 5) and the elongation time was decreased to 2 min (instead of 3 min). Library A was paired-end sequenced for 100 bases (PE100) on half a lane of Illumina HiSeq2000. Library B was PE100 sequenced on four lanes, each time constituting one quarter of the lane. Data of the lanes of population B were merged to create combinations from one to four lanes together (lane 1, lanes 1+2, lanes 1+2+3, lanes 1+2+3+4; Fig 1.).

Library preparation: GBS

Genomic DNA (50 ng per individual) was digested at 37°C for 65 min in a 20 µL reaction volume with 5 U *Msp*I (NEB, Ipswich, MA, USA) in 10× NEBuffer 4. Following heat inactivation of the restriction enzyme (65°C, 20 min), tubes were allowed to cool slowly to room temperature covered with tinfoil. Adapter ligation was then performed immediately, using the following reaction mixture: 5 µL 10× NEBuffer 2, 1.93 µL P1 adapter (10 µM; sequence as per Elshire et al. [39] but with a CG instead of a CWG sticky end, and containing a 4-9 base barcode sequence), 1.93 µL P2

adapter (10 µM; sequence as per Elshire et al. [38] but with a CG instead of CWG sticky end), 1.8 µL rATP (100 mM), 1.5 µL T4 DNA ligase (2×10^6 U/mL), made up to 50 µL with ddH_2O. Ligation reactions were incubated at room temperature for 45 min, then heat-inactivated at 65°C for 20 min. Tubes were allowed to cool slowly as before.

To multiplex barcoded samples, 5 µL of each ligation mix was pooled. The mixture was cleaned with a Clean and Concentrator -5 Kit (Zymo Research, Irvine, CA, USA), eluted in 50 µL Buffer EB. The pooled and cleaned DNA was used as template in 25 parallel PCR amplifications (replicated to minimise template bias). Each well included 2 µL template DNA, 2.5 µL of each PCR primer (as per Elshire et al. [38]), 5 µL dNTPs (2 mM), 0.5 µL Taq polymerase (Promega, Madison, WI, USA), 5x GoTaq buffer (Promega) and ddH_2O to a final volume of 50 µL. Cycling protocol was as follows: 72°C for 5 min, 96°C for 30 s, 18 cycles of [96°C for 30 s, 65°C for 30 s, 72°C for 30 s], and a final extension of 72°C for 5 min. All replicate PCR reactions were pooled, and cleaned a second time as before, eluting in 30 µL of buffer per ~200 µL of PCR product. Size selection was performed with the Caliper LapChip XT (PerkinElmer, Waltham, MA, USA), set to collect two peaks (first peak: 350 bp, second peak: 455 bp), which effectively collected fragments between 301-519 bp due to the machine's size accuracy limit of 14%. A third cleanup was performed, eluting in 17 µL Buffer EB. Sequencing was performed in a single Illumina HiSeq2000 lane.

Bioinformatics pipelines and SNP frequency comparison

The bioinformatics pipelines for Pool-seq and GBS sequence data were kept as similar as possible to minimize differences due to software used (pipelines accessible at: http://github.com/fraca). The sequences are stored at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) with the accession number PRJEB8335. Demultiplexing of the GBS data was performed with the preprocess_radtags script of Stacks [40], which retains reads with the proper barcode and restriction cut sites.

The Pool-seq and GBS sequences were trimmed using the script trim-fastq.pl of the software program PoPoolation [34] with a base quality threshold of 20, trimmed only from the 3' end to allow the subsequent removal of duplicates. Reads were mapped with BWA-MEM using default parameters [41]. The first 8 scaffolds of the published genome of *A. lyrata* v1.0 [42] were used as the reference genome. Data of the Pool-seq lanes of population B were merged to create the different combinations. Duplicate reads were removed with the MarkDuplicates tool of Picard [43]. Only proper paired reads with a mapping quality score above 20 were retained to create a pileup file with SAMtools [44]. The pileup file of Pool-seq data was filtered to retain regions with depth of coverage per site of 14-500 for population A and 25-500 for population B. The pileup file of GBS data was filtered for regions with depth of coverage per site of 5-500 for an individual and for data available for at least 90% of the individuals of a population. The regions near insertions and deletions were identified (identify-genomic-indel-regions.pl) and removed (filter-pileup-by-gtf.pl) with PoPoolation [34]. The genomic interspersed repeats were identified in the reference genome with RepeatMasker [45] using the default settings for "arabidopsis" and removed from the pileup files.

Finally, the filtered pileup files were used to call SNPs with the program VarScan with a significance (P) value ≤ 0.05, minimum base quality of 20 and a minimum allele count of two reads. For the Pool-seq data, SNPs were additionally called with Snape [35]. We retained SNPs with a posterior probability of segregation > 0.9 and minimum allele count of two reads. The nucleotide diversity and the genetic differentiation from the reference genome that are needed to set prior probabilities in the Bayesian model of Snape were calculated by NPStat [37]. We used the BEDTools software [46] to calculate sequencing depth or depth of coverage per site, defined as the number of times each base was sequenced per individual or per population pool. We applied the same thresholds for SNP calling and genome coverage calculation. Figure 1 presents the final 12 data sets used for further analysis. Allele frequency estimates were calculated as the fraction of

reads carrying the non-reference allele for Pool-seq data, and the fraction of the non-reference allele across GBS-derived genotypes.

Three statistics were used to compare Pool-seq-based SNP frequencies with those obtained by GBS. First, the concordance correlation coefficient (CCC) was calculated using the epiR package [47]. This test statistic can be used to evaluate the agreement between two variables [48]. The CCC combines precision (deviation from best-fit-line) and accuracy (deviation of best-fit-line from 45° line through origin) to determine how far the observed data deviate from the line of perfect concordance. Second, the absolute value of the difference between the estimated SNP frequencies with the two methods ($|\Delta f|$) was calculated and its distribution investigated. Third, a false negative rate was calculated as the fraction of SNPs called in GBS but not in the pooled sample, relative to the total number of SNPs called by GBS. This calculation included only genomic regions covered by both GBS and Pool-seq data, and considered SNP frequencies estimated from GBS to represent the true population frequencies. Because sequencing depth of GBS reads did not meet the minimum threshold of five reads for all the individuals, data did not allow the reliable estimation of a false positive rate of SNP calling.

Results

Sequencing statistics

Pooled sequencing of population A yielded 34 million paired-end reads. Prior to restricting the reads falling within an informative range of coverage depth, 50% of reads mapped unambiguously to 74% of the *A. lyrata* nuclear genome, at a mean depth of 27×. After applying the read depth cutoff (min 14×, max 500×) and removing duplicates, 46% of the reads mapped to 41% of the *A. lyrata* nuclear genome. The mean sequencing depth of population A was 37× in the final data set, which is equivalent to a mean depth of 2.6× per individual. Pooled sequencing of population B was performed on four lanes, each of which yielded ~40 million paired-end reads. We unambiguously

mapped 59% of the total reads to cover 80% of the *A. lyrata* nuclear genome, at a mean depth of 25× per lane. After applying the read depth cutoff (min 25×, max 500×) and removing duplicates, the percentage of the genome covered by one lane was on average 36%, while the four lanes together covered 70%. The mean sequencing depth (post-cutoff) of population B depended on the number of lanes merged; depth was on average 36× for one lane and 103× for four lanes (Fig 1). Accordingly, sequencing depth per individual varied between 1.4× and 4.1×. Individual sequencing by GBS yielded 105 million paired-end reads (population A and B together) that were correctly barcoded and trimmed. We unambiguously mapped 40% of reads to cover 2% of the *A. lyrata* nuclear genome. Once the read depth cutoff (min 5×, max 500×) was applied, the mean sequencing depth per individual for population A in the final data set was 17× (range across individuals: 10×-30×). For population B the mean sequencing depth per individual was 30× (range across individuals: 11×-113×).

Number of SNPs

Table 1 shows the number of SNPs called by GBS and Pool-seq. For the Pool-seq protocol and population A, the software VarScan called 0.50 million SNPs, while Snape called 0.72 million SNPs. Increasing the depth from 1.4× to 4.1× (from one to four lanes) for population B increased the number of SNPs called. Using VarScan, the SNPs called increased from 0.68 million to 1.95 million. Using Snape, the SNPs called increased from 1.04 million to 2.54 million. Figure 2 shows the fraction of SNPs called with both VarScan and Snape in population B using one or four lanes. Almost all the SNPs called by VarScan were also called by Snape. The percentage of SNPs called by both programs relative to the total number of SNPs called by either Snape or VarScan, increased from 65% to 76% when the input data were increased from one lane to four lanes.

GBS led to more SNPs for population A than for population B. The smaller sample of individuals in population A (14 instead of 25 in population B) made it easier to attain the processing

threshold of five or more reads for at least 90% of individuals. Therefore, population A had higher overlap among individuals in genomic regions with sufficient sequencing depth and a higher total number of called SNPs. Moreover, the number of SNPs identified by both GBS and Pool-Seq was low (column SNP_{both}, Table 1) because GBS revealed SNP information for a small fraction of the genome and that fraction overlapped incompletely with genomic regions also covered with acceptable depth by Pool-seq.

Comparison of SNP frequencies revealed by Pool-seq versus GBS

First, SNP frequencies obtained with Pool-seq and GBS were compared by the use of the concordance correlation coefficient (CCC), which captures the agreement between two variables by accounting for precision and accuracy and which can range from 0 to 1. Figure 3 illustrates CCC values with upper and lower 95% confidence ranges for all library/lane combinations studied. CCC values for population A were 0.827 for SNPs called with VarScan and 0.864 for those called with Snape (Table 1). For population B, CCC values increased with increasing depth of coverage per site from 0.887 (1.4×) to 0.952 (4.1×) with VarScan and from 0.911 (1.4×) to 0.955 (4.1×) with Snape. Figure S1 illustrates the correlation between SNP frequency estimates of Pool-seq and those of GBS. The correlation between the two increased when more samples were pooled, and when the depth of coverage per site was increased.

Second, SNP frequencies revealed with Pool-seq and GBS were compared based on the absolute difference between the SNP frequency estimates of the two methods ($|\Delta f|$ in Table 1). The mean $|\Delta f|$ for population A was 0.109 with VarScan and 0.103 with Snape. The mean $|\Delta f|$ for population B decreased with increasing sequencing depth, from 0.092 to 0.058 with VarScan and from 0.083 to 0.055 with Snape. Figure 4 shows the distribution of $|\Delta f|$ for each library/lane combination, and Fig S2 presents the distribution of the difference between the SNP frequency estimates of the two methods across the achieved read depth at SNP sites for each library/lane

combination. The difference in SNP frequencies between methods was generally lower when read depth was high, both across and within library/lane combinations. Furthermore, the distribution of the difference was not appreciably biased towards either negative or positive values (Fig S2).

Third, the false negative rate (FN rate in Table 1) decreased with increasing sequencing depth, from 0.385 (1.4×) to 0.170 (4.1×) with VarScan and from 0.212 (1.4×) to 0.101 (4.1×) with Snape. At the same time, the mean frequency of minor alleles at GBS SNPs that were missed by Pool-seq (FN MAF in Table 1) decreased from 0.077 to 0.045 with VarScan and from 0.054 to 0.036 with Snape. Figure 5 illustrates that the minor allele frequency at SNP sites missed by Pool-seq was mostly lower than 5% when the number of sequenced individuals and the sequencing depth per individual were both high.

Discussion

Pooled whole-genome re-sequencing (Pool-seq) has only recently been adopted for population genomics in eukaryotes, so validation studies are needed, together with test of aspects of the wet-lab protocol and effects of the bioinformatics pipeline on results. Several studies have addressed the validation of this method (see Table 1 in [28]) but very few have examined the kind of large data sets now common in population genomics, containing more than a few thousand SNPs [22,27,32,51]. Here we analysed two populations of *Arabidopsis lyrata* by sequencing pools of individuals, and sequencing the same individuals separately by GBS. The main objective was to compare SNP frequencies obtained by Pool-seq with GBS-based SNP frequencies. Overall, we found that concordance correlation coefficients between SNP frequencies based on the two methods were high, between 0.827 and 0.955. These values are well within the range of other validation studies of pooled sequencing (e.g. Table 1 in [28]). Concordance increased with the pool size, with mean individual sequencing depth in the pool, and with the use of Snape as compared to VarScan as SNP calling software for the pooled samples.

The comparison of different numbers of individuals pooled was based on comparing 14 individuals with sequencing depth per individual of 2.6× and 25 individuals sequenced on two lanes with sequencing depth per individual of 2.3×. With the frequentist SNP caller VarScan, the concordance correlation coefficient increased from 0.827 to 0.931, while the mean absolute difference between SNP frequency estimates from the two methods decreased from 0.109 to 0.073 (Table 1, Fig 3, Fig 4). With the Bayesian-based SNP caller Snape, the concordance correlation coefficient increased from 0.864 to 0.941, while the mean absolute difference between SNP frequency estimates from the two methods decreased from 0.103 to 0.067. These results clearly show that an increase in the number of individuals that are pooled – at least for the range we worked with – improves the accuracy of SNP frequency estimation, as predicted by several theoretical studies [8,10,33]. Similar to our results, those of another study on pooling different

numbers of isofemale lines of *Drosophil*a revealed increases in concordance correlation coefficients from 0.822-0.867 with 22 lines to 0.906-0.934 and 0.911-0.936 with 42 lines [27]. Aside from this, we found that increasing the number of pooled individuals did not greatly increase the chance of detecting SNPs, at least not with sequencing depth per individual used here. The false negative rate remained almost unchanged, increasing slightly from 0.270 to 0.287 with VarScan, and from 0.137 to 0.146 for Snape.

The comparison of varying depth of coverage per site revealed further potential for improving SNP frequency estimates. An increase of the depth of sequencing per individual from 1.4× to 2.3×, 3.2×, and 4.1×, led to an increase in concordance of Pool-seq with GBS (Fig 3) and a decrease in the absolute difference between SNP frequency estimates between methods (Fig 4) and false negative rate (Table 1). In line with our results, a sequencing study on a pool of 30 individuals of the pine processionary moth [32] revealed improved frequency estimates when the sequencing depth was increased from a range of 10×-50× to >200×, equivalent to a depth per individual of 0.3×-1.7× to >6.7×. The authors observed an increase in the correlation coefficient from 0.93 to >0.99 (across different sequencing depths per individual for individual sequencing) and a decrease of the median of the absolute difference between individual-based and pooled-based frequency estimates from 0.067 to 0.007.

A major issue with the Pool-seq technique is a lack of power to detect rare alleles [9,27,33], which is unimportant for some applications but important for others. For example, rare alleles may be important for explaining phenotypic variation within populations [52] and therefore desirable to detect in genome-wide association studies. We investigated this issue by analyzing the minor allele frequency of false negative SNPs (SNPs that were called only in GBS but not in the Pool-seq samples). In all library/lane combinations, the majority of false negative SNPs had low minor allele frequencies (Fig 5). At the sequencing depth of 4.1× per individual in the pool with 25 individuals the majority of GBS SNPs not detected by Pool-seq had a frequency below 0.05 (mean = 0.045 for

VarScan and mean = 0.036 for Snape; Table 1). For higher GBS-based SNP frequencies, the number of SNPs missed by Pool-seq rapidly decreased. This result supports the utility of our upper pool size and maximum depth of sequencing. It has been suggested that to detect a minor allele with near-certainty, its frequency must be larger than 10 divided by the number of pooled diploid individuals [33], which in our study would have been 0.4 for the larger population. We appeared able to detect all minor alleles with frequency > 0.15 at the largest pool size and sequencing depth tested (Fig. 5). The discrepancy is likely due to the difference in variant calling approaches and the fact that we used a $P = 0.05$ threshold for detection as opposed to the $P = 0.001$ level used by Lynch et al. [33]. For some population genetics studies this detection threshold is likely to be acceptable and our results confirm that this kind of pooled data is useful for detecting common minor alleles. Of course, those considering Pool-seq should be aware of the limitation of this approach in detecting rare alleles.

Several SNP callers can be applied to pooled data (reviewed in [8]). We used VarScan [38], which uses a frequentist approach, and Snape [35], which uses a Bayesian approach. Both take into account sequencing depth, base quality, and statistical significance, while Snape includes information on nucleotide diversity and divergence from the reference genome to detect SNPs. Our results show that Snape called considerably more SNPs than VarScan (Fig 2). The number of SNPs called by Snape that were confirmed by GBS was on average 20% higher than the number of SNPs called by VarScan confirmed by GBS (column SNP_{both} in Table 1). Furthermore the false negative rate was found to be systematically lower with Snape. Therefore, it can be argued that Snape is more powerful at detecting SNPs than is VarScan. This may however be accompanied by an increase in the false positive rate, which is an important avenue for further investigation. Also, the concordance correlation coefficients between GBS and Pool-seq SNP frequencies were slightly higher with Snape than with VarScan, although this difference between SNP callers declined with increasing sequencing depth (Table 1, Fig 3). The absolute difference in SNP frequencies between

methods was lower with Snape than with VarScan. These results indicate that the use of priors for nucleotide diversity and divergence contribute positively to the calling of SNPs.

In conclusion, we have presented a method that uses low input DNA (1.6 ng per individual) and widely-available commercial kits to perform pooled whole-genome re-sequencing. Thanks to the tagmentation step, we avoided fragmentation by sonication, which requires more input DNA. We validated SNP frequencies by comparison with GBS data. Our study strengthens the conclusion that the quality of pooled sequencing data sets relies on two critical parameters: the number of individuals that are pooled, and sequencing effort. In a recent review on Pool-seq [8], the authors recommend pools of at least 40 individuals with sequencing depth of more than 50× per pool. Lynch et al. [33] used a maximum likelihood estimator and suggested more than 100 individuals and a sequencing depth of 100× per pool to obtain high confidence in allele frequency estimates. Based on the empirical comparison we performed, we find that a pool of 25 individuals combined with a sequencing depth of 100× produces SNP frequency data with satisfactory precision and accuracy. We confirm that Pool-seq is a useful method to detect genomic variants with a frequency of about 0.05 and larger.

Acknowledgements

Christian Beisel and Daniel Berner gave advice on wet lab protocols, Luca Ferretti and Robert Kofler on bioinformatics pipelines and data analysis. We thank two anonymous reviewers for helpful comments that improved this article. Wet lab work and sequencing was done at: the Genetic Diversity Centre, ETH Zürich; the Functional Genomics Centre Zürich, ETH Zürich and University of Zürich; the Quantitative Genomics Facility Basel, ETH Zürich-Basel and University of Basel.

References

1. Dorfman R. The detection of defective members of large populations. Ann Math Stat. 1943;14: 436–440.

2. Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA pooling: a tool for large-scale association studies. Nat Rev Genet. 2002;3: 862–871. doi:10.1038/nrg930

3. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380. doi:10.1038/nature03959

4. Pandey V, Nutter RC, Prediger E. Applied Biosystems SOLiD™ System: ligation-based sequencing. In: Janitz M, editor. Next Generation Genome Sequencing: Towards Personalized Medicine. 2008. pp. 29–42. doi:10.1002/9783527625130.ch3

5. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456: 53–59. doi:10.1038/nature07517

6. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323: 133–138. doi:10.1126/science.1162986

7. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475: 348–352. doi:10.1038/nature10242

8. Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014;15: 749–763. doi:10.1038/nrg3803

9. Cutler DJ, Jensen JD. To pool, or not to pool? Genetics. 2010;186: 41–43. doi:10.1534/genetics.110.121012

10. Anderson EC, Skaug HJ, Barshis DJ. Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. Mol Ecol. 2014;23: 502–512. doi:10.1111/mec.12609

11. Holt KE, Teo YY, Li H, Nair S, Dougan G, Wain J, et al. Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. Bioinformatics. 2009;25: 2074–2075. doi:10.1093/bioinformatics/btp344

12. Burke MK, Liti G, Long AD. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of *Saccharomyces cerevisiae*. Mol Biol Evol. 2014;31: 3228–3239. doi:10.1093/molbev/msu256

13. Clément JAJ, Toulza E, Gautier M, Parrinello H, Roquis D, Boissier J, et al. Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways under selection in two inbred *Schistosoma mansoni* strains. PLoS Negl Trop Dis. 2013;7: e2591. doi:10.1371/journal.pntd.0002591

14. Pespeni MH, Sanford E, Gaylord B, Hill TM, Hosfelt JD, Jaris HK, et al. Evolutionary change during experimental ocean acidification. Proc Natl Acad Sci. 2013;110: 6937–6942. doi:10.1073/pnas.1220673110

15. Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin S V. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet. 2010;42: 260–263. doi:10.1038/ng.515

16. Fischer MC, Rellstab C, Tedder A, Zoller S, Gugerli F, Shimizu KK, et al. Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. Mol Ecol. 2013;22: 5594–5607. doi:10.1111/mec.12521

17. Kofler R, Betancourt AJ, Schlötterer C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. 2012;8: e1002487. doi:10.1371/journal.pgen.1002487

18. Fabian DK, Kapun M, Nolte V, Kofler R, Schmidt PS, Schlötterer C, et al. Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Mol Ecol. 2012;21: 4748–4769. doi:10.1111/j.1365-294X.2012.05731.x

19. Bastide H, Betancourt AJ, Nolte V, Tobler R, Stöbe P, Futschik A, et al. A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. PLoS Genet. 2013;9: e1003534. doi:10.1371/journal.pgen.1003534

20. Corander J, Majander KK, Cheng L, Merilä J. High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. Mol Ecol. 2013;22: 2931–2940. doi:10.1111/mec.12174

21. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464: 587–591. doi:10.1038/nature08832

22. Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods. 2008;5: 247–252. doi:10.1038/nmeth.1185

23. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013;495: 360–364. doi:10.1038/nature11837

24. Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. 2012;109: 19529–19536. doi:10.1073/pnas.1217149109

25. Esteve-Codina A, Paudel Y, Ferretti L, Raineri E, Megens H-J, Silió L, et al. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. BMC Genomics. 2013;14: 148. doi:10.1186/1471-2164-14-148

26. Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P. Who is eating what: diet assessment using next generation sequencing. Mol Ecol. 2012;21: 1931–1950. doi:10.1111/j.1365-294X.2011.05403.x

27. Zhu Y, Bergland AO, González J, Petrov DA. Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. PLoS One. 2012;7: e41901. doi:10.1371/journal.pone.0041901

28. Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC. Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. PLoS One. 2013;8: e80422. doi:10.1371/journal.pone.0080422

29. Futschik A, Schlötterer C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics. 2010;186: 207–218. doi:10.1534/genetics.110.114397

30. Pérez-Enciso M, Ferretti L. Massive parallel sequencing in animal genetics: wherefroms and wheretos. Anim Genet. 2010;41: 561–569. doi:10.1111/j.1365-2052.2010.02057.x

31. Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A. Detecting selective sweeps from pooled next-generation sequencing samples. Mol Biol Evol. 2012;29: 2177–2186. doi:10.1093/molbev/mss090

32. Gautier M, Foucaud J, Gharbi K, Cézard T, Galan M, Loiseau A, et al. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Mol Ecol. 2013;22: 3766–3779. doi:10.1111/mec.12360

33. Lynch M, Bost D, Wilson S, Maruki T, Harrison S. Population-genetic inference from pooled-sequencing data. Genome Biol Evol. 2014;6: 1210–1218. doi:10.1093/gbe/evu085

34. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, et al. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One. 2011;6: e15925. doi:10.1371/journal.pone.0015925

35. Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S, Pérez-Enciso M. SNP calling by sequencing pooled samples. BMC Bioinformatics. 2012;13: 239. doi:10.1186/1471-2105-13-239

36. Boitard S, Kofler R, Françoise P, Robelin D, Schlötterer C, Futschik A. Pool-hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. Mol Ecol. 2013;13: 337–340. doi:10.1111/1755-0998.12063

37. Ferretti L, Ramos-Onsins SE, Pérez-Enciso M. Population genomics from pool sequencing. Mol Ecol. 2013;22: 5561–5576. doi:10.1111/mec.12522

38. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6: e19379. doi:10.1371/journal.pone.0019379

39. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22: 568–576. doi:10.1101/gr.129684.111

40. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013;22: 3124–3140. doi:10.1111/mec.12354

41. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;

42. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011;43: 476–481. doi:10.1038/ng.807

43. Picard: a set of tools for working with next generation sequencing data in the BAM format. [Internet]. Available: http://broadinstitute.github.io/picard/

44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

45. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0 [Internet]. 2010. Available: http://www.repeatmasker.org

46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

47. Stevenson M, Nunes T, Heuer C, Marshall J, Sanchez J, Thornton R, et al. epiR: tools for the analysis of epidemiological data. R package version 0.9-62 [Internet]. 2015. Available: http://cran.r-project.org/package=epiR

48. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45: 255–268.

49. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 2011;12: 35. doi:10.1186/1471-2105-12-35

50. Carr D, Lewin-Koh N, Maechler M. Hexbin: hexagonal binning routines. R package version 1.27.0. In: 2014 [Internet]. Available: http://cran.r-project.org/web/packages/hexbin/index.html

51. Bansal V, Tewhey R, Leproust EM, Schork NJ. Efficient and cost effective population resequencing by pooling and in-solution hybridization. PLoS One. 2011;6: e18353. doi:10.1371/journal.pone.0018353

52. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, et al. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A. 2006;103: 1810–1815. doi:10.1073/pnas.0508483103

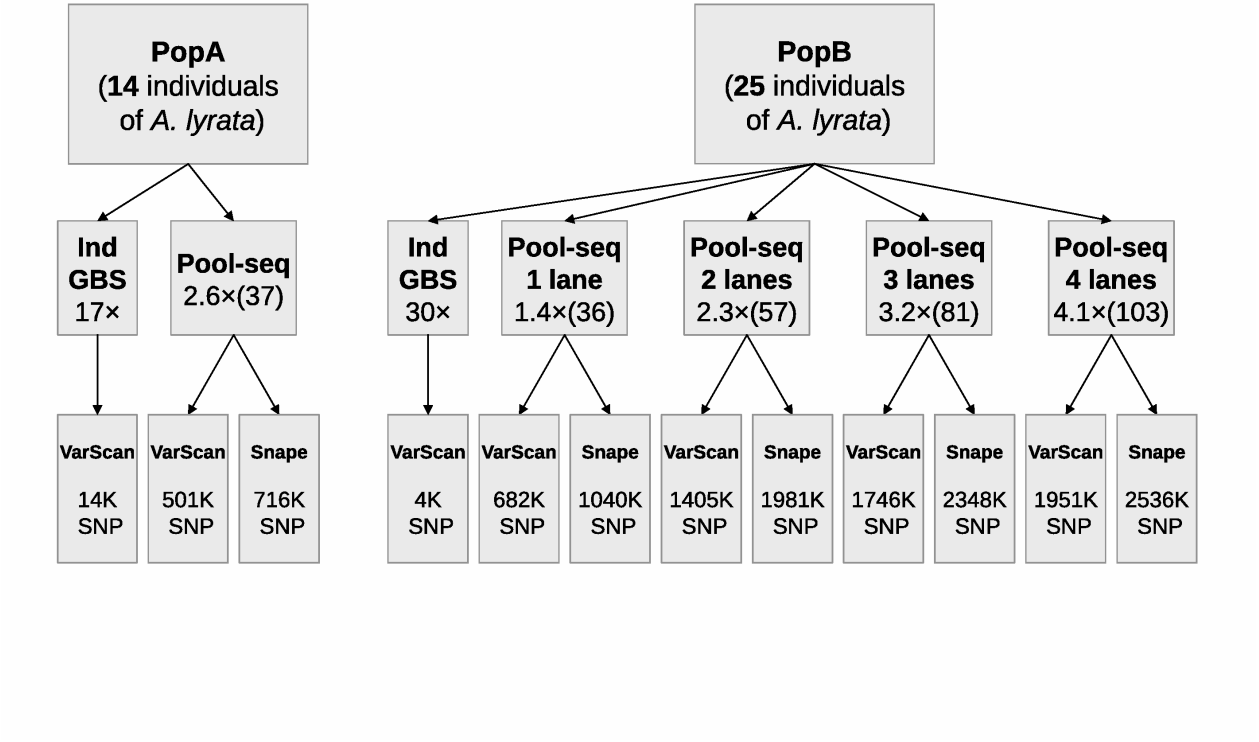


Fig 1. Diagram presenting the data sets produced to validate pooled whole-genome re-sequencing (Pool-seq) by individual-based Genotyping By Sequencing (GBS).

The three rows of boxes contain the following information: top row: name of *Arabidopsis lyrata* population and number of individuals per population; second row: sequencing method, number of lanes merged (Pool-seq, population B only), the sequencing depth per individual and per pool (in parentheses); third row: the number of SNPs called by VarScan and Snape for each data set. Note that for GBS data, only the SNP caller VarScan was used.

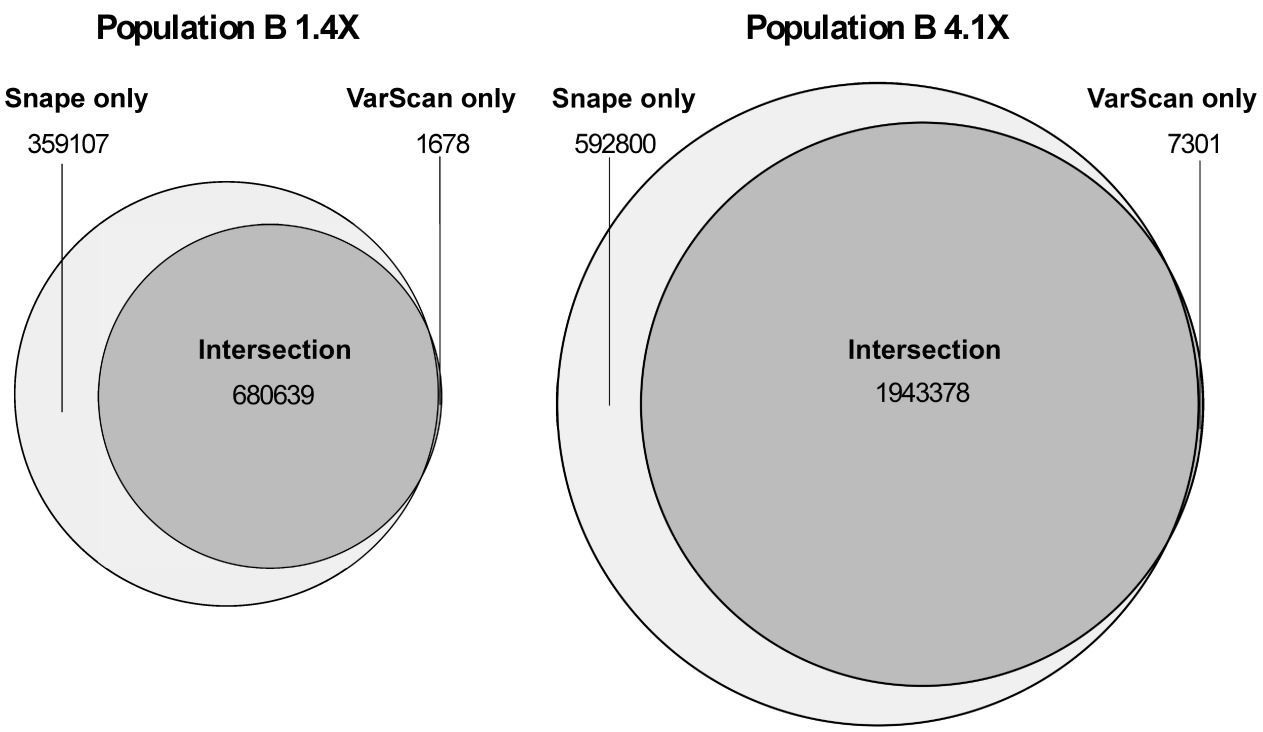


Fig 2. Venn diagram of Pool-seq SNPs called with VarScan (dark grey) and Snape (light grey). The left-hand panel shows the SNPs called for population B using data from lane 1 only. The right-hand panel shows the SNPs called for population B with the data from all four lanes. The figure was produced with the R package VennDiagram [49].

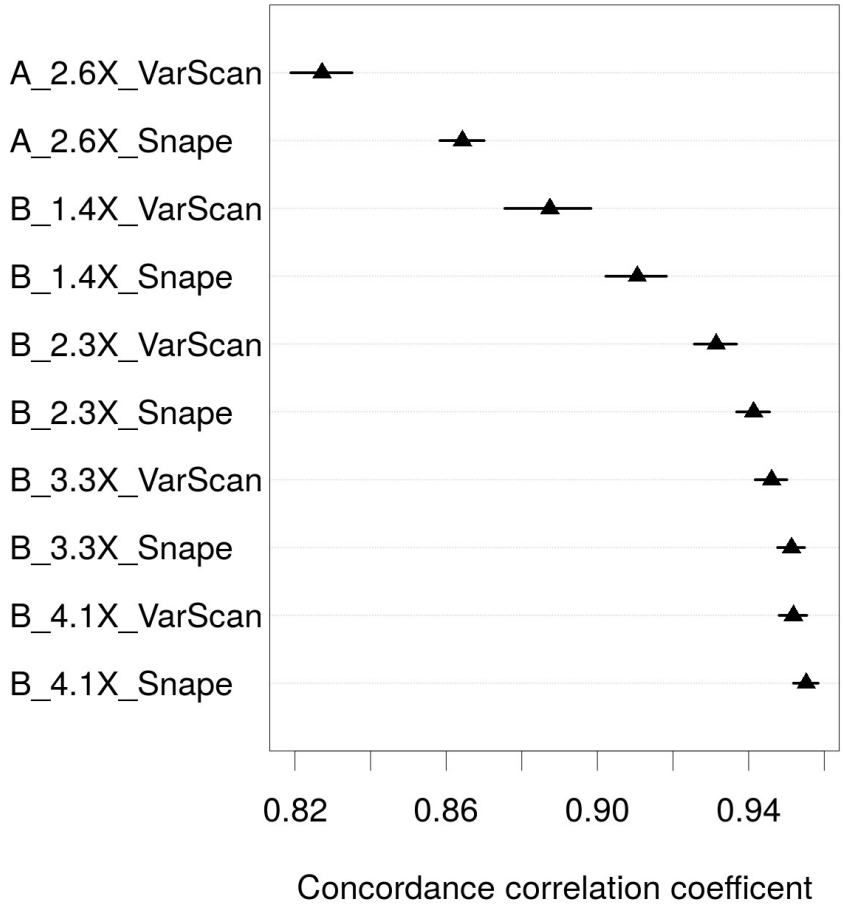


Fig 3. Concordance correlation coefficient between SNP frequencies estimated with Pool-seq and GBS for each library/lane combination and SNP caller.

Mean CCC values with upper and lower 95% confidence ranges are shown. The name of a library/lane combination contains information on: the population (A or B), sequencing depth per individual by Pool-seq, and the software used to detect SNPs for Pool-seq (either VarScan or Snape; for GBS, only VarScan was used).

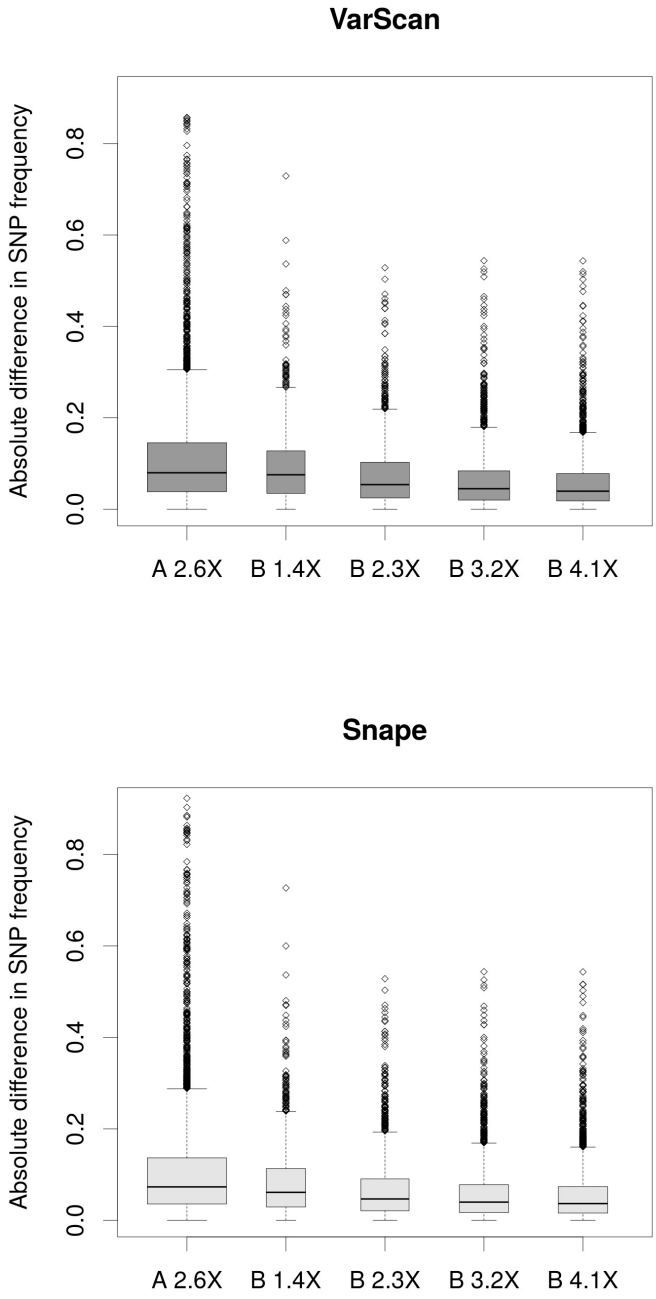


Fig 4. Box plot illustrating the distribution of the absolute difference in SNP frequency estimates between Pool-seq and GBS.

The upper panel (dark grey) shows distributions when SNPs were called with VarScan for Pool-seq, the lower panel (light grey) shows distributions with Snape. Library names contain information on: the population (A or B), and the sequencing depth by Pool-seq. The band inside each box shows the median, while the lower and upper ends indicate the first and third quartile, respectively. The lower

whisker is -1.5x the interquartile range from the first quartile, while the upper whisker is +1.5x the interquartile range from the third quartile. The diamonds represent outliers.

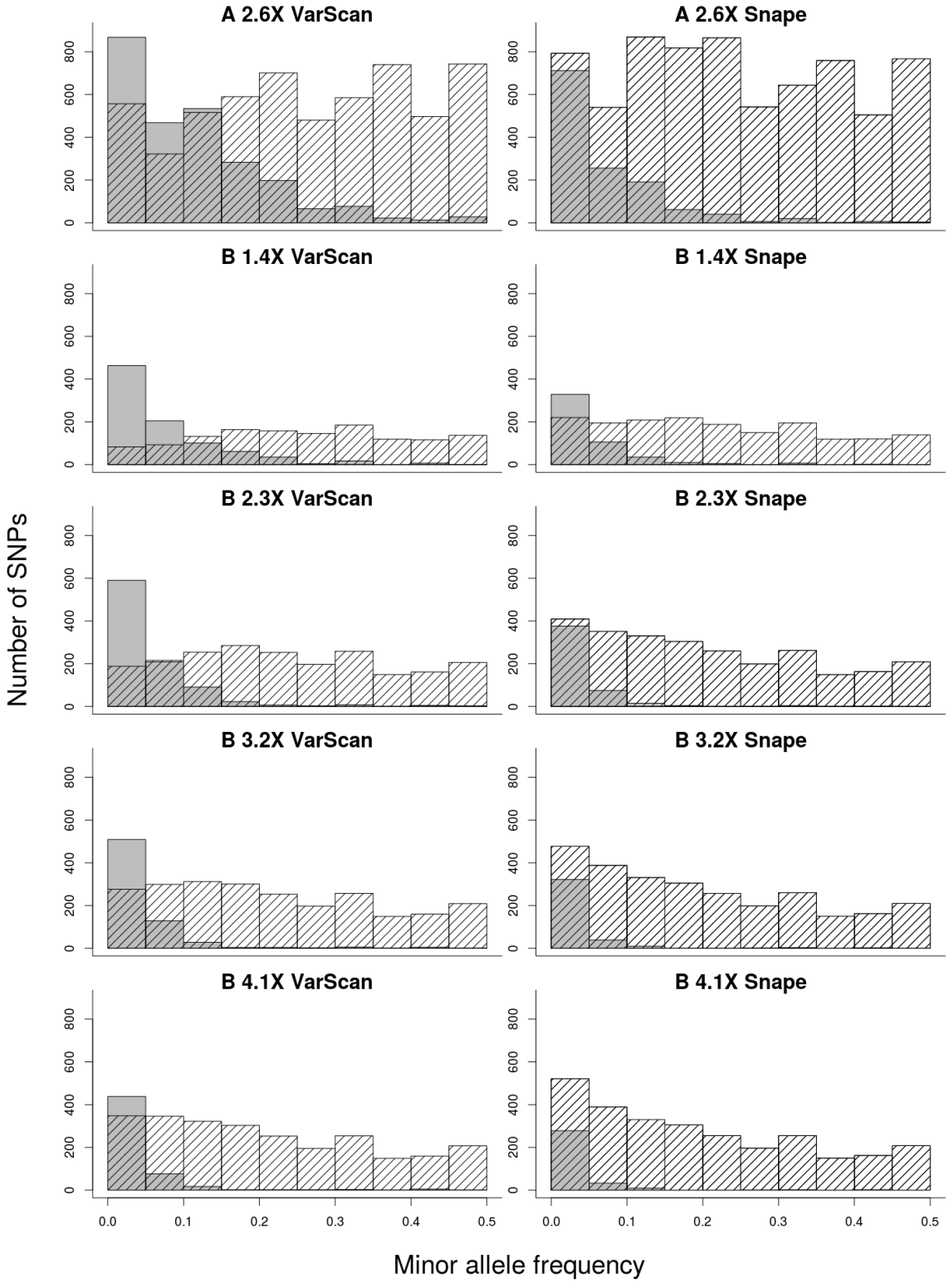


Fig 5. Histogram of minor allele frequency of GBS.

The grey bars represent the SNPs present only in GBS. The striped bars represent the SNPs sequenced in the GBS and Pool-seq samples. The 10 panels show the results for the various Pool-seq library/lane combinations and the two SNP callers. The name of a library/lane combination contains information on: the population (A or B), sequencing depth per individual by Pool-seq, and the software used to detect SNPs for Pool-seq (either VarScan or Snape; for GBS, only VarScan was used).

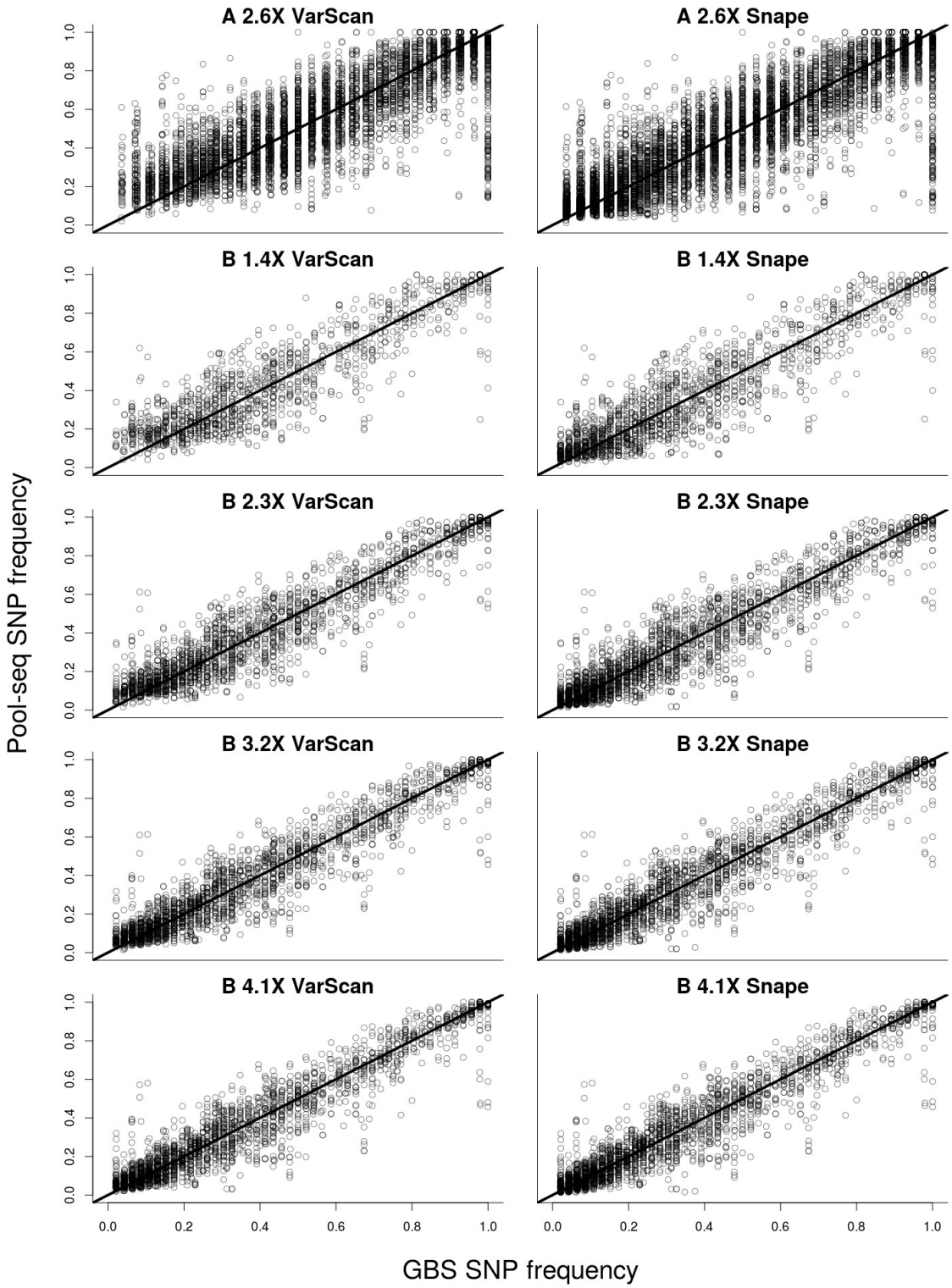


Fig S1. Scatter plots of SNP frequency estimates based on GBS and Pool-seq for the various library/lane combinations and the two SNP callers.

The name of a library/lane combination contains information on: the population (A or B), sequencing depth per individual by Pool-seq, and the software used to detect SNPs for Pool-seq

(either VarScan or Snape; for GBS, only VarScan was used).The solid line indicates the expectation of equal frequency with both sequencing approaches.

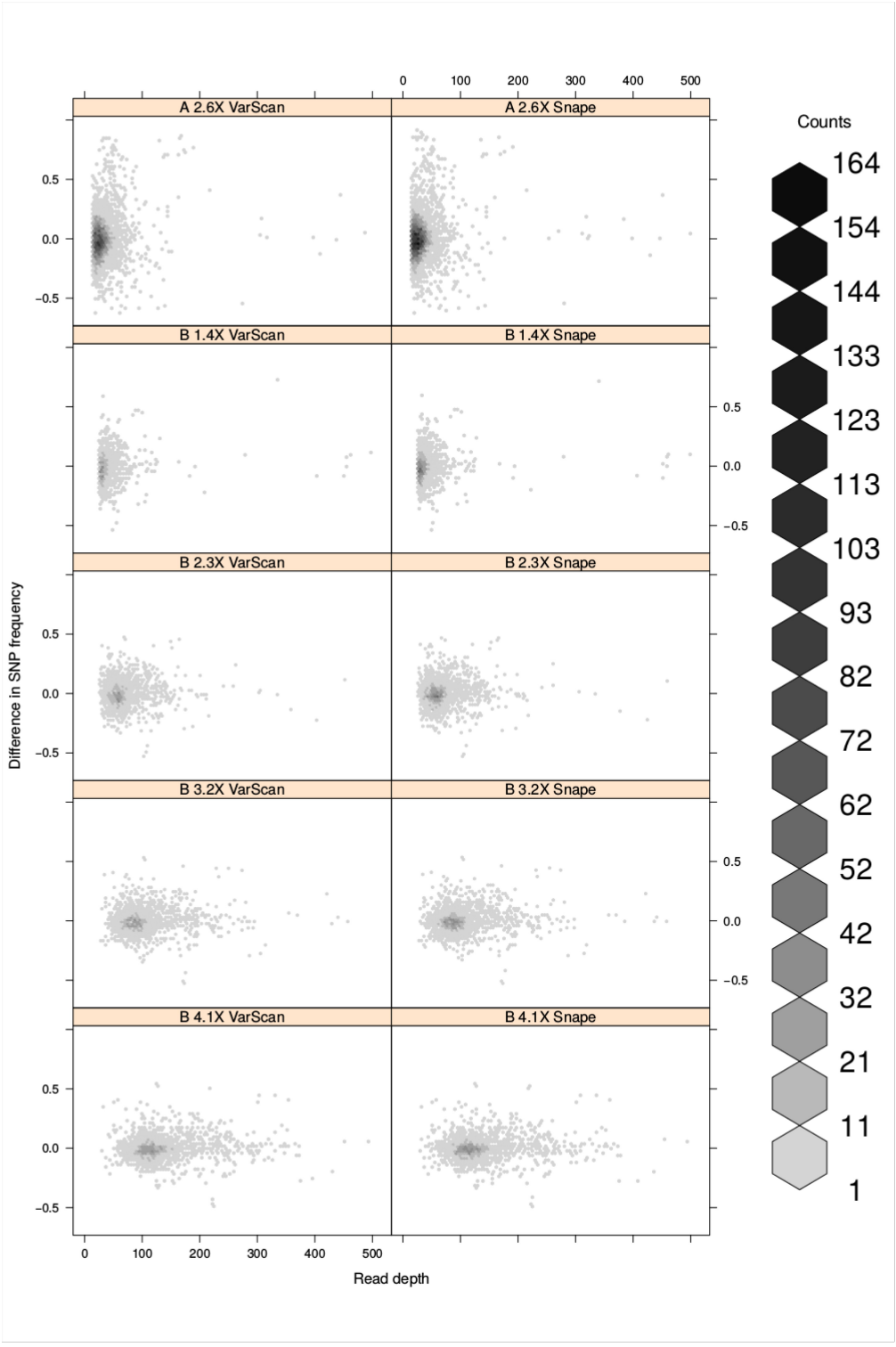


Fig S2. Hexbin plots of the difference in SNP frequency estimates between Pool-seq and GBS with respect to the total read depth at SNP sites of Pool-seq.

The name of a library/lane combination contains information on: the population (A or B), sequencing depth per individual by Pool-seq, and the software used to detect SNPs for Pool-seq (either VarScan or Snape; for GBS, only VarScan was used). Hexagons are shaded by SNP count according to the scale shown on the right. The figure was produced with the hexbin package in R [50].

Table 1. Comparison of SNP numbers and frequency estimate accuracy revealed by Pool-seq and by GBS.

Columns report: library/lane identity (population A or B, estimation of sequencing depth per individual in Pool-seq, and software used to detect SNPs of Pool-seq data set), number of SNPs detected by GBS (SNPGBS) and Pool-seq ($SNP_{Pool\text{-}seq}$), overlapping number of SNPs detected (SNP_{both}), concordance correlation coefficient (CCC) with lower and upper 95% confidence limit (LCL; UCL) of CCC, the mean of the absolute difference in SNP frequency estimates of the two methods (|Δf|), false negative rate (FN rate), that is, the fraction of SNPs called by GBS but not by Pool-seq, and their mean minor allele frequency (FN MAF).

Library/lane ID	**$SNP_{Pool\text{-}seq}$**	**SNP_{GBS}**	**SNP_{both}**	**CCC**	**LCL**	**UCL**	**|Δf|**	**FN rate**	**FN MAF**
A 2.6× VarScan	500’515	13’843	5731	0.827	0.819	0.835	0.109	0.27	0.115
A 2.6× Snape	716’483	13’843	7102	0.864	0.858	0.87	0.103	0.137	0.075
B 1.4× VarScan	682’317	4177	1333	0.887	0.876	0.898	0.092	0.385	0.077
B 1.4× Snape	1’039’746	4177	1754	0.911	0.902	0.918	0.083	0.212	0.054
B 2.3× VarScan	1’405’122	4177	2166	0.931	0.926	0.937	0.073	0.287	0.059
B 2.3× Snape	1’981’376	4177	2636	0.941	0.937	0.946	0.067	0.146	0.043
B 3.2× VarScan	1’745’682	4177	2413	0.946	0.942	0.95	0.063	0.211	0.049
B 3.2× Snape	2’348’269	4177	2738	0.951	0.948	0.955	0.059	0.116	0.038
B 4.1× VarScan	1’950’679	4177	2536	0.952	0.948	0.955	0.058	0.17	0.045
B 4.1× Snape	2’536’178	4177	2771	0.955	0.952	0.958	0.055	0.101	0.036

CHAPTER 2: The role of historic, species-scale and recent local-scale demographic processes in explaining population genomic diversity

Authors: M. Fracassetti[1,2]*, Y. Willi[1,2]

Affiliations:

[1] Institute of Biology, University of Neuchâtel, Rue Emile Argand 11, CH-2000 Neuchâtel, Switzerland

[2] Department of Environmental Sciences, University of Basel, Schönbeinstrasse 6, CH-4056 Basel, Switzerland

*Correspondence to: marco.fracassetti@unibas.ch

TO BE SUBMITTED TO MOLECULAR ECOLOGY

Abstract: Genetic diversity is the raw material on which evolution acts and is therefore of key interest of many fields of biology. Albeit much research has been conducted to understand the role of local demographic factors such as census size and mating system on within-population diversity, it remains often unclear what role historic, species-scale demographic processes play compared to the more recent, local demographic processes. Furthermore, little is known how strongly such putatively neutral demographic processes may differentially affect intergenic and coding regions of the genome. In this study, we estimated genomic diversity of 52 populations of North American *Arabidopsis lyrata* across its entire distribution at different functional regions of the genome. Analysis on the relatedness of populations confirmed the presence of a historic split between an eastern genetic cluster along the Appalachian Mountains and a western genetic cluster west of Lake Erie with evidence of two bouts of past gene flow between clusters. Within clusters, expansion routes since the end of the last glaciation cycle were traced from out of Wisconsin to the north and to Lake Erie, and from the central Appalachians to the north and south. Best predictors of genomic diversity were mating system (selfing compared to outcrossing) followed by historic range dynamics since the last glaciation cycle. Historic demographic processes pre-dating the last glaciation cycle and admixture between clusters had a much smaller impact and was significant only for intergenic regions. Census size did not explain significant variation. The study highlights that for a species with a relatively recent expansion history, this history is one of the most important factors explaining genome-wide genetic diversity.

One Sentence Summary: In a northern-temperate-zone herbaceous plant, current within-population genomic diversity is strongest related to recent mating system shifts to selfing and range dynamics associated with the last glaciation cycle, while old species-scale demographic processes left little imprint.

Keywords: Center-margin hypothesis, genomic diversity, expected heterozygosity, population genetic structure, population history.

Introduction

Within-population genetic diversity is the basis of evolutionary change, for both neutral and adaptive evolution (Barrett & Schluter 2008). Given the importance of genetic diversity, it has been an important entity in evolutionary biology, applied conservation biology and animal and plant breeding. The assessment of genetic diversity is particularly crucial in conservation biology, where it is used to determine the conservation status of a population, with low diversity indicating low current and future fitness that endangers a population's persistence (Reed & Frankham 2003). Similarly plant breeders aim to preserve genetic diversity in order to develop new varieties and hybrids (Govindaraj *et al.* 2015). In these fields, single factors have been estimated for their effect of determining population genetic diversity such as population size or mating system, that explain a considerable amount of variation in genetic diversity (Schoen & Brown 1991; Leimu & Fischer 2008). However, within-population genetic diversity may bear also an important signature of historic demographic processes (Wright & Gaut 2005; Duncan *et al.* 2015). In this study, we assessed the extent to which historic processes across a species range can explain genetic diversity for different regions of the genome, i.e. intergenic and coding regions and the extent to which more recent and local demographic factors such as census size and mating system are important.

Theory predicts that the diversity of neutral genomic regions is affected by population size, mutation rate and gene flow (Wright 1931; Kimura 1955; Slatkin 1981). Population size has a positive effect on the amount of genetic diversity, as it is inversely proportional to genetic drift (Kimura 1955). Genetic drift is defined as the random change in allele frequency, which leads to the fixation and loss of diversity in small populations (Wright 1931). Furthermore, both increased mutation rate and gene flow positively affect genetic diversity (Ohta & Kimura 1973; Slatkin 1981). Diversity of genomic regions may further directly or indirectly – via linkage – be affected by selection, whose impact on diversity depends then on the strength and type of selection. Indeed, directional selection is generally predicted to reduce genetic variation (Smith & Haigh 1974),

whereas balancing selection tends to maintain variation (Dobzhansky 1943), but the impact of selection also depends on the effective population size and the relative magnitude of genetic drift (Wright 1931; Frankham 1996). Also the mating system is predicted to have a strong effect on within-population genetic diversity, where e.g., selfing reduces the effective population size (N_e) by one half (Pollak 1987). In summary, classic equilibrium-based theoretical models predict a positive effect of population size, mutation rate, gene flow and outcrossing on genetic diversity and mixed effects of selection depending on their type and strength (reviewed in (Willi *et al.* 2006)).

Contemporary within-population genetic diversity may not only be shaped by recent and local demographic parameters, but also by historic, large-scale processes such as species retractions due to major disturbance events, long-term isolation, re-colonization, and admixture between formerly separated clusters. Such dynamics were found to have been particularly important to species in the northern-temperate zones that were affected by Quaternary ice ages (Hewitt 2000; Fussi *et al.* 2010). Indeed, ancient pollen data (Bennett & Parducci 2006) and phylogeographic studies (Schönswetter *et al.* 2005; Soltis *et al.* 2006) suggest that many plant species survived the last glacial maximum (LGM) by retreating to refugia, where the climatic conditions allowed species persistence. At the end of the ice ages many of these species recolonized the newly ice-free areas. The prediction is that post-glacial range dynamics led to the situation of high genetic diversity in areas where the species persisted, including the refugia from which recolonization occurred (Keppel *et al.* 2012). Along the expansion route or leading species edge, a series of founder events is expected to have left a signature of genetic drift and a decline in heterozygosity (Pannell & Dorken 2006; Hallatschek *et al.* 2007). In parallel, rare genetic variants – either standing or new - are predicted to have increased in frequencies at the front wave of expansion, a phenomenon called "gene surfing" (Klopfstein *et al.* 2006; Excoffier *et al.* 2009). The opposite edge of the species distribution, called trailing or rear edge, is predicted to have been affected by a dynamics of small population size and prolonged isolation that reduced within-population genetic diversity and

increased genetic differentiation among them (Petit *et al.* 2003; Hampe & Petit 2005). In summary, species affected by Quaternary Ice Ages are predicted to have generally reduced genetic diversity at both leading and trailing edges.

Based on the observation that many species have smaller population sizes at the edges compared to the center of distribution a somewhat related hypothesis, the center-margin hypothesis was formulated. Originally, the hypothesis was motivated based on the observation that for many species, densities decline towards the edge of distribution (Hengeveld & Haeck 1982; Brussard 1984). Eckert et al. (Eckert *et al.* 2008) investigated whether the hypothesis was generally supported by published neutral marker studies. The authors found that 64.2% of the studies detected a decline in within-population genetic diversity and 70.2% of the studies detected an increase in among-population genetic differentiation. However, most of these empirical studies estimated genetic diversity using only neutral portions of the genome and results were interpreted to reflect local long-term population size, gene flow, and evolutionary potential. The next generation sequencing (NGS) revolution allows the estimation of genetic diversity on the level of the entire genome (e.g. in plants: (Cao *et al.* 2011; Branca *et al.* 2011; Mace *et al.* 2013)). Hence, the hypothesis about the effect of different players affecting genetic diversity can be tested for distinct parts of the genome that may be differentially affected by selection and thus allows determining how well estimates of presumably neutral genomic variation reflect genomic variation in expressed genes.

We quantified the relative importance of historic range dynamics compared to current local demographic parameters in explaining genetic diversity in *Arabidopsis lyrata* spp. *lyrata*. The subspecies is a short-lived perennial plant, pre-dominantly outcrossing and closely related to the plant model species *A. thaliana* (Hu *et al.* 2011). We analyzed 52 populations across its entire geographic range in North America, which extends from North Carolina and Missouri to upstate New York and Ontario. Previous studies based on microsatellites data identified an old split

between an eastern and a western genetic cluster (Hoebe *et al.* 2009; Willi & Määttänen 2010). We first reconstructed the phylogeographic history of *A. lyrata* based on nuclear single nucleotide polymorphism (SNP) frequencies and identified possible refugia during the LGM. We then compared genomic diversity estimates based on genome-wide SNP frequencies and published microsatellite-based genetic diversity estimates (Griffin & Willi 2014). Lastly, we tested how the phylogeographic history, admixture events, local census size and the mating system affected genome-wide genetic diversity for intergenic regions, introns and coding regions (CDS).

Material and methods

Population sampling and library preparation.

Populations of *Arabidopsis lyrata* ssp. *lyrata* were collected during the reproductive season in 2007, 2011 and 2014 (Table S1). In this manuscript, populations were named by the state/province abbreviation followed by a number that sorted populations along latitude for US populations and longitude for Ontario (ON) populations. The sampled populations covered the whole known range of the species (Schmickl *et al.* 2010; Paccard *et al.* 2016). In total 52 populations were analyzed, of which 50 populations had been previously analyzed at 19 microsatellite loci (Griffin & Willi 2014). For the remaining two populations, microsatellite genotyping was done as described in Griffin & Willi. Subsequent analyses revealed that ON1 was selfing, with an F_{IS} value of 0.7. For each population one library was prepared with the Nextera Kit (Illumina, San Diego, CA, USA) from 25 equimolarly pooled DNA samples. We followed the library preparation protocol described in Fracassetti et al. 2015. Each library was paired-end sequenced for 100 bases (PE100) on four Illumina HiSeq2000 lanes, using one quarter of the lane each time. This approach of estimating SNP frequencies was previously compared with SNP frequencies based on individual-level representation sequencing and very high congruence in estimates was found (Fracassetti *et al.* 2015).

Bioinformatic pipeline

Lane-by-lane, raw sequences were trimmed with a base quality threshold of 20 using the Perl script *trim-fastq.pl* that is part of the software package PoPoolation (Kofler *et al.* 2011) Trimming was done only from the 3' end to allow the subsequent removal of duplicates. Reads were mapped with BWA-MEM (Li 2013) against the reference using default parameters. The reference was the nuclear genome of *A. lyrata* v1.0 (Hu *et al.* 2011) and the chloroplast and mitochondrial genomes of *Arabidopsis thaliana* (Lamesch *et al.* 2012). Two regions of scaffold 2 of the *A. lyrata* reference genome were masked (position ranges: 8746475-8835273 and 9128838-9212301) because these regions shared very high similarity with the *A. thaliana* chloroplast genome, suggesting an assembly error in the *A. lyrata* genome. Data of the different lanes were subsequently merged and only reads that mapped against scaffolds I-VIII – representing the eight chromosomes of *A. lyrata* – were retained for the further analyses.

Further filtering steps were applied: duplicate reads were removed with the MarkDuplicates tool of Picard v.2.5.0 (http://broadinstitute.github.io/picard/) and only proper paired reads with a mapping quality score above 20 were retained. The reads belonging to the three different regions (intergenic, introns and coding DNA sequencing) were filtered with BEDTools (Quinlan & Hall 2010). The selection of intergenic regions, introns and coding DNA sequences (CDS) was based on the newest annotation of *A. lyrata* (Rawat *et al.* 2015). Intergenic regions were defined as regions 1000 bp away from the 5' and 3' untranslated regions (UTR) of each gene. Pileup files for each scaffold and for each region were created with SAMtools (Li *et al.* 2009). Each pileup file was filtered to retain regions with depth of coverage per site of 25-500. Indels (inserts, deletions) were called with the command pileup2indel of the program VarScan (Koboldt *et al.* 2012) for each population. The regions near insertions and deletions were identified (identify-genomic-indel-regions.pl) and removed (filter-pileup-by-gtf.pl) with PoPoolation (Kofler *et al.* 2011). The genomic interspersed

repeats were identified in the reference genome with RepeatMasker (Smit *et al.* 2010) using the default settings for "arabidopsis" and removed from the pileup files. SNPs were called with the command pileup2snp of the program VarScan (Koboldt *et al.* 2012) for each population. We retained only bi-allelic SNPs, with a minimum count of the variant allele of 3, a minimum frequency of the variant allele of 0.015, a P-value lower than 0.15, and minimum mapping quality of 20. Cut-off parameters were chosen to balance the removal of false positives and retaining true rare variants**.** Finally, SNPs with a strand bias of more than 90% were filtered out.

Population relatedness and genomic diversity esimates

The relatedness tree of the *A. lyrata* populations was estimated with TreeMix (Pickrell & Pritchard 2012) on 127,726 SNPs present at nucleotide sites sequenced in all populations. The SNP frequencies of the *Arabidopsis halleri* population Ha31 (Fischer *et al.* 2013) were used for rooting the tree. The analysis was conducted with a SNP window size (-K) of 500. This corresponds to a windows size of 0.2 million bp which exceeds the known extent of linkage disequilibrium of *A. lyrata* (Ross-Ibarra *et al.* 2008). We allowed seven migration events and we tested them with the four-population test (Reich *et al.* 2009) implemented in TreeMix (Pickrell & Pritchard 2012). Additional migrations events were tested but they were not significant with the four-population test. TreeMix was run 100 times and the tree with the highest maximum likelihood value was selected. The most basal split separated the Ozark populations (MO) from all others, and the next split the previously found western and eastern ancestral cluster. For these two clusters separately, the location of ancestral populations - relative to each internal node in the Treemix tree topology - was determined with the phylo.to.maps function of the R package phytools (Revell 2012). The time calibration of the tree was performed with the chronos function in the R package *ape* (Paradis 2013) using a "correlated" model with a smoothing parameter (λ) equal to 0 and 10 branch categories, which permit different mutation rate between branches. One calibration point was used, the

presumable split time between *A. lyrata* and *A. halleri* of (334'400 years) estimated based on 29 nuclear loci (Roux *et al.* 2011). Usually time calibration requires that branch lengths reflect the number of substitutions. Based on coalescent theory, the time to fixation should take $2N_e$ generations. However, our relatedness tree had branch lengths which were estimated based on frequency data and represented the drift parameter, which is time divided by $2N_e$. Therefore we assumed that the two types of branch lengths were proportional.

Three estimates of genomic diversity were calculated. We analyzed the pileup files with NPStat (Ferretti *et al.* 2013) in 5000 bp windows using only biallelic SNPs called by VarScan. For intergenic regions, introns and coding sequence regions (CDS) separately, we calculated nucleotide diversity, π (Nei & Li 1979), Watterson's Theta, θ (Watterson 1975), and Tajima's D (Tajima 1989). We then took the weighted median across windows based on the number of sequenced bps (table S1). Interpolation maps (fig 7-9-11) were generated with the akima R package (Akima *et al.* 2015). For the production of maps, further sources were: state lines (http://gadm.org/), waterways (http://www.naturalearthdata.com/), maximum extent of the ice sheet (Clark *et al.* 2009).

Determinants of genomic diversity

To identify the underlying processes that shaped patterns of genomic diversity, we assessed the relation of five explanatory variables (table S2) with the weighted medians of π, θ, and Tajima's D of intergenic regions and CDS regions in linear models. Analysis for intron regions were not done as genomic estimates were highly correlated with those of exons/coding regions (all $R^2 > 0.96$, table 1). A first explanatory variable of the linear model was the ancestral cluster membership based on the population relatedness tree (Figure 2; blue and purple shapes in figure 4), reflecting the oldest historic split in this species. Since the MO1 and MO2 populations showed signature of admixture with the southern populations of the western genetic cluster, they were assigned to the western cluster. The second explanatory variable was the geographic distance from a core point of each of

the two ancestral clusters, defined as the node from which expansions happened since about the last glacial maximum (blue and purple circle in figure 3). This ancestral population can be thought as having given raise to the leading edge of distribution, while populations that had split before where considered rear-edge relative to the core site. For leading edge populations (blue tips in figure 3), we calculated the sum of the great-circle distances back to all ancestral populations until the presumable refugium (core) population was reached, considering the entire expansion route. For rear-edge populations, we calculated the direct great-circle distance to the core population.

The third explanatory variable was census size, it was estimated by the surface area occupied by the plants of a population multiplied by a measure of mean local density (Willi & Määttänen 2011), which was $\log_{10}$-transformed. The fourth explanatory variable was the role of mating system, i.e. whether a population was predominantly outcrossing or selfing (two mixed-mating populations were considered as selfing). The mating system was inferred from the population inbreeding coefficient (F_{IS}) of 19 microsatellite markers (Griffin & Willi 2014). It had been demonstrated that F_{IS} is strongly correlated with multi-locus outcrossing rate assessed by progeny-array ($N = 18$, $R^2 = 0.929$, $P < 0.001$, figure S1 in Griffin & Willi 2014). The fifth explanatory variable depicted admixture events (binary: 0/1) between the two genetic clusters suggested by TreeMix. The relative importance of these variables was assessed with the R package relaimpo (Grömping 2006) using the averaging over orderings (Lindeman *et al.* 1980).

The pipelines for analysis were written in BASH (Fox 1989) and R (R Core Team 2015) and are accessible at: http://github.com/fraca. The sequences were stored at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) with the accession number XXXXXX.

Results

Sequencing statistics

Sequencing of all populations yielded more than 13 billion mapped paired-end reads. After applying

the read depth cutoff (25-500X) and removing duplicates, on average 220 millions paired-end reads per populations (range: 128-323 millions) mapped unambiguously to 67% (range: 62-72%) of the *A. lyrata* nuclear genome with a mean depth of 128X (range: 72-188X). The mean number of biallelic SNPs called per population was 1.5 million (range: 0.7-2.8 millions). Overall 1.5% of the sequenced base pairs were SNPs, the percentage was higher in intergenic regions (1.9%), followed by introns (1.2%) and CDS (1%), see (figure 1).

Relatedness tree and historic range dynamics across the species distribution

The topology of the relatedness tree indicated two main genetic clusters: one in the west and the other in the east (figures 2&3). Two populations located in Missouri (MO1 and MO2) formed a third clade that diverged about 253,050 years ago, suggesting that this first split within the species occurred before the Illinoian glaciation 130,000-191,000 years ago (Bowen & Frye 1970). The topology of the tree was in agreement with the relative geographical position of populations and suggested the expansion routes of the *A. lyrata* populations since the LGM (figure 4).

In the eastern cluster the most basal population was on the New Jersey coast line (NJ1, figure 4) with a divergence time about 84,350 years ago (figure 3), suggesting that this population belonged to a lineage that predated the LGM. Other eastern populations emerged from ancestors located in Pennsylvania (blue square in figure 4). The populations that appeared next were located towards the north in New York (NY1-NY3) and in Pennsylvania (PA1). From the ancestor of PA1 some populations seemed to have expanded northwards (PA3, NY4-NY6), whereas others expanded southward along the Appalachian mountains (WV1, MD3, NC1-NC4); and eastwards towards the Atlantic coast (PA2, MD1, MD2, MD4, VA1 and VA2).

In the western cluster the most basal populations were found in Missouri and Iowa (MO3, IA1, IA2 figure 4). Since all these populations had high divergence times (126,525 and 84,350 years ago respectively, figure 3), they seemed to have persisted south of the Laurentide ice sheet

during the LGM. The other western populations emerged from an ancestor located in Wisconsin (purple square in figure 4). The recolonization of regions that became free of ice started in Wisconsin (WI1-WI3) towards the shore of Lake Superior and further northwest to Lake of the Woods (MI5, MI6, ON8-ON12), towards the western shore and later the eastern shore of Lake Michigan and up to Manitoulin Island on northern Lake Huron (IL1, IL2, IN1, MI1-MI4, ON5-ON7). The expansion of the species continued from the eastern shore of Lake Michigan to southern Lake Huron and Lake Erie (ON1-ON4, OH1, and PA4). The populations around Lake Erie showed signature of genomic admixture (figure 2,4). There were two migration events: one from the ancestral population of NY1 and NY2 to ON1, ON2, ON4 and PA1 with a migration weight (w_m) of 41.8% and one from eastern PA1 (w_m = 40%) to ON1 and PA1. We investigated this signature of admixture with the four-population test for treeness (Reich *et al.* 2009). We tested the tree [[NJ1, PA1],[ON3, X]], where X is a western population located north of the Laurentide ice sheet. We obtained significant positive Z scores only for the four eastern populations of Lake Erie (table S3). These results were consistent with a signature of admixture between PA1 and ON1, ON2, ON4 and PA4.

Additional migration events were detected from MO1 to MO3 (w_m = 39.3%), IA1 (w_m = 24.2%), the ancestor of IA2 and MO3 (w_m = 20%) and WI1 (w_m = 15.3%). We investigated also these admixtures with the four-population test. We tested the tree [[MO1, MO2],[WI3, X]], where X was a western population. We obtained significant positive Z scores for three populations with the exception of WI1 (Z score = -0.587, P = 0.557). Another admixture event was found from eastern VA2 to the ancestor of the Missouri populations of MO1 and MO2 (w_m = 28.6%), which could not be tested for significance with the four-population test. The most basal populations in the calibrated relatedness tree showed the longest branch lengths (red tips in figure 3). These populations were located at the rear-edge of the species distribution: on the Atlantic coast in New Jersey for the eastern genetic cluster (NJ1) and in Missouri and Iowa for the western genetic cluster (MO1-MO3,

IA1, IA2). The populations that occurred in areas formerly covered by the Laurentide ice sheet (figure 4) must have appeared after the beginning of the glacial retreat ~18,000-19,000 years ago (Clark *et al.* 2009). Indeed, the estimated age to the most common ancestor for most northern populations was 14,060 years for both genetic cluster (figure 3). Dating their common ancestor supported the hypothesis that these northern populations had been part of the leading-edge of the species distribution. Another part of the leading edge were the populations towards the southern Appalachians.

Comparison of heterozygosity and allelic diversity across genomic regions

Pearson correlation analysis among genomic and microsatellite diversity estimates were overall highly positive (Table 1). Heterozygosity estimates (nucleotide diversity π) were highly correlated across genomic regions (ρ = 0.98-0.99), as well as those estimated based on SNPs and microsatellites (ρ = 0.88-0.92). Also allelic diversity (Watterson's θ) was similar for SNPs of the different genomic regions (ρ = 0.98-0.99) and for the estimate based on SNPs and microsatellites (ρ = 0.89-0.93). Overall SNP-based heterozygosity and allelic diversity were relatively highly correlated for different genomic regions (ρ = 0.95-0.97).

Determinants of genomic diversity

Nucleotide diversity, π. The mating system, the distance from the ancestral core determined by the ancestor that gave rise to populations splitting after the withdrawal of the ice sheet at the end of the last glaciation period and the ancestral cluster membership could explain most of the variation in π across all the populations (table 2). Particularly, the factors that significantly explained the variation of π in intergenic regions were mating system (29.80%), distance from the core (25.90%), ancestral cluster membership (17.30%) and admixture (2.10%). The fractions of variation in π of coding

regions were higher for mating system (32%) and distance from the core (28.80%). Instead, ancestral cluster membership and presence of admixture were not significantly related to genomic diversity in coding regions. Census size was not related to π either of intergenic or coding regions. Selfing populations had lower π than outcrossing populations and eastern populations showed higher π than western populations (figure 5) and π decreased with distance from the core site (figure 6).

Watterson's θ. The mating system, distance from the core and the ancestral cluster membership were correlated with Watterson's θ to similar degrees as they were with π (table 2). Factors that significantly explained variation in θ in intergenic regions were mating system (23.10%), distance from the core (22.10%) and ancestral cluster membership (14.20%). The fractions of variation in θ explained in coding regions were mating system (24.30%) and distance from the core (23.10%). Ancestral cluster membership was not significantly related to Watterson's θ of coding regions, and admixture and census size were also not significantly related to θ of intergenic and coding regions. Selfing populations had lower θ than outcrossing populations and eastern populations showed higher θ than western populations (figure 7), and θ decreased with distance from the core (figure 8).

Tajima's D. Linear model analysis on Tajima's D of intergenic regions showed that the overall modell was not significant ($R^2 = 0.3$, $P = 0.247$), and none of the explanatory variables were significantly associated with the Tajima's D. For the coding region the only explanatory variable significantly associated with Tajima's D was the mating system (17.10%). Selfing populations had lower values of Tajima's D compared to the outcrossing populations (figure 9), indicating a selective sweep in these populations.

Discussion

Genetic diversity estimates have often been studied either for a link with recent and relatively local

demographic factors such as census size (Willi *et al.* 2006; Leimu & Fischer 2008) or a link with historic events such as the reconstruction of population history (Wright & Gaut 2005; Duncan *et al.* 2015). The goal of our study was to compare the relative importance of both historic species-range dynamics and recent, local demographic factors in explaining patterns of within-population genetic diversity. We found that current within-population genetic diversity among north American *Arabidopsis lyrata* had been shaped strongest by recent shifts in the mating system from outcrossing to selfing and relatively recent historic range dynamics going back to the last glaciation cycle. Older historic subdivisions were confirmed, namely the east-west split, but this split did not contribute as much to contemporary patterns of genetic diversity. The allelic diversity (θ) and nucleotide diversity (π) of rear- and leading edge populations decreased from the core of distribution to the margin. These findings were generally consistent for both intergenic regions of the genome as well as for coding DNA sequence (CDS) regions. In contrast, admixture events during range expansion left only a small imprint in π, and census size could not explain genomic diversity.

The main factor shaping genetic diversity was mating system. Selfing populations had both decreased nucleotide diversity, θ, and heterozygosity, π, in intergenic and coding regions (Table 3). Theory predicts that self-fertilization reduces the effective population size (N_e) by one half (Pollak 1987), or possibly more when selfing populations experience founder events and linked selection (Charlesworth & Wright 2001). Empirical evidence for the aforementioned theoretical predictions stem from several plant species (Glémin & Muyle 2014), which includes studies on the same populations as studied here (Willi & Määttänen 2010; Griffin & Willi 2014). The selfing populations studied here mostly occur at range margins of genetic clusters where bottlenecks facilitated the shit to selfing (Baker 1955; Stebbins 1957; Pannell & Dorken 2006). Furthermore, we found in this study that selfing populations had low values of Tajima's D (table 3), which is a sign of increased purifying selection (Nielsen 2005). Increased purifying selection was also found in selfing populations of *Eichhornia paniculata* (Ness *et al.* 2010; Arunkumar *et al.* 2015).

The second most important factor shaping genomic diversity (θ and π) was postglacial range expansion. Dating of the relatedness tree revealed that the majority of populations had diverged since about 14,060 years ago, hence after the start of the withdrawal of the Laurentide ice sheet of the last (Wisconsinan) glaciation cycle 85,000-11,000 years ago (Clark *et al.* 2009). The genomic imprint of geographic range expansion was well visible in the relatedness tree where lineages appeared that were further and further away from the location of the most common ancestor. Or in other words, the tree suggested a geographic route of range expansion that makes sense in the context of a moving wave-front (figure 4). Based on the location of the most recent common ancestor within each genetic cluster, we were able to track down the possible locations of the last glacial refugia. The western cluster most likely had a refugium in the Driftless Area in central Wisconsin, which is consistent with the importance of this region as refugium for many species (Li *et al.* 2013). Recolonization in the east was short in distance and likely happened from the central Appalachians in Pennsylvania. The southern-most populations in the eastern genetic cluster showed short branch lengths in the time calibrated tree (figure 3). Hence, the colonization of the southern Appalachians seem to be a young event, of about the same age as the recolonization of the north. Older populations were found in the west in Missouri and in the east on the Atlantic coast. The recent expansion of the species range dating to the end of the LGM in north and south as well as rear-edge dynamics is most likely responsible for the consistent decline in genomic diversity from core to edge as predicted by theory (Hampe & Petit 2005).

The third factor was the historic species-range dynamics that affected genomic diversity estimates of intergenic regions. The subspecies *A. lyrata* subsp. *lyrata* was suggested to originate from a colonization event of North America via the Bering Strait (Schmickl *et al.* 2010; Novikova *et al.* 2016). The divergence time between the American *A. lyrata* subsp. *lyrata* and the European *A. lyrata* subsp. *petrae* ranges between 130,000 and 300,000 years, assuming a generation time of two

years (Pyhäjärvi *et al.* 2012). Consequently *A. lyrata* subsp. *lyrata* very likely occurred in North America before the Illinoian glaciation period of 191,000-130,000 years ago (Bowen & Frye 1970). Our finding support this statement and The very old divergence time we estimated for (253,050 year ago) the populations located in Missouri may thus reflect a very early split following the initial colonization. The history of our study organism *Arabidopsis lyrata* subsp. lyrata was further impacted by the split between eastern and western populations. This split was previously confirmed using microsatellite data (Willi & Määttänen 2010; Griffin & Willi 2014). An east-west split is a common phenomenon for many North American plant and animal species (reviewed in (Soltis *et al.* 2006)) and likely reflects isolation and differentiation during Pleistocene glaciations. Our time calibrated tree (figure 3) suggested that the split between the two clusters was due to isolation during the Illinoian glaciation 130,000-191,000 years ago (Bowen & Frye 1970). This split was a predictor of genomic diversity of intergenic regions but not of coding regions (table 2). This may be due to the fact that coding regions were more subjected to processes linked to mating system shifts, recent expansion dynamics (table 3), and natural selection that overrode older processes. Gene flow and admixture between populations left only a small signature of increased nucleotide diversity in intergenic regions. Two admixture events were detected in the western cluster: in Missouri, Iowa and in the region of Lake Erie. The first was form the oldest *A. lyrata* populations (MO1, MO2 fig 3) to other southern populations of the western cluster. The second was between the eastern and western cluster, and the direction of admixture was most likely from east to west. Such admixture must have happened after the end of the last glaciation cycle, as the populations involved appeared only then. The fact that it is still associated with increased nucleotide diversity in intergenic regions supports their young age.

Census size showed no impact on the genomic diversity. This is in concordance with a previous study on 18 *A. lyrata* populations that overlapped with our studied sites and which found only a small amount of variation in expected heterozygosity to be explained (6%) variation by

microsatellite marker (Willi & Määttänen 2011). In a meta-level analysis, it was found that census size is overall related to estimates of genetic diversity (Willi *et al.* 2006). We hypothesize that for our species, current census size is a weak estimate of size for a period of some dozen generations because the species occurs in generally disturbed places by fire or erosion by water and wind, where census size can vary a lot over short periods of time.

To conclude, we find genomic diversity to be strongly affected by mating system and more recent processes of range dynamics that occurred since the last glaciation event. A less important factor is the deep species-scale historic demographic that affect intergenic regions. Assuming that these processes are predominantly neutral in effect and not linked to selection, it means that genome-wide diversity, whether in intergenic or in coding regions, is strongly affected by genetic drift. Similar processes are likely to be at play for many other species affect by range dynamics during the quaternary ages. This work contributes to the knowledge of how to genetic diversity is shaped at genome-wide level.

Acknowledgments: We thank the many institutions that granted collection permits, and B. Mable, J. Proffitt, and J. Van Buskirk for help with collecting seeds or plant tissue. Olivier Bachmann gave support with the wet-lab work, and J. Vieu and K. Lucek gave feedback on drafts of this manuscript. Sequencing was done at the Quantitative Genomics Facility Basel, ETH Zürich-Basel and University of Basel, and the Genetic Diversity Centre of ETH Zürich. This work was supported by the Swiss National Science Foundation (PP00P3-123396 and PP00P3_146342), and the Fondation Pierre Mercier pour la Science, Lausanne.

REFERENCES

Akima H, Gebhardt A, Petzoldt T, Maechler M (2015) akima: Interpolation of irregularly spaced data. R package version 0.5-12.

Arunkumar R, Ness RW, Wright SI, Barrett SCH (2015) The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics*, **199**, 817–29.

Baker HG (1955) Self-compatibility and establishment after "long-distance" dispersal. *Evolution*, **9**, 347.

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, **23**, 38–44.

Bennett KD, Parducci L (2006) DNA from pollen: principles and potential. *The Holocene*, **16**, 1031–1034.

Bowen WH, Frye JC (1970) *Pleistocene stratigraphy of Illinois*.

Branca A, Paape TD, Zhou P *et al.* (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, E864-70.

Brussard PF (1984) Geographic patterns and environmental gradients: the central-marginal model in *Drosophila* revisited. *Annual Review of Ecology and Systematics*, **15**, 25–64.

Cao J, Schneeberger K, Ossowski S *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature genetics*, **43**, 956–63.

Charlesworth D, Wright SI (2001) Breeding systems and genome evolution. *Current Opinion in Genetics & Development*, **11**, 685–690.

Clark PU, Dyke AS, Shakun JD *et al.* (2009) The Last Glacial Maximum. *Science*, **325**, 710–4.

Dobzhansky T (1943) Genetics of Natural Populations IX. Temporal Changes in the Composition of Populations of Drosophila Pseudoobscura. *Genetics*, **28**, 162–86.

Duncan SI, Crespi EJ, Mattheus NM, Rissler LJ (2015) History matters more when explaining genetic diversity within the context of the core-periphery hypothesis. *Molecular Ecology*, **24**, 4323–4336.

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. *Molecular Ecology*, **17**, 1170–1188.

Excoffier L, Foll M, Petit RJ (2009) Genetic Consequences of Range Expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

Ferretti L, Ramos-Onsins SE, Pérez-Enciso M (2013) Population genomics from pool sequencing. *Molecular Ecology*, **22**, 5561–5576.

Fischer MC, Rellstab C, Tedder A *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Molecular Ecology*, **22**, 5594–5607.

Fox B (1989) Bash is in beta release!

Fracassetti M, Griffin PC, Willi Y (2015) Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata* (U Melcher, Ed,). *PLOS ONE*, **10**, e0140462.

Frankham R (1996) Relationship of genetic variation to population size in wildlife. *Conservation Biology*, **10**, 1500–1508.

Fussi B, Lexer C, Heinze B (2010) Phylogeography of *Populus alba* (L.) and *Populus tremula* (L.) in Central Europe: secondary contact and hybridisation during recolonisation from disconnected refugia. *Tree Genetics & Genomes*, **6**, 439–450.

Glémin S, Muyle A (2014) Mating systems and selection efficacy: a test using chloroplastic sequence data in Angiosperms. *Journal of Evolutionary Biology*, **27**, 1386–1399.

Govindaraj M, Vetriventhan M, Srinivasan M (2015) Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genetics research international*, **2015**, 431487.

Griffin PC, Willi Y (2014) Evolutionary shifts to self-fertilisation restricted to geographic range margins in North American *Arabidopsis lyrata*. *Ecology Letters*.

Grömping U (2006) Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, **17**.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR (2007) Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences*, **104**, 19926–19930.

Hampe A, Petit RJ (2005) Conserving biodiversity under climate change: the rear edge matters. *Ecology letters*, **8**, 461–7.

Hengeveld R, Haeck J (1982) The distribution of abundance. I. Measurements. *Journal of Biogeography*, **9**, 303–316.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hoebe PN, Stift M, Tedder A, Mable BK (2009) Multiple losses of self-incompatibility in North-American *Arabidopsis lyrata*?: Phylogeographic context and population genetic consequences. *Molecular Ecology*, **18**, 4924–4939.

Hu TT, Pattyn P, Bakker EG *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, **43**, 476–481.

Keppel G, Van Niel KP, Wardell-Johnson GW *et al.* (2012) Refugia: identifying and understanding safe havens for biodiversity under climate change. *Global Ecology and Biogeography*, **21**, 393–404.

Kimura M (1955) Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences of the United States of America*, **41**, 144–50.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–90.

Koboldt DC, Zhang Q, Larson DE *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**, 568–576.

Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE*, **6**, e15925.

Lamesch P, Berardini TZ, Li D *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**, D1202-10.

Leimu R, Fischer M (2008) A meta-analysis of local adaptation in plants. (A Buckling, Ed,). *PloS one*, **3**, e4010.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li P, Li M, Shi Y *et al.* (2013) Phylogeography of North American herbaceous *Smilax* (Smilacaceae): Combined AFLP and cpDNA data support a northern refugium in the Driftless Area. *American journal of botany*, **100**, 801–14.

Lindeman RH, Merenda PF, Gold RZ (1980) *Introduction to bivariate and multivariate analysis*.

Mace ES, Tai S, Gilding EK *et al.* (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature communications*, **4**, 2320.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–73.

Ness RW, Wright SI, Barrett SCH (2010) Mating-system variation, demographic history and patterns of nucleotide diversity in the Tristylous plant *Eichhornia paniculata*. *Genetics*, **184**, 381–92.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Novikova PY, Hohmann N, Nizhynska V *et al.* (2016) Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, **48**, 1077–1082.

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, **22**, 201.

Paccard A, Van Buskirk J, Willi Y (2016) Quantitative genetic architecture at latitudinal range boundaries: reduced variation but higher trait independence (CG Eckert, JL Bronstein, Eds,). *The American Naturalist*, 000–000.

Pannell JR, Dorken ME (2006) Colonisation as a common denominator in plant metapopulations and range expansions: effects on genetic diversity and sexual systems. *Landscape Ecology*, **21**, 837–848.

Paradis E (2013) Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion. *Molecular Phylogenetics and Evolution*, **67**, 436–444.

Petit RJ, Aguinagalde I, de Beaulieu J-L *et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science (New York, N.Y.)*, **300**, 1563–5.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. (H Tang, Ed,). *PLoS Genetics*, **8**, e1002967.

Pollak E (1987) On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics*, **117**, 353–60.

Pyhäjärvi T, Aalto E, Savolainen O (2012) Time scales of divergence and speciation among natural populations and subspecies of *Arabidopsis lyrata* (Brassicaceae). *American Journal of Botany*, **99**, 1314–22.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

R Core Team (2015) R: A language and environment for statistical computing.

Rawat V, Abdelsamad A, Pietzenuk B *et al.* (2015) Improving the annotation of *Arabidopsis lyrata* Using RNA-Seq Data. *PloS one*, **10**, e0137391.

Reed DH, Frankham R (2003) Correlation between fitness and genetic diversity. *Conservation Biology*, **17**, 230–237.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing indian population history. *Nature*, **461**, 489–94.

Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.

Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of polymorphism and demographic history in natural populations of Arabidopsis lyrata. *PLoS ONE*, **3**, e2411.

Roux C, Castric V, Pauwels M *et al.* (2011) Does speciation between *Arabidopsis halleri* and *Arabidopsis lyrata* coincide with major changes in a molecular target of adaptation? (PK Ingvarsson, Ed,). *PLoS ONE*, **6**, e26872.

Schmickl R, Jørgensen MH, Brysting AK, Koch MA (2010) The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evolutionary Biology*, **10**, 1–18.

Schoen DJ, Brown AH (1991) Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proceedings of the National Academy of Sciences*, **88**, 4494–4497.

Schönswetter P, Stehlik I, Holderegger R, Tribsch A (2005) Molecular evidence for glacial refugia of mountain plants in the European Alps. *Molecular Ecology*, **14**, 3547–3555.

Slatkin M (1981) Estimating levels of gene flow in natural populations. *Genetics*, **99**, 323–35.

Smit AFA, Hubley R, Green P (2010) RepeatMasker Open-3.0.

Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23.

Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular ecology*, **15**, 4261–93.

Stebbins GL (1957) Self Fertilization and Population Variability in the Higher Plants. *The American Naturalist*, **91**, 337–354.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Willi Y, Buskirk J Van, Hoffmann AA (2006) Limits to the adaptive potential of small populations. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 433–458.

Willi Y, Määttänen K (2010) Evolutionary dynamics of mating system shifts in *Arabidopsis lyrata*. *Journal of Evolutionary Biology*, **23**, 2123–31.

Willi Y, Määttänen K (2011) The relative importance of factors determining genetic drift: mating system, spatial genetic structure, habitat and census size in *Arabidopsis lyrata*. *The New Phytologist*, **189**, 1200–1209.

Wright S (1931) Evolution in mendelian populations. *Genetics*, **16**, 97–159.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular biology and evolution*, **22**, 506–19.

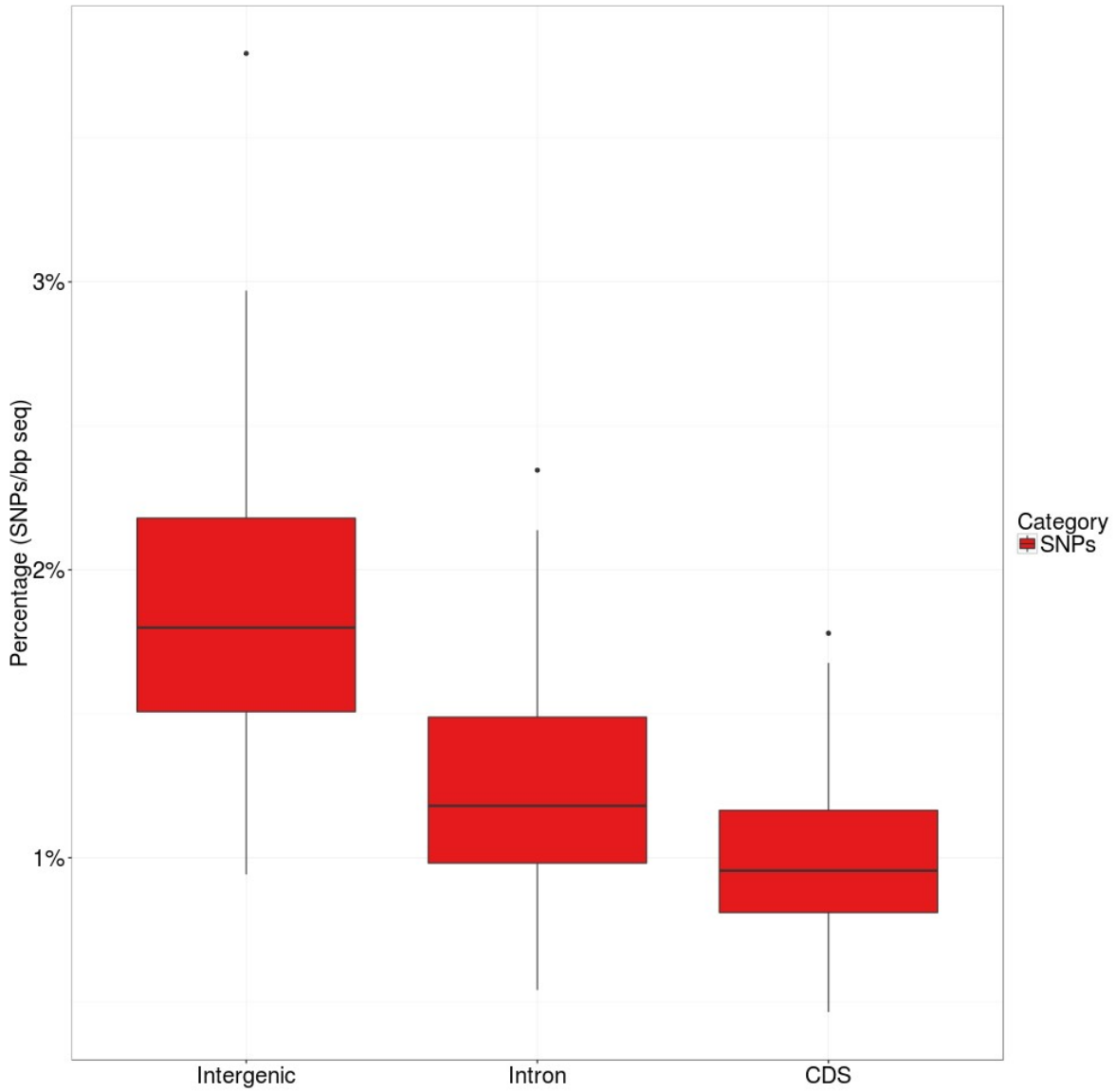


Figure 1. Boxplot of the percentage of SNPs relative to the base pairs sequenced for intergenic regions, introns and coding DNA sequences (CDS).

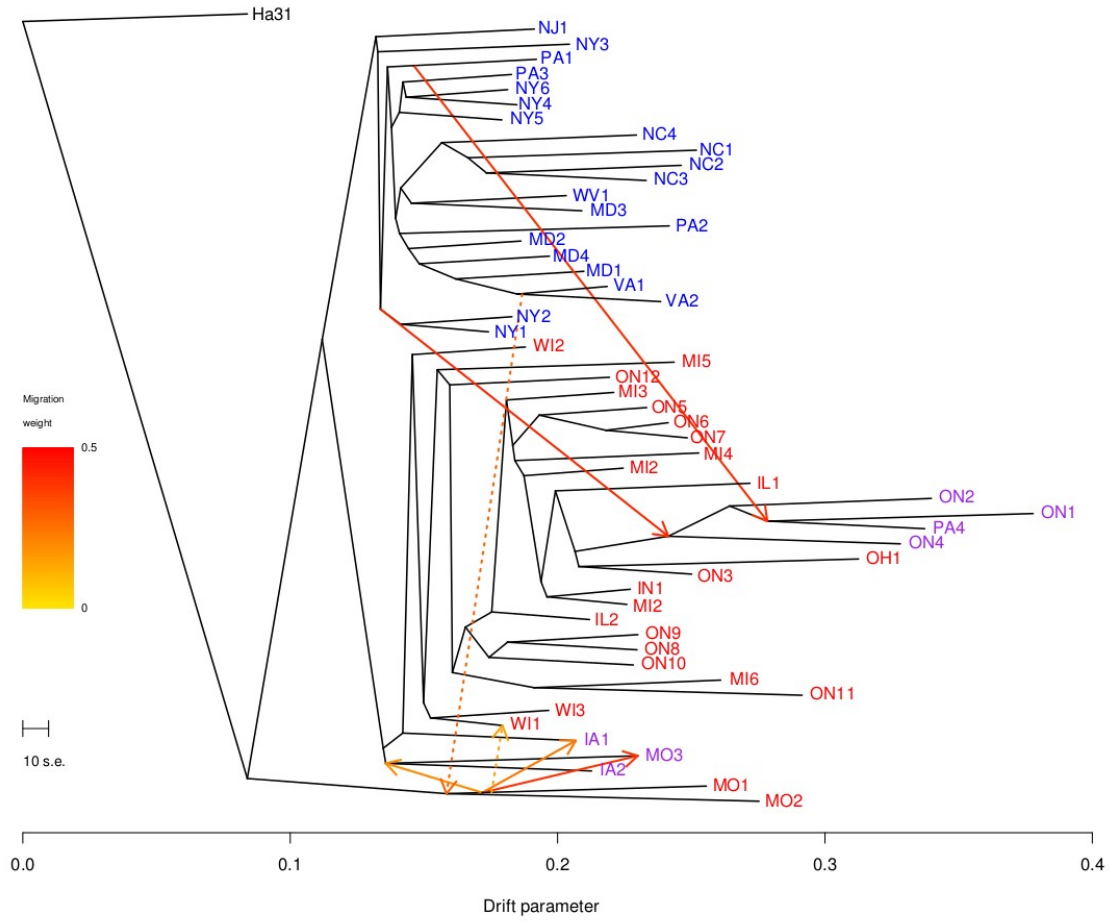


Figure 2. Relatedness tree of *A. lyrata* populations. The branch lengths are proportional to the genetic drift parameter of time over $2N_e$. The populations of the eastern cluster are indicated in blue, those of the western cluster in red and those with admixture in purple. The full arrows indicate migration events with support by the four population test for treeness. The dotted arrows indicate migration events without support by the four population test for treeness.

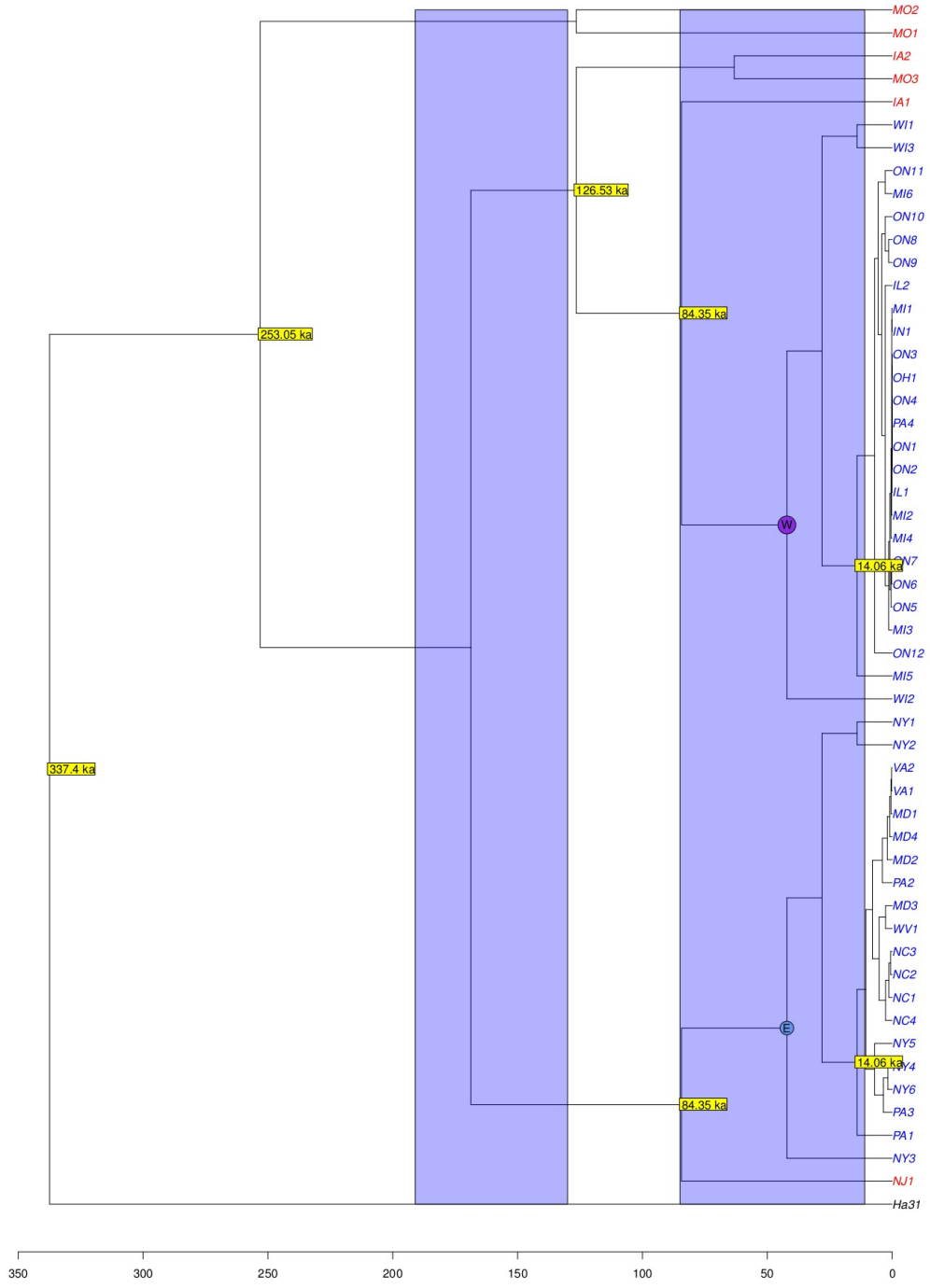


Figure 3. Time calibrated tree of *A. lyrata* populations. The populations that are north of the Laurentide ice sheet are indicated in blue. The blue rectangles show the duration of the Illinoian glaciation (191,000-130,000 years ago) and Wisconsin glaciation (85,000-11,000 years ago). In the yellow boxes is written the estimated divergence time. The red tips point to the populations that

survived the LGM. The blue tips point to the populations that emerged after the LGM.

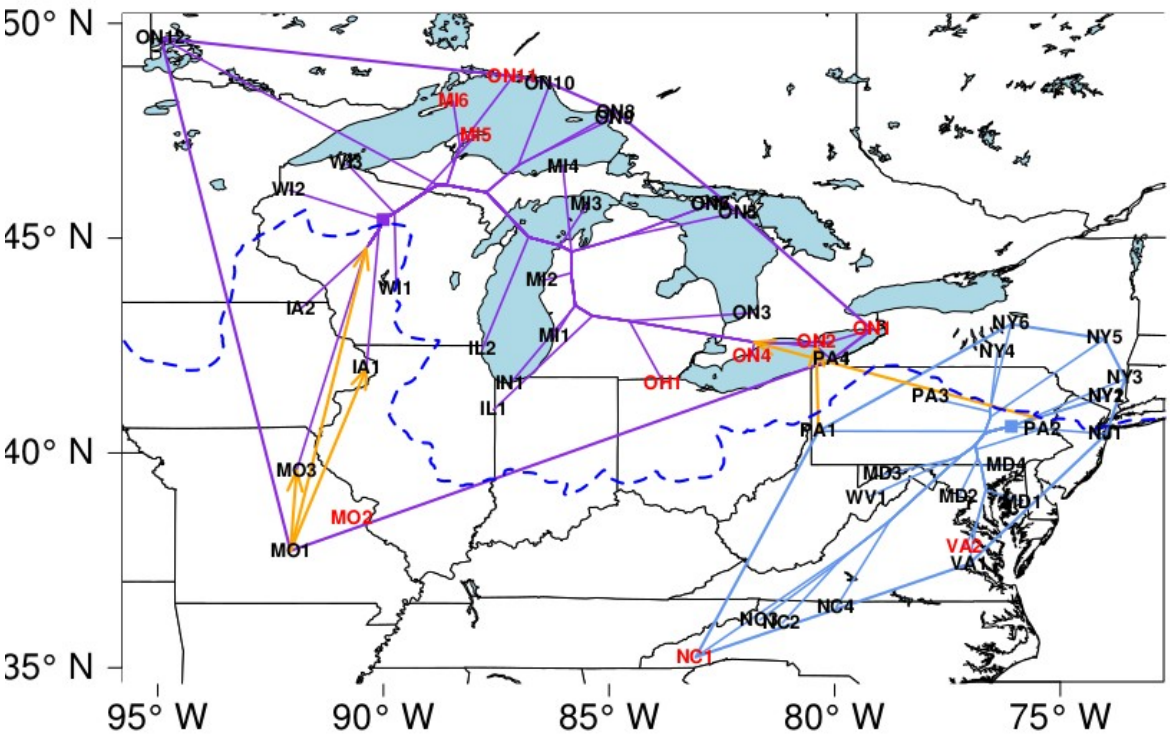


Figure 4. Map of the *A. lyrata* populations in North America. The blue dashed line shows the maximum extend of the ice at the LGM (Dike et al. 2003). The relatedness tree is plotted on the map, in blue for the eastern genetic cluster and in purple for the western genetic cluster. The position of the most common ancestor of populations that appeared after the LGM is represented with a squares. Outcrossing populations are indicated in black, the selfing populations in red.

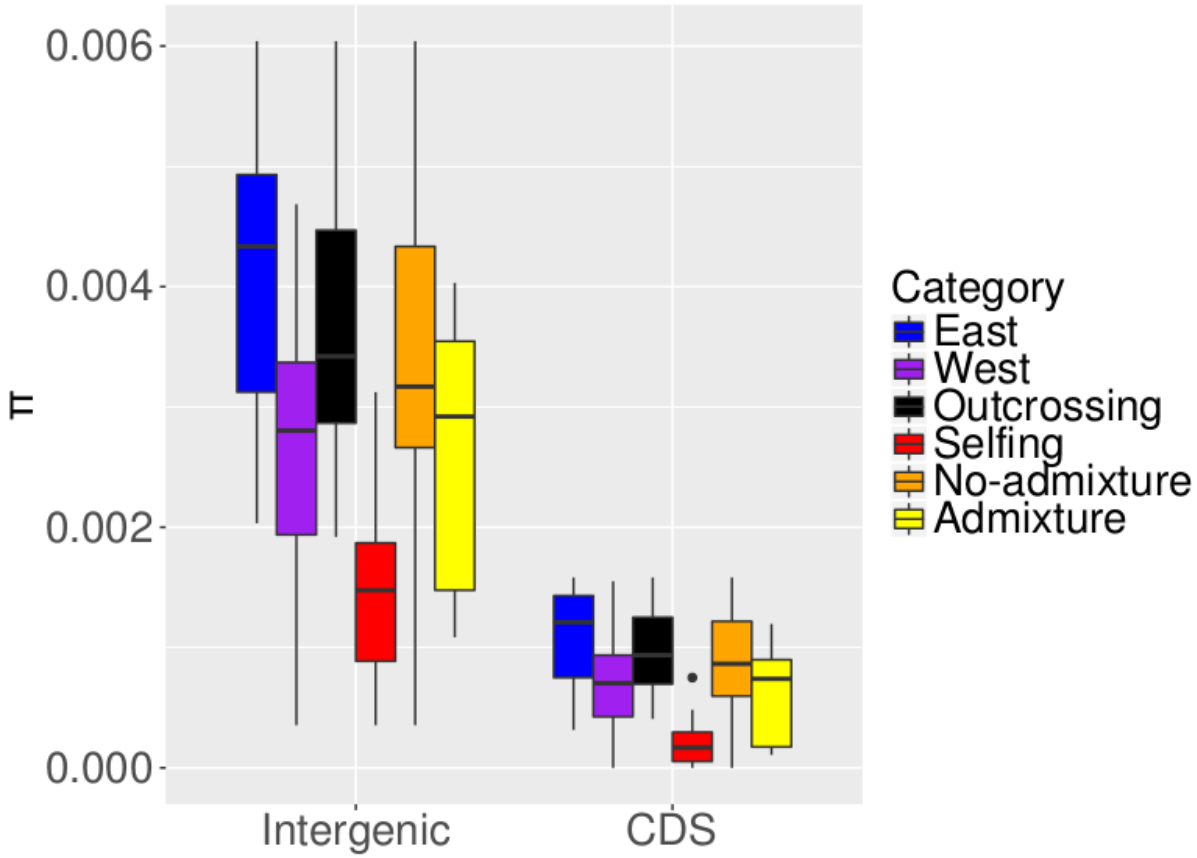


Figure 5. Boxplot illustrating the weighted median of π of the intergenic regions and coding DNA sequences (CDS). Populations are grouped by genetic cluster, mating system and signature of admixture. Results for the eastern cluster are indicated in blue, those of the western cluster in purple, for outcrossing in black and selfing populations in red, for population with no signature of admixture in orange and for those with a signature of admixture in yellow.

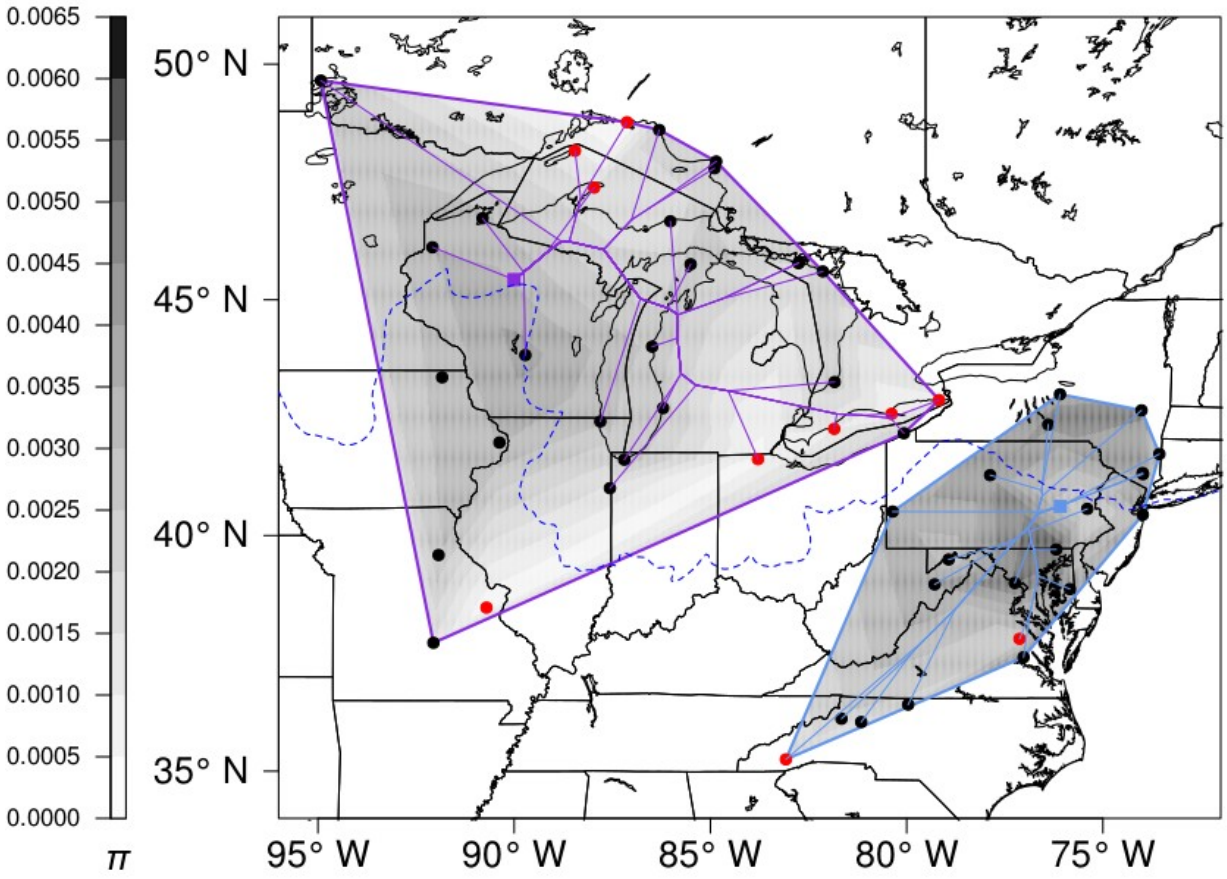


Figure 6. Map of interpolated nucleotide diversity π of intergenic regions. The blue dashed line shows the maximum extend of the ice at the LGM. The minimum convex polygon hull of the eastern cluster and the relatedness tree of its leading-edge populations is indicated in blue, those for the western cluster are indicated in purple. The position of the most common ancestor of populations that appeared after the LGM are represented with a square. The circles represent the populations assessed, in black the outcrossing populations and in red the selfing populations.

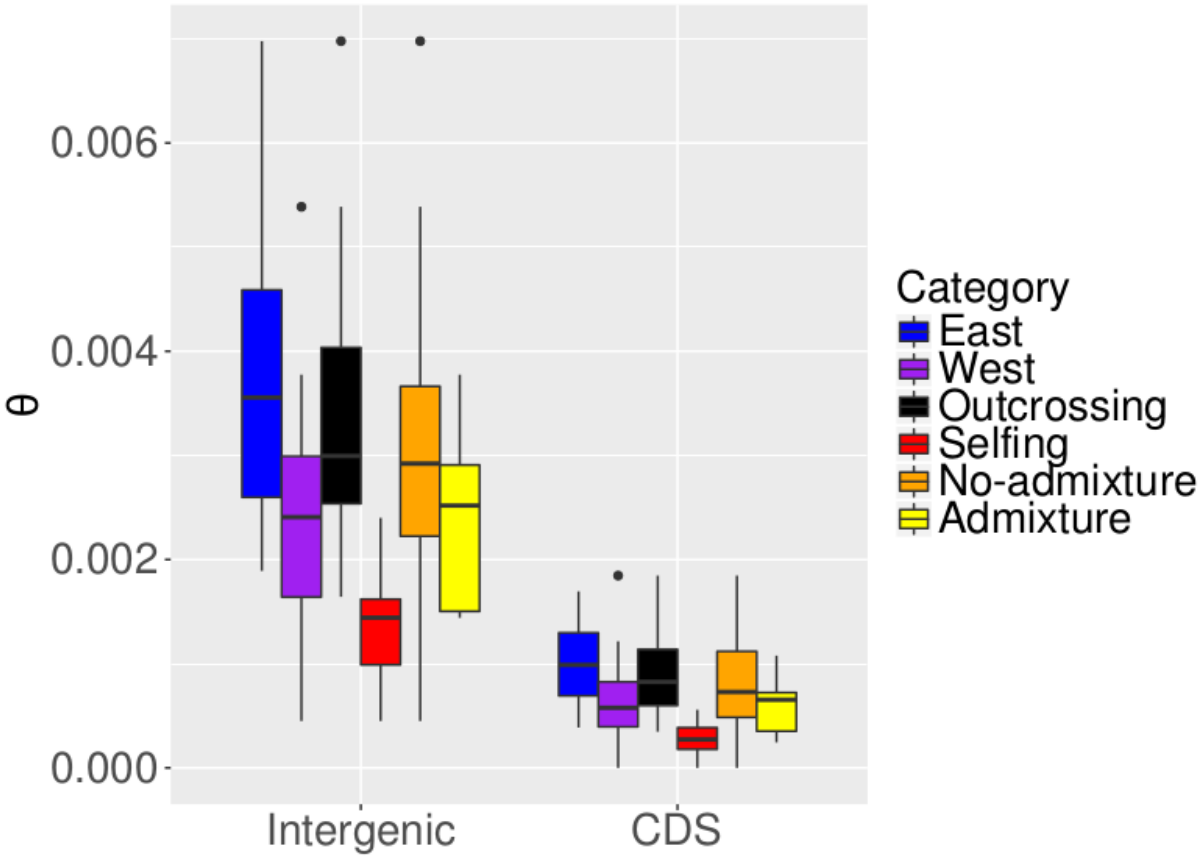


Figure 7. Boxplot illustrating the weighted median of Watterson's θ of the intergenic regions and coding DNA sequences (CDS). Populations are grouped by genetic cluster, mating system and signature of admixture. Results for the eastern cluster are indicated in blue, those of the western cluster in purple, for outcrossing in black and selfing populations in red, for population with no signature of admixture in orange and for those with a signature of admixture in yellow.

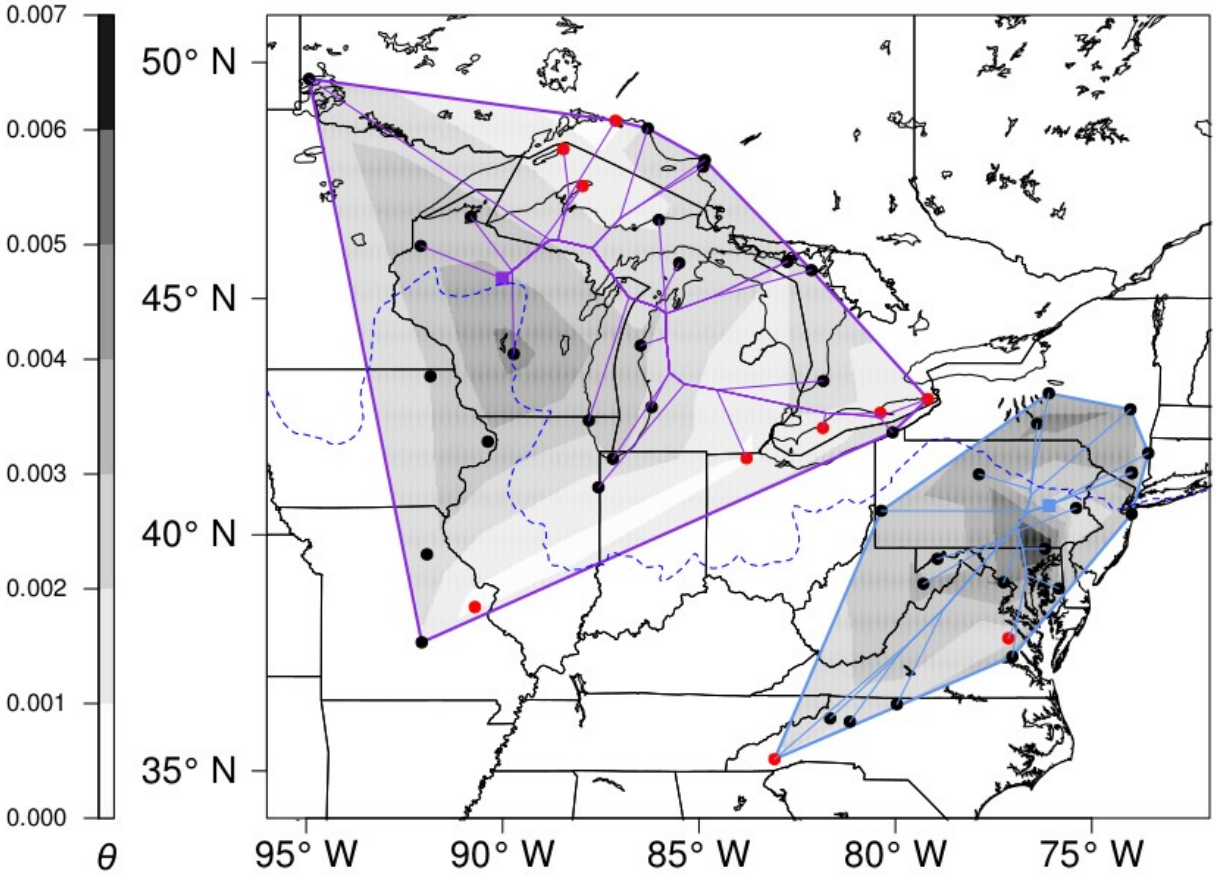


Figure 8. Map of interpolated Watterson's θ of intergenic regions. The blue dashed line shows the maximum extend of the ice at the LGM. The minimum convex polygon hull of the eastern cluster and the relatedness tree of its leading-edge populations is indicated in blue, those for the western cluster are indicated in purple. The position of the most common ancestor of populations that appeared after the LGM are represented with a square. The circles represent the populations assessed, in black the outcrossing populations and in red the selfing populations.

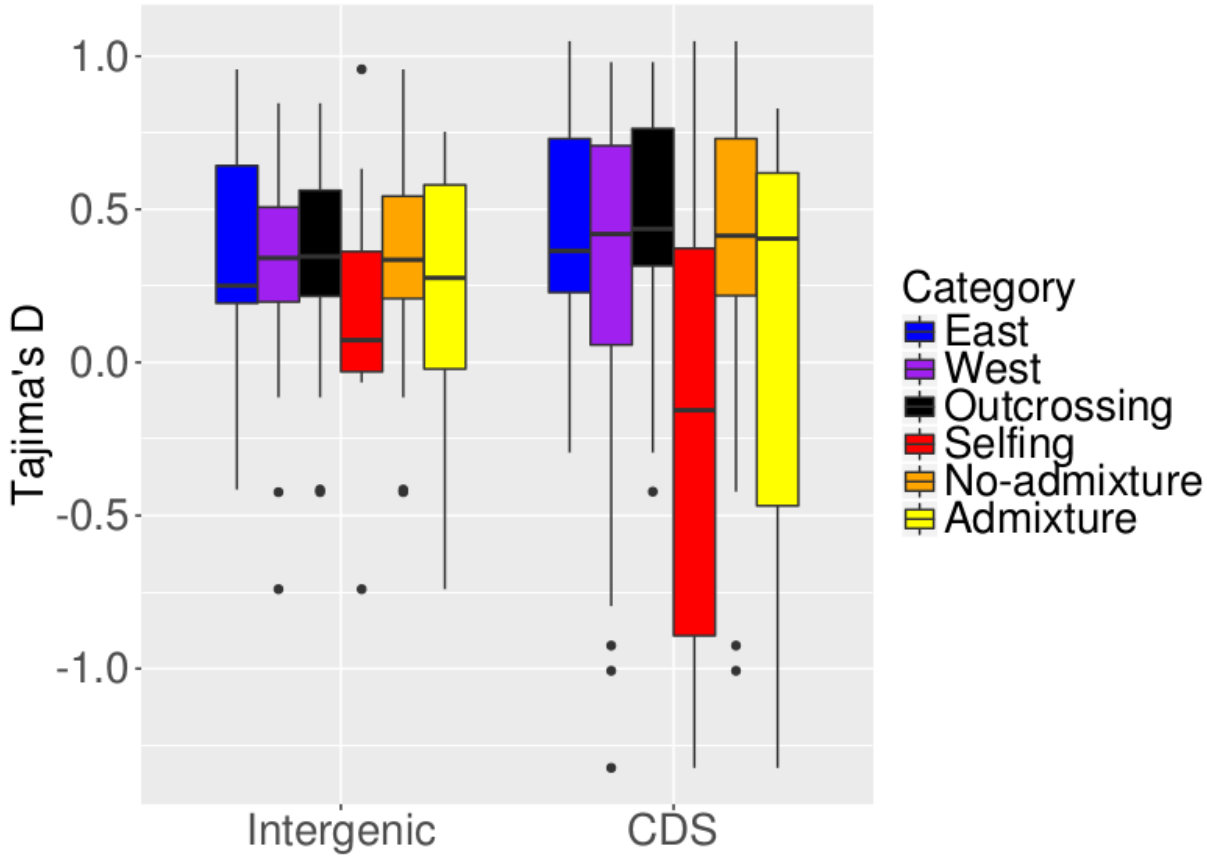


Figure 9. Boxplot illustrating the weighted median of Tajima's D of the intergenic regions and coding DNA sequences (CDS). Populations are grouped by genetic cluster, mating system and signature of admixture. Results for the eastern cluster are indicated in blue, those of the western cluster in purple, for outcrossing in black and selfing populations in red, for population with no signature of admixture in orange and for those with a signature of admixture in yellow.

Table 1. Pearson correlation coefficients for pairs of genetic diversity estimates based on microsatellites (expected heterozygosity, He, and allelic richness, A) and single nucleotide polymorphisms for intergenic regions, introns and coding regions, CDS (nucleotide diversity, π depicting heterozygosity, and Watterson's θ, depicting SNP richness). P-values were all <0.001.

	π, interg.	π, introns	π, CDS	A	θ, interg.	θ, introns	θ, CDS
Microsatellite He	0.877	0.924	0.922	0.919	0.85	0.891	0.876
π, intergenic		0.981	0.981	0.842	0.969	0.955	0.945
π, introns			0.999	0.885	0.948	0.969	0.956
π, CDS				0.886	0.951	0.972	0.961
Microstellite A					0.893	0.933	0.927
θ, intergenic						0.979	0.978
θ, introns							0.997

Table 2. Linear model testing the relation of ancestral cluster membership, distance from cluster core, census size and mating system on genomic diversity: nucleotide diversity π, allelic diversity θ and Tajima's D of intergenic regions and CDS regions. Sample size for all models was 52 populations. Most models explained a large amount of variation, with significant p-values <0.0001: π, intergenic: $R^2 = 0.76$; π, CDS: $R^2 = 0.73$; θ, intergenic: $R^2 = 0.62$; θ, CDS: $R^2 = 0.59$. For Tajima's D, variation explained was little and not significant: D, intergenic: $R^2 = 0.13$; D, CDS: $R^2 = 0.30$. Output statistics reported are: t-values and variation explained (var, in percent). Significance is indicated: *, $0.01 < P \leq 0.05$; **, $0.001 < P \leq 0.01$; ***, $P \leq 0.001$.

		Intergenic regions			**CDS regions**	
Diversity	**Source**	**t**	**var. expl.**		**t**	**var. expl.**
π	Ancestral cluster (west)	-3.81***	17.30		-1.86	9.51
π	ancestor distance	-4.77***	25.90		-5.16***	28.80
π	log10(census size)	0.82	0.93		0.75	0.90
π	Mating system (selfing)	-6.15***	29.80		-6.02***	32.00
π	Admixture (yes)	2.48*	2.10		1.77	1.86
θ	Ancestral cluster (west)	-2.76**	14.20		-1.69	8.92
θ	ancestor distance	-3.62***	22.10		-3.78***	23.10
θ	log10(census size)	1.05	0.99		0.97	0.97
θ	Mating system (selfing)	-4.19***	23.10		-4.17***	24.30
θ	Admixture (yes)	1.76	1.72		1.52	1.55
Tajima's D	Ancestral cluster (west)	0.6	0.45		0.95	0.96
Tajima's D	ancestor distance	-0.79	3.38		-1.78	9.40
Tajima's D	log10(census size)	-1.36	3.55		-0.61	0.75
Tajima's D	Mating system (selfing)	-1.51	4.83		-2.96**	17.10
Tajima's D	Admixture (yes)	-0.38	0.90		-0.41	1.85

Table S1. Median of genetic diversity estimates weighted based on the number of base pairs sequenced for each windows of 5000 bp. The first column lists the population identity, the following columns report nucleotide diversity, π, Watterson's θ, and Tajima's D of intergenic regions, introns and coding DNA sequences (CDS).

Pop ID	inter π	inter θ	inter D	intro π	intro θ	intro D	CDS π	CDS θ	CDS D
IA1	0.0040	0.0038	0.2761	0.0016	0.0014	0.4463	0.0012	0.0011	0.4043
IA2	0.0036	0.003	0.6788	0.0013	0.001	0.821	0.001	0.0008	0.7746
IL1	0.0019	0.0016	0.5681	0.0006	0.0005	0.8625	0.0004	0.0003	0.8293
IL2	0.0036	0.0033	0.2842	0.0014	0.0012	0.4971	0.0011	0.001	0.4057
IN1	0.0034	0.0031	0.2819	0.0013	0.0011	0.5595	0.001	0.0008	0.4644
MD1	0.0043	0.0035	0.7942	0.0015	0.0012	0.9507	0.0011	0.0009	0.9111
MD2	0.0055	0.0052	0.2098	0.002	0.0018	0.4098	0.0016	0.0014	0.3536
MD3	0.0037	0.003	0.7531	0.0012	0.0009	0.9447	0.0009	0.0007	0.893
MD4	0.006	0.007	-0.4152	0.0021	0.0022	-0.1021	0.0016	0.0017	-0.1125
MI1	0.0032	0.0029	0.3613	0.0012	0.0011	0.5319	0.0009	0.0008	0.453
MI2	0.0034	0.0032	0.2272	0.0012	0.0011	0.4603	0.0009	0.0008	0.3781
MI3	0.0032	0.003	0.3096	0.0012	0.0011	0.5048	0.0009	0.0008	0.4136
MI4	0.0022	0.0019	0.4726	0.0007	0.0006	0.7086	0.0005	0.0004	0.6414
MI5	0.002	0.0016	0.6323	0.0006	0.0005	0.9256	0.0005	0.0004	0.8871
MI6	0.0015	0.0014	0.3713	0.0003	0.0003	0.4617	0.0002	0.0002	0.4457
MO1	0.002	0.0018	0.1683	0.0005	0.0005	0.0778	0.0004	0.0005	-0.1664
MO2	0.0006	0.0006	0.3351	0	0	0.1418	0	0	0.1531
MO3	0.0035	0.0028	0.7531	0.0011	0.0009	0.8168	0.0008	0.0007	0.8294
NC1	0.002	0.0019	0.1502	0.0004	0.0004	-0.0546	0.0003	0.0004	-0.1719
NC2	0.0022	0.002	0.2504	0.0005	0.0005	0.1331	0.0004	0.0005	-0.0568
NC3	0.0029	0.0026	0.3861	0.0008	0.0007	0.4803	0.0006	0.0006	0.3643
NC4	0.0029	0.0027	0.1473	0.0008	0.0009	-0.0552	0.0007	0.0008	-0.2943
NJ1	0.004	0.0033	0.8001	0.0015	0.0012	1.0074	0.0012	0.0009	0.9623
NY1	0.0051	0.0048	0.2161	0.002	0.0018	0.3839	0.0015	0.0014	0.3341
NY2	0.0047	0.0041	0.4786	0.0017	0.0015	0.6224	0.0013	0.0011	0.5657
NY3	0.0032	0.0026	0.7048	0.0011	0.0009	0.8692	0.0009	0.0007	0.837
NY4	0.0051	0.0052	-0.0778	0.0018	0.0018	0.0771	0.0014	0.0015	-0.1109

NY5	0.0051	0.0049	0.108	0.0018	0.0017	0.3376	0.0014	0.0013	0.2553
NY6	0.0049	0.0045	0.3103	0.0019	0.0016	0.5137	0.0014	0.0013	0.4528
OH1	0.0008	0.0009	-0.0655	0	0.0001	-0.6456	0	0.0002	-1.0069
ON1	0.0014	0.0014	-0.0051	0.0001	0.0002	-0.4757	0.0001	0.0002	-0.7956
ON2	0.0011	0.0014	-0.7405	0.0001	0.0004	-1.1361	0.0001	0.0004	-1.3239
ON3	0.0028	0.0029	-0.1142	0.0009	0.0009	0.0816	0.0007	0.0007	-0.0388
ON4	0.0015	0.0016	-0.0373	0.0003	0.0004	0.0133	0.0002	0.0003	-0.1408
ON5	0.003	0.0027	0.3511	0.001	0.0008	0.5624	0.0008	0.0007	0.3695
ON6	0.0027	0.0024	0.3672	0.0009	0.0008	0.5574	0.0007	0.0006	0.5125
ON7	0.0023	0.002	0.5333	0.0008	0.0006	0.8089	0.0006	0.0005	0.7802
ON8	0.0028	0.0023	0.609	0.0009	0.0007	0.9597	0.0007	0.0005	0.8921
ON9	0.0028	0.0022	0.8465	0.0009	0.0007	1.0446	0.0007	0.0005	0.9805
ON10	0.0027	0.0022	0.7327	0.0009	0.0007	0.9621	0.0007	0.0006	0.9003
ON11	0.0004	0.0004	-0.0097	0	0	-0.9131	0	0	-0.9247
ON12	0.0029	0.0026	0.388	0.0011	0.0009	0.6053	0.0008	0.0007	0.5082
PA1	0.0042	0.0036	0.6424	0.0016	0.0013	0.7754	0.0012	0.001	0.7305
PA2	0.0029	0.0024	0.5428	0.0007	0.0006	0.5906	0.0005	0.0004	0.542
PA3	0.0049	0.0046	0.2169	0.0019	0.0017	0.4114	0.0014	0.0013	0.3856
PA4	0.0029	0.0025	0.4806	0.0009	0.0008	0.6017	0.0007	0.0007	0.4627
VA1	0.0049	0.0045	0.1933	0.0016	0.0015	0.3071	0.0013	0.0012	0.2284
VA2	0.0031	0.0024	0.9572	0.001	0.0007	1.0919	0.0008	0.0006	1.0492
WI1	0.0047	0.0054	-0.4237	0.002	0.0023	-0.3877	0.0016	0.0018	-0.422
WI2	0.004	0.0037	0.3409	0.0016	0.0015	0.3148	0.0013	0.0012	0.2181
WI3	0.0036	0.0034	0.2916	0.0014	0.0013	0.4773	0.0011	0.001	0.4192
WV1	0.0045	0.0042	0.2086	0.0015	0.0014	0.4197	0.0012	0.0011	0.309

Table S2. List of population information: population identity, membership to ancestral genetic cluster, the geographic distance from the core (ancestor that gave raise to populations emerging after the withdrawal fo the ice of the last glacial cycle; for a detailed description see Methods), the predominant mating system, and evidence for admixture.

Pop ID	Cluster	Distance [km]	log Census size	Mating system	Admixture
IA1	west	410.43	2.16	outcrossing	yes
IA2	west	305.07	2	outcrossing	yes
IL1	west	542.48	1.9	outcrossing	no
IL2	west	387.7	6.79	outcrossing	no
IN1	west	490.35	5.17	outcrossing	no
MD1	east	195.29	2.01	outcrossing	no
MD2	east	202.12	2.91	outcrossing	no
MD3	east	267.48	2.08	outcrossing	no
MD4	east	100.24	3.62	outcrossing	no
MI1	west	428.57	5.2	outcrossing	no
MI2	west	311.41	7.05	outcrossing	no
MI3	west	326.52	4.61	outcrossing	no
MI4	west	306.03	2.85	outcrossing	no
MI5	west	235.28	3.06	selfing	no
MI6	west	296.75	3.02	selfing	no
MO1	west	900.14	3.26	outcrossing	no
MO2	west	799.91	2.08	selfing	no
MO3	west	696.2	3.58	outcrossing	yes
NC1	east	850.33	2.56	selfing	no
NC2	east	669.19	4.37	outcrossing	no
NC3	east	692.79	3.21	outcrossing	no
NC4	east	571.8	3.77	outcrossing	no
NJ1	east	184.53	2.08	outcrossing	no
NY1	east	198.55	2.62	outcrossing	no
NY2	east	198.97	2.26	outcrossing	no
NY3	east	250.56	5.57	outcrossing	no
NY4	east	195.07	2.83	outcrossing	no
NY5	east	288.93	2.21	outcrossing	no

NY6	east	265.52	2.48	outcrossing	no
OH1	west	652.19	4.23	selfing	no
ON1	west	893.3	1.78	selfing	yes
ON2	west	817.98	3.52	selfing	yes
ON3	west	678.5	5.92	outcrossing	no
ON4	west	732.3	6.26	selfing	yes
ON5	west	589.13	4.1	outcrossing	no
ON6	west	540.58	2.24	outcrossing	no
ON7	west	540.46	2.15	outcrossing	no
ON8	west	449.35	2.88	outcrossing	no
ON9	west	438.16	2.68	outcrossing	no
ON10	west	418.51	4.68	outcrossing	no
ON11	west	399.08	1.49	selfing	no
ON12	west	594.8	3.11	outcrossing	no
PA1	east	356.29	2.07	outcrossing	no
PA2	east	63.99	3.88	outcrossing	no
PA3	east	163.16	3	outcrossing	no
PA4	west	862.7	5.45	outcrossing	yes
VA1	east	361.77	2.78	outcrossing	no
VA2	east	321.6	1.79	selfing	no
WI1	west	200.62	3.35	outcrossing	no
WI2	west	193.04	1.73	outcrossing	no
WI3	west	149.2	3.8	outcrossing	no
WV1	east	324.68	2.76	outcrossing	no

Table S3. Test statistics of the four-population test for treeness. Clade 1 was [NJ1,PA1], clade 2 was[ON3,X], where X is one western population formerly covered by the Laurentide ice sheet.

Populations	F4 statistic	Standard error	Z score	P
NJ1,PA1;ON3,PA4	0.0087	0.00133	6.52	6.90E-011
NJ1,PA1;ON3,ON2	0.00659	0.00139	4.73	2.29E-006
NJ1,PA1;ON3,ON1	0.00655	0.00139	4.7	2.61E-006
NJ1,PA1;ON3,ON4	0.00434	0.00115	3.77	0.000166
NJ1,PA1;ON3,IL1	0.000974	0.000978	0.996	0.319
NJ1,PA1;ON3,OH1	0.000794	0.001	0.791	0.429
NJ1,PA1;ON3,WI2	0.000147	0.000917	0.161	0.872
NJ1,PA1;ON3,IL2	-0.000284	0.00112	-0.253	0.8
NJ1,PA1;ON3,IN1	-0.000383	0.00083	-0.462	0.644
NJ1,PA1;ON3,MI2	-0.000453	0.000845	-0.537	0.592
NJ1,PA1;ON3,MI1	-0.000589	0.000873	-0.675	0.5
NJ1,PA1;ON3,WI3	-0.00102	0.001	-1.02	0.306
NJ1,PA1;ON3,MI3	-0.0011	0.000875	-1.25	0.21
NJ1,PA1;ON3,ON5	-0.00188	0.000932	-2.02	0.0437
NJ1,PA1;ON3,ON8	-0.00229	0.00112	-2.04	0.0413
NJ1,PA1;ON3,MI5	-0.00239	0.00114	-2.1	0.0361
NJ1,PA1;ON3,ON7	-0.00233	0.0011	-2.12	0.0344
NJ1,PA1;ON3,ON12	-0.0024	0.000981	-2.45	0.0143
NJ1,PA1;ON3,ON6	-0.00261	0.001	-2.6	0.00927
NJ1,PA1;ON3,MI4	-0.00252	0.000945	-2.66	0.00777
NJ1,PA1;ON3,ON9	-0.00346	0.00102	-3.41	0.000647
NJ1,PA1;ON3,MI6	-0.00545	0.00143	-3.82	0.000134
NJ1,PA1;ON3,ON10	-0.00381	0.000925	-4.12	3.87E-005
NJ1,PA1;ON3,ON11	-0.00609	0.00139	-4.37	1.23E-005

Table S4. Test statistics of the four-population test for treeness. Clade 1 was [MO1,MO2], clade 2 was [WI3,X], where X is one western population.

Populations	F4 statistic	Standard error	Z score	P
MO2,MO1;WI3,MO3	0.00595	0.00175	3.4	0.000663
MO2,MO1;WI3,IA1	0.00295	0.00133	2.22	0.0262
MO2,MO1;WI3,IA2	0.00308	0.00149	2.07	0.038
MO2,MO1;WI3,PA4	0.00108	0.00155	0.692	0.489
MO2,MO1;WI3,WI2	0.000523	0.000954	0.548	0.583
MO2,MO1;WI3,MI5	-0.000426	0.00133	-0.321	0.748
MO2,MO1;WI3,WI1	-0.000552	0.00094	-0.587	0.557
MO2,MO1;WI3,ON2	-0.00112	0.00165	-0.679	0.497
MO2,MO1;WI3,ON1	-0.00148	0.00168	-0.882	0.378
MO2,MO1;WI3,ON12	-0.00279	0.00143	-1.96	0.0502
MO2,MO1;WI3,ON4	-0.00383	0.00153	-2.5	0.0123
MO2,MO1;WI3,ON8	-0.00441	0.00127	-3.47	0.000523
MO2,MO1;WI3,ON10	-0.00444	0.00119	-3.73	0.000195
MO2,MO1;WI3,ON9	-0.0048	0.00126	-3.82	1.36E-004
MO2,MO1;WI3,ON11	-0.00628	0.00157	-4	6.35E-005
MO2,MO1;WI3,MI3	-0.00529	0.00129	-4.1	4.21E-005
MO2,MO1;WI3,MI6	-0.00603	0.00142	-4.24	2.23E-005
MO2,MO1;WI3,MI4	-0.00739	0.00141	-5.22	1.74E-007
MO2,MO1;WI3,IL2	-0.00619	0.00113	-5.46	4.85E-008
MO2,MO1;WI3,ON5	-0.00742	0.00128	-5.79	7.10E-009
MO2,MO1;WI3,ON7	-0.00897	0.0015	-6	2.02E-009
MO2,MO1;WI3,ON6	-0.0083	0.00138	-6.01	1.91E-009
MO2,MO1;WI3,MI2	-0.00835	0.00116	-7.23	4.98E-013
MO2,MO1;WI3,IN1	-0.00893	0.00119	-7.53	4.98E-014
MO2,MO1;WI3,OH1	-0.0131	0.00173	-7.54	4.56E-014
MO2,MO1;WI3,ON3	-0.0104	0.00134	-7.74	9.69E-015
MO2,MO1;WI3,IL1	-0.0107	0.00131	-8.15	3.52E-016
MO2,MO1;WI3,MI1	-0.00989	0.00116	-8.53	1.50E-017

CHAPTER 3: Environmental marginality and geographic range limits in *Arabidopsis lyrata* spp. *lyrata*

Julie A. Lee-Yaw[1, 2, 3], **Marco Fracassetti**[1,2] **and Yvonne Willi**[1,2]

[1] *Institute of Biology, University of Neuchâtel, 2000 Neuchâtel, Switzerland*

[2] *Department of Environmental Sciences, University of Basel, 4056 Basel, Switzerland*

[3] *Current Address: Biodiversity Research Centre and Department of Botany, University of British Columbia, Vancouver, Canada*

Running title: Environmental marginality and range limits

Keywords: adaptation; ecological niche model; genetic variation; genomic diversity; habitat suitability; niche limit; species distribution model

SUBMITTED TO ECOGRAPHY

Abstract

Understanding the factors that govern the distribution of species is a central goal in evolutionary ecology. It is commonly assumed that geographic range limits reflect ecological niche limits and that species experience increasingly marginal conditions towards the edge of their ranges. Using spatial data and ecological niche models we tested these hypotheses in *Arabidopsis lyrata*. Specifically, we asked whether range limits coincide with predicted niche limits in this system and whether the suitability of sites declines towards the edge of the species' range in North America. We further explored patterns of environmental change towards the edge of the range and asked whether genome-wide patterns of genetic diversity decline with increasing peripherality and environmental marginality. Our results suggest that latitudinal range limits coincide with niche limits. Populations experienced increasingly marginal environments towards these limits—though patterns of environmental change were more complex than most theoretical models for range limits assume. Genomic diversity declined towards the edge of the species' range and with increasing distance from the estimated centre of the species' niche in environmental space, but not with the suitability of sites based on niche model predictions. Thus while latitudinal range limits in this system are broadly associated with niche limits, the link between environmental conditions and genetic diversity, and thus the adaptive potential of populations, is less clear.

Introduction

Understanding species' geographic ranges is a major goal in ecology. There are two main ways in which the distribution of environmental conditions is thought to shape species' ranges. The first is based on Hutchison's "n-dimensional hypervolume" niche concept (Hutchinson 1957), which describes the region of multivariate environmental space in which conditions permit population persistence (i.e. non-negative growth rates; Hutchinson 1978; Holt 2009). When transferred across geographic space, the niche defined in this way delineates the potential geographic range of a species in a binary fashion (e.g. Jackson & Overpeck 2000). Not mutually exclusive to this concept is the model championed by Brown (1984) in which environmental conditions deteriorate towards the edge of a species' range. This scenario sets up a gradient within species' ranges (i.e. within the geographical realization of Hutchinson's niche: Brown 1984), whereby populations become smaller and increasingly isolated towards the edge of the range (sometimes referred to the "abundant-centre hypothesis"; though environmental factors may not always drive such a pattern where it is observed: Sagarin & Gaines 2002; Gaston 2003). These premises have greatly influenced both ecological and evolutionary perspectives on range limits (e.g. Table 1 from Sagarin & Gaines 2002; Sexton *et al.* 2009) and empirical evaluation of these models is thus of fundamental interest. In this study, we test the relationship between environmental marginality and range limits in *Arabidopsis lyrata* sub. *lyrata* (hereafter *A. lyrata* for simplicity).

Hutchinson's niche concept gives us an initial framework with which to think about species' ranges, centered around the question of whether species' fill their potential niche on the landscape. Additional factors, namely biotic interactions and constraints on dispersal may preclude species' from occupying all regions of geographic space that are suitable for them and a biotic-abiotic-migration (i.e. BAM) classification scheme is commonly used in ecology to think about range limits (e.g. Soberón 2007; Peterson *et al.* 2011). From an evolutionary standpoint, assessing the relative importance of these different types of factors for range limits provides critical insight into the types

of traits that may be limiting the range and whether range expansion requires subsequent adaptation to abiotic conditions (i.e. "niche expansion").

Brown's model of increasing environmental marginality towards the edge of the range also has implications for understanding the extent to which constraints on adaptation govern species' range limits. First, this model assumes that there is a single set of optimal conditions in which population growth and/or carrying capacity is highest. Observing a strong association between population size and environmental gradients in accordance with this model suggests that local adaptation has not allowed this optimum to vary from place to place (Gaston 2003) and thus speaks to the readiness with which adaptation has occurred within the range. Second, an extension of Brown's model is that marginal conditions, in keeping populations small, may actually constrain adaptation. For instance, small (effective) population size decreases the efficacy of selection relative to drift (Kimura 1983; Charlesworth 2009), leads to inbreeding and potentially inbreeding depression (reviewed by Keller & Waller 2002), and may establish density gradients that promote swamping gene flow from elsewhere in the range (e.g. Garcia-Ramos & Kirkpatrick 1997). These factors diminish the ability of populations to respond to selection and, in the context of range limits, may limit adaptation to novel conditions at the range edge (Soule 1973; Kawecki 2008).

A number of empirical studies have addressed the question of whether range limits are niche limits (see studies included in Hargreaves *et al.* 2014 and Lee-Yaw *et al.* 2016) or, in line with Brown's model, the extent to which population size (reviewed by Sagarin & Gaines 2002) or genetic diversity (Eckert *et al.* 2008; Lira-Noriega & Manthey 2014) decline towards the edge of the range. Yet these questions are rarely addressed simultaneously. Several alternative scenarios are possible and have different implications for understanding the relative importance of and nature of constraints on adaptation at the edge of the range. For instance, a range limit may be a niche limit in the absence of a gradual decline in habitat quality and population size from centre to edge if the environment changes abruptly (Gaston 2003). Although adaptation is still required to overcome

such a range limit, this scenario creates different expectations for the number and effect size of mutations necessary to permit range expansion, and for the effects of gene flow on adaptation, than a scenario whereby range limits fall along a smooth environmental gradient (e.g. Kawecki 2008; Gomulkiewicz et al. 2010; Holt & Barfield 2011). Species may also fail to demonstrate demographic patterns consistent with Brown's model in cases where there has been a history of asymmetric range expansion. For instance, in the northern hemisphere, post-glacial colonization from southern refugia, may mean that the spatial centers of many contemporary ranges are displaced from historical areas of high abundance and diversity (see also Micheletti & Storfer 2015). Colonization from multiple refugia in many cases may also limit the utility of models based on a single range centre. Finally, even where species conform to Brown's model, a limited amount of time for colonization in these areas may mean that many species have yet to colonize the full spatial extent of their Hutchison niche (i.e. a form of dispersal limitation). Thus historical events may override niche dynamics in structuring many species' ranges and negate one or both of the models for range limits discussed above.

The difficulty of empirically quantifying environmental suitability and population size with the amount of replication and across the scales necessary for understanding range limits has traditionally limited studies evaluating the relationship between suitable habitat and range limits. Here we used ecological niche modeling and a large genomic dataset to overcome some of these limitations and to address these questions in *A. lyrata* across its North American range. We specifically asked: 1) Are range limits niche limits in this system? 2) How do environmental conditions and habitat suitability change towards the edge of the range? 3) In line with an effect of environmental conditions on population size and genetic diversity, is there a relationship between environmental suitability and genome-wide patterns of genetic diversity or is genetic diversity explained by range position alone, particularly in light of recent range changes associated with the Pleistocene glaciations?

Methods

Study System

Arabidopsis lyrata sub. *lyrata* occurs in the eastern USA and southern Canada (Fig. 1a) and is one of the most southern members of the circumpolar *Arabidopsis lyrata* species complex (Schmickl *et al.* 2010). This short-lived perennial is associated with disturbed habitats—primarily sand dunes, deposits and rocky outcrops along lakes and riverbanks. Previous work suggests that the species consists of two genetic lineages (Willi & Määttänen 2010; Griffin & Willi 2014), likely reflecting repeated contraction into and expansion from distinct eastern and western refugia during the Pleistocene glaciations (Griffin & Willi 2014). Genomic data suggest divergence between lineages is low and that some amount of gene flow has occurred in the recent past (Fracassetti & Willi unpublished). Crosses between the two lineages also result in viable offspring with no evidence of reduced fitness (Willi unpublished). For these reasons, we treat North American populations of *A. lyrata* as a single species and ask about range limits and the niche of the species as a whole. At the same time, evidence for post-glacial range expansion from at least two refugia prompts us to consider this history in our analysis of the distribution of genome-wide diversity below.

Quantifying environmental suitability

We used ecological niche models to assess the distribution of suitable habitat for *A. lyrata* within and beyond its range in North America. Close to 600 localities were collected from published studies, herbaria and government agencies and were georeferenced to within 1 km (average estimated error: 0.25 km). Records along lakeshores or on small islands that resulted in missing data (i.e. ended up in water) at the resolution of our raster dataset (below) were moved a maximum distance of 5 km to their nearest "onshore" grid cell or were discarded. In order to reduce the potential effects of sampling bias on our niche models (Varela *et al.* 2014; Boria *et al.* 2014), we

thinned records using an approach that considered both the density of points (geographic filtering) and environmental novelty (environmental filtering; full details in Appendix S1). After thinning, our final locality dataset included 279 records (including 55 that had been moved a median distance of 1.28 km "onshore"; Fig. 1).

The environmental data in our models included eight climatic, topographical and vegetation variables at a resolution of 30 arc-seconds (Table 1; variable selection explained in Appendix S1). Niche models were built using MAXENT (Phillips *et al.* 2006; Phillips & Dudík 2008) and the *dismo* package (Hijmans *et al.* 2013) in R. The background area for model calibration included all ecotones (Commission for Environmental Cooperation Working Group 1997) occupied by the species with the maximum study extent set as -100 to -70 degrees longitude and 32 to 53 degrees latitude. Prior to model calibration, we used the approach described by Warren & Seifert (2011; see also Warren *et al.*, 2014) to tune both the features and regularization coefficient (ß; Appendix S1). Following feature and ß tuning, we used ten-fold cross-validation (Fielding & Bell 1997) to calibrate a final set of models, assessing the ability of each model to discriminate between withheld presence data and background data based on the area under the curve statistic (AUC; of the receiver operating characteristic). For each model, we also compared observed AUC values to a null distribution of values generated from 99 niche models built using random points (Raes & Steege 2007) and calculated the true test statistic (TSS; Allouche *et al.*, 2006) using the threshold that maximized the sum of sensitivity and specificity *(*Liu *et al.* 2013). Models that had an AUC score that was both ≥ 0.7 and that fell outside the 95th quantile of this null distribution (i.e. performed significantly better than random), and for which TSS > 0.5, were used to predict environmental suitability across the study region, with the mean of these predictions taken as an estimate of the suitability of each cell for the species (Peterson *et al.* 2011).

As a second metric of environmental suitability that relies less heavily on the niche models, we used the method of Lira-Noriega & Manthey (2014) to calculate the distance of sites to an

estimate of the centre of the species' niche in environmental space. Specifically, we converted continuous niche model predictions to a binary representation of the species' niche in geographic space using two cutoffs for defining suitable locations: the minimum suitability score of any *A. lyrata* locality (minPres; Fig. 1c) and the 5th percentile of locality suitability scores (PercentPres; Fig. 1d). For each of these distributions, we took a random sample of 5000 points and used these points, as well as the locality data, to conduct a principal component analysis (PCA) of the eight environmental variables used in the niche models. The species' niche centroid in each case was defined as the vector of median scores of all points used in the PCA for PC axes one to six (which captured > 99% of the variance in environment). The distance between any location and this niche centroid serves as a measure of the niche centrality (or conversely marginality) of that point. As a sensitivity test, we reran the niche centroid calculation for both thresholds with different numbers of random sites in the PCA (500, 1000, 2500).

Are range limits niche limits?

To determine whether range limits reflect the limits of suitable habitat for the species (i.e. niche limits), we asked whether observed range limits correspond to those predicted by the niche models. Apart from qualitative comparisons between model prediction surfaces and the species' range, we used a novel range-fitting test to evaluate the statistical significance of the fit between observed range limits and predicted niche limits. The steps of this test were as follows: 1) Continuous niche model predictions were converted to a binary map of habitat suitability representing the species' niche limits; 2) Depending on the range limit in question (e.g. northern limit, etc.) this binary raster was then subset to include only cells to the north (or south etc.) of the centroid of the MCP. 3) Treating all of the cells inside the MCP in this subsetted raster as "in-range" and all cells outside the MCP as "out-of-range", we calculated as the proportion of in-range cells predicted as suitable + proportion of out-of-range cells predicted as non-suitable -1 (TSS_{range}); 4) The species' MCP was

then shifted randomly in the direction of the range limit being evaluated (e.g. to the north etc.); 5) Based on the centroid of the shifted MCP, the original binary raster was subset anew and TSS_{range} was recalculated; 6) Steps 3 and 4 were repeated 99 times to generate a null distribution of TSS_{range} scores for each range limit. The fit between a given range limit and the niche limit of the species in the same direction was considered significant if the observed values of TSS_{range} were more extreme (indicating a better fit) than 95% of the values in the null distribution.

We conducted this test to evaluate the fit of the species' northern, southern and western range limits to predicted niche limits (the species' eastern limit coincides with the Atlantic ocean making tests of niche limits irrelevant to range limits in this direction). The MCP was repeatedly shifted in each direction by a random amount between 50 and 500 km (i.e. reasonable for our study region). We repeated this test using binary suitability maps based on two cutoffs for considering a site suitable: the minimum suitability score of any *A. lyrata* locality and the 5th percentile of locality suitability scores (Fig. 1c, d). We note that using a cutoff that maximizes the sum of sensitivity and specificity *(*Liu *et al.* 2013) produced a binary map that was almost identical to our second cutoff and so this third, commonly used cutoff was not considered further.

Environmental changes towards the edge of the range

We used the average predicted suitability of sites based on our niche models, as well as our index of niche centrality, to test the hypothesis that environmental conditions become more marginal (i.e. less suitable) towards the edge of the species' range. Brown's hypothesis specifically posits that conditions are most suitable at the geographic centre of species' ranges. Thus as an index of peripherality, we took the minimum great circle distance from each locality to the centroid of the MCP encompassing all localities. These values were standardized by dividing by the sum of this distance and the minimum distance between the locality and the MCP hull (following Griffin & Willi 2014). Localities were also assigned to non-overlapping groups according to their primary

direction (north, south, east, west) away from the MCP centroid. We then used linear models to test whether environmental suitability declines and distance to the niche centroid increases with increasing peripherality. Direction was included as a second, potentially interacting, explanatory variable. Peripherality scores were arcsin transformed and distance to the niche centroid was log transformed to better meet model assumptions. We repeated the test for the estimates of distance to niche centroid based on different thresholds for defining the extent of suitable habitat and based on the PCAs involving different numbers of points to estimate the niche centroid (above). One of the sites towards the eastern edge of the species' range (a coastal population from Sandy Hook, NJ, USA) was an environmental outlier in both datasets. We ran the models with and without this point included.

To understand environmental patterns underlying changes in suitability at the species' range limits on a smaller scale, we used a model selection approach to evaluate the nature of the relationship between range position and the two variables that were most important for predicting presence in the niche models. Focusing on 200 km transects spanning the range edge in different places and centered on different peripheral populations, we specifically tested whether minimum temperature of early spring and precipitation of the wettest quarter (see results) change towards or across the range edge and whether changes in these variables tend to be gradual or abrupt. Fourteen sites were chosen to center each transect, with transects running from these points 100 km towards and 100 km away from interior parts of the range (Fig. 1a; see Appendix S1 for full details of transect designation). We extracted environmental values at 5 km intervals along each transect and used model selection based on AIC to determine whether an intercept-only (no change), linear (environmental gradient) or four parameter logistic (abrupt change) model better describe the relationship between the two environmental variables and range position for each transect. Finally, because these simple model types may fail to account for more complex changes in the environment towards the range limit, we conducted a breakpoint analysis to determine the number of significant

structural changes in the relationship between each environmental variable and transect position. The optimal number of breakpoints per transect was estimated using the *strucchange* (Zeileis *et al.* 2002; Achim *et al.* 2003) package in R, with the minimum number of observations per segment set to five (permitting a maximum of seven breaks per transect).

Environmental marginality and patterns of genomic diversity

Small population size associated with increasingly marginal habitat is expected to result in a decrease in genetic diversity towards the edge of species' ranges (i.e. Eckert *et al.* 2008; Lira-Noriega & Manthey 2014). To test this hypothesis, we looked at the relationship between environmental suitability and genome-wide estimates of genetic diversity. Following the protocol and pipeline outlined by Fracassetti *et al.* (2015), we sequenced pools of 25 individuals from 42 outcrossing populations from across the species' range (Fig. 1a). After aligning reads to the published nuclear genome of *A. lyrata* (v.1.0; Hu *et al.* 2011) and filtering (see Fracassetti *et al.* 2015), single nucleotide polymorphisms (SNPs) were called for each population using Varscan (Koboldt *et al.* 2012). We retained biallelic SNPs with a minimum variant allele frequency of 0.015, a P-value lower than 0.15, a minimum mapping quality score of 20 and a minimum allele count of three reads. Our final genomic dataset consisted of a mean of > 1.6 million SNPs per population with an average depth coverage of 125X. Nucleotide diversity (π) was estimated in 5000 bp windows across the genome of each population using NPStat v. 0.99c (Ferretti *et al.* 2013). The weighted median value across windows was taken as an estimate of genomic diversity for each population.

We used Pearson correlation tests to examine the relationship between genomic diversity and environmental marginality. However, genome-wide patterns of genetic diversity are known to decline with increasing peripherality as the result founder effects during colonization (Griffin & Willi 2014). We thus used linear models to test for an effect of environmental marginality on

genetic diversity, above and beyond any effects associated with range position. In this case, range position was defined in terms of distance away from putative refugial areas based on the results of Griffin & Willi (2014). Specifically, Griffin & Willi (2014) described a pattern of declining microsatellite diversity away from the centroids of the western and eastern lineages that they described. These centroids roughly coincide with the driftless area of Wisconsin and the central Appalachian Mountains respectively (Fig. 1a), both of which have previously been proposed as refugial areas for a number of other taxa (Jackson *et al.* 2000; Soltis *et al.* 2006; Lee-Yaw *et al.* 2008; Beatty & Provan 2011; Li *et al.* 2013). We assigned sites to genetic group (based on a modified version of the boundaries from Griffin & Willi 2014: see Appendix S1) and used the formula above to calculate the peripherality of each site with respect to the centroid of its genetic group. We regressed nucleotide diversity against peripherality and then asked whether environmental marginality explained any of the residual variation from this relationship. Models were run separately for the two measures of environmental marginality (suitability and distance from niche centre). We note that the suitability outlier was removed for this analysis, as it was an extreme outlier that polarized the values of all other sites. All analyses were done in R v. 3.2.3.

Results

Are range limits niche limits?

The niche models from all ten rounds of model calibration in MAXENT had high predictive performance when applied to withheld data (AUC ranged from to 0.84 to 0.91; mean AUC = 0.88) and outperformed models built using random locations from within the species' range. Models also performed reasonably well based on the threshold-dependent evaluation metric, with TSS > 0.64 in all cases. Of the eight variables included in our models, average minimum temperatures during the early spring (March and April) made the largest percent contribution to the model and resulted in the largest drop in model performance when values were randomly permutated across the training

dataset (Table 1). The suitability of sites tended to be highest at intermediate to higher values of this variable and drop off towards very high or very low values within the study region (Fig. S2-1a). Precipitation of the wettest quarter and the Priestly-Taylor coefficient (representing moisture availability) were also of moderate importance in the latter regard (Table 1), with suitability tending to decrease with very high values of spring precipitation (Fig. S2-1b) and towards very high or low values of the Priestly-Taylor coefficient. Non-climate variables, associated with topography and vegetation, were the least important for explaining the occurrence of the species across the landscape (Table 1).

Examination of average model predictions across the study region revealed that range limits are closely aligned with the predicted distribution of suitable conditions for the species—although areas of moderate to high suitability extend beyond the species' range to the west and to the northeast along the St. Lawrence River (Fig. 1). Our range fitting tests generally supported these conclusions, though only speak to the fit of model predictions to range limits in three broad directions: north, south and west. Observed TSS_{range} was significantly higher than values calculated after randomly shifting range limits 50 to 500 km to the north, regardless of whether the minimum suitability score of any locality or the 5^{th} percentile of locality suitability scores was used as a cutoff for defining suitable habitat ($p = 0.01$ in both cases). In the southward direction, TSS_{range} was significantly higher than values calculated for shifted range limits when the 5^{th} percentile of locality suitability scores was used as a cutoff for defining suitable habitat ($p = 0.01$) but not when the minimum suitability score of any locality was used ($p = 0.33$). In the westward direction, observed values were non-significant regardless of the cutoff used, supporting the qualitative observation that range limits are not well-predicted by niche limits in this direction.

Environmental changes towards the edge of the range

Patterns of predicted suitability suggested that conditions generally become more marginal towards the edge of the species' range. Average predicted suitability based on the niche models declined with distance from the centre of the MCP (linear model: estimate ± SE = -0.309 ± 0.032, F = 93.16, df = 1, 274, p <<0.001; Fig. 2). Although eastern populations tended to have higher suitability scores (Fig. 2a), the main effect of the direction of sites relative to the geographic centre (i.e. north, south, east or west) was not significant (F = 0.33, df = 3, 274, p = 0.80; Type III sums of squares). However, the interaction between peripherality and direction was marginally significant (F = 2.50, df = 3, 274, p =0.059), with suitability declining slightly more steeply towards the northern edge of the range; Fig. 2a). These results were robust to the inclusion or exclusion of a single high suitability outlier near the eastern edge of the species' range (outliner shown in Fig. 2a).

The relationship between our index of niche centrality and geographic peripherality depended on direction away from the range centre (Fig. 2b). As expected if conditions become more marginal towards the edge of the range, distance to the niche centre increased moving away from the geographic centre of the species' range to the north and south. However, distance to the niche centre increased only gradually to the west and tended to decline (i.e. conditions became less marginal) towards the eastern edge of the species' range (Fig. 2b). These results were consistent regardless of the threshold of suitability and number of points in the PCA used to calculate the niche centroid (the significance of the main effect of peripherality varied across these iterations, however we note that it is difficult to interpret main effects in the presence of a significant interaction, and in all cases, results were qualitatively the same; Table S2-1). Results were once again robust to inclusion or exclusion of the outlier.

The variables that best predicted presence in the niche models exhibited a diversity of patterns at the edge of the species' range (Fig. 3). As expected based on latitude, average minimum spring temperatures tended to decline across northern range limits and increase across southern range limits. Model selection suggested that these changes were more abrupt than simple linear

models would predict (i.e. the four parameter logistic model was chosen in seven out of twelve transects; the linear model was chosen for the remaining five transects). Examination of the data revealed that change was most abrupt or step-like for transect 6 in the northeastern part of the species' range. The direction of change in precipitation of the wettest quarter was more variable across transects, although a consistent pattern of increasing precipitation towards the most southern range limits was observed (transects 8, 10, 11, 12; Fig. 3b). Our breakpoints analysis suggested that the three model forms tested may have been overly simplistic as, for both variables, the estimated number of breakpoints was usually higher than the number expected for the chosen model (i.e. expected 0 for intercept and linear models, 2 for four parameter logistic model; Table S2-2). Examination of the data for each transect also suggested that patterns of change were more complex than these simple models in many cases, although the models chosen were reasonable in some cases (Fig. 3 a,b).

Environmental marginality and patterns of genomic diversity

In line with the results of Griffin & Willi (2014), genomic diversity was negatively correlated with distance from the centroids of the two genetic groups that comprise the species' range ($r = -2.42$, $df = 40$, $p = 0.02$; Fig. 4a). Genomic diversity was not significantly associated with the predicted suitability of sites based on our niche models ($r = 1.39$, $df = 40$, $p = 0.17$; Fig. 4b). However, genomic diversity did decline with our distance to the centre of the species' niche in environmental space ($r = -3.70$, $df = 39$, $p = 0.0006$; Fig. 4c). This latter measure of marginality also explained a significant proportion of the residual variation in genomic diversity after accounting for range position.

Discussion

We evaluated niche limits and the link between environmental marginality and range limits in *Arabidopsis lyrata*. The species' northern and, to some extent, southern range limits were well predicted by our niche models suggesting that latitudinal range limits represent niche limits. In contrast, suitable habitat extended continuously beyond the species' western and northeastern range limits indicating that other factors limit range expansion in these directions. Environmental suitability tended to decline in all directions away from the centre of the species' range, with the underlying environment changing in a variety of ways at the edge of the species' range. However, the relationship between the environmental suitability of sites and genome-wide patterns of genetic diversity depended on the measure of suitability used. Thus although latitudinal range limits appear to be niche limits in this system, the specific consequences of environmental marginality for individual populations are less clear. We discuss the implications of these results below.

Niche limits, environmental marginality and range limits in Arabidopsis lyrata

Species' geographic ranges reflect the interplay of different ecological and evolutionary processes (reviewed by Sexton *et al.* 2009), as well as the contingency of historical events experienced by populations (e.g. (Hewitt 1996). Studying the relative importance of these different factors is necessary to fully understand species' range limits and the distribution of biodiversity more generally. Our niche modeling approach allowed us to evaluate the importance of several variables on the range limits of *A. lyrata* in North America. Consistent with the idea that range limits are manifestations of species' ecological niches (i.e. *sensu* Hutchinson 1957; e.g. Jackson & Overpeck 2000; Soberón 2007), we found that the northern and southern-most range limits of *A. lyrata* were well-predicted by our niche models. Furthermore, inline with Brown's model of declining conditions towards the edge of species' range (Brown 1984), we found that the predicted suitability of sites declined towards range limits in these directions. These results suggest that populations at the northern and southern edges of the species' range experience increasingly marginal (i.e. novel)

conditions and that range expansion in these directions may be precluded by constraints on adaptation to the variables considered here.

However, not all range limits were clearly associated with niche limits in this system. In contrast to latitudinal limits, the species' western range limit was not predicted by our niche models, especially when the minimum suitability score of any locality was used to gauge the overall suitability of sites. Suitable habitat also extended continuously beyond the species' range to the northeast. Although the predicted suitability of sites did decline towards the western and eastern limits of the species' range, the distance of sites to the centre of the species' niche in environmental space (i.e. the marginality of sites) showed only a moderate increase towards the species' western limit, and actually decreased towards the eastern limit. Thus it is less clear that peripheral populations in these directions face increasingly marginal conditions. Whereas most of the species' eastern range limit coincides with the Atlantic Ocean and is thus readily explained by a hard physical barrier, determinants of the species' western range limit remain ambiguous. The availability of moderately suitable habitat beyond this limit suggests that some other form of dispersal limitation may constrain the range in this direction. However, it is also possible that other abiotic or biotic factors not incorporated in our models set the western bounds of the species' range. All together, these findings underscore the potential for range limits to be shaped by different factors in different places.

Implications for adaptation at the edge of the range

Where niche models predict range limits, the variables with the greatest influence on model predictions can shed light on the dimensions along which adaptation at the range edge may be constrained. In *A. lyrata*, the variables that were most important to predicting presence on the landscape were average minimum temperature during the early spring and precipitation during the wettest quarter. Other annual and seasonal variables added very little to our models, although

general moisture availability, as represented by the Priestley-Taylor coefficient, was somewhat important. *A. lyrata* breaks seed dormancy during the early spring and undergoes a period of growth before flowering in May and June. Our results suggest that extreme temperatures and levels of precipitation during this period are detrimental to populations. This conclusion is consistent with previous work showing that frost and drought stress compromise plant size and thus may impact fitness in this system (Paccard *et al.* 2014; Wos & Willi 2015).

Apart from identifying variables associated with range limits, our analysis of environmental change at the edge of the range has implications for understanding how different ecological and evolutionary factors may impact adaptation. In several places, the range limits of *A. lyrata* were associated with roughly linear changes in the variables that were most important for predicting presence. In contrast, step-like changes in these variables were rarely observed. These results suggest that theoretical models for adaptation and range limits based on the assumption that the environment changes linearly (e.g. Kirkpatrick & Barton 1997; Bridle *et al.* 2010) may be more relevant to understanding range limits in this system than models based on discreet habitat patches (e.g. Holt & Barfield 2011). At the same time, changes in environment observed at the range edge were often better characterized by a four-parameter logistic model than a simple linear model and in many cases, range limits were associated with complex, non-linear patterns of environmental change. Recent studies have demonstrated that non-linear landscapes can alter the conditions necessary for trait evolution towards local optima (García-Ramos & Huang 2012; Schiffers *et al.* 2014). However, the number of alternative landscape scenarios and ecological and evolutionary processes considered in theoretical studies is limited to date. The complex patterns of environmental change observed at the range limits of *A. lyrata* argue for further investigation into the sensitivity of predictions from existing range limit theory to assumptions about the underlying environment.

Finally, the patterns of genomic diversity revealed in our analyses have implications for thinking about the adaptive potential of peripheral populations. A long-held explanation for the

failure of adaptation at range limits is that peripheral populations have limited genetic variation upon which selection can act (Hoffmann & Blows 1994; Blows & Hoffmann 2005). Griffin & Willi (2014) previously documented two genetic groups within *A. lyrata* and a pattern of declining genetic diversity away from the geographic centers of these groups based on microsatellite data. We followed up on those results in this study with genome-wide patterns of genetic diversity. Our results confirm a decline in genetic diversity towards the range edge and thus clearly establish a link between range position and diversity in this system. These low levels of diversity at the edge of the species' range may compromise the ability of populations to respond to the environmental changes discussed above.

The impact of environmental marginality on genomic diversity

In addition to range position, we tested the importance of marginal environmental conditions for explaining levels of genomic diversity. Following from the abundant centre hypothesis, genetic diversity is expected to decline towards the edge of the range as populations become smaller and more isolated in response to increasingly marginal conditions (Brown 1984; Sagarin & Gaines 2002; reviewed by Gaston 2003). We found mixed support for this hypothesis. Although genomic diversity declined significantly with increasing distance away from the estimated centre of the species' niche in environmental space and this measure of suitability explained a significant amount of the residual variation between genomic diversity and range position, niche model estimates of suitability were unrelated to genomic diversity. The reasons for these differences are unclear but we note that these two measures of suitability have different strengths and weaknesses. Whereas the niche modeling algorithm of MAXENT weights variables based on their importance in explaining presence on the landscape, the niche centrality measure of Lira-Noriega & Manthey (2014) emphasizes the variables that explain the most variation among in any point within the species' putative range. Likewise, whereas MAXENT can incorporate nonlinear relationships between

environmental conditions and suitability, the niche centrality method assumes suitability is a function of the linear distance of sites away from the center of the set of (PCA-transformed) environment conditions underlying all places where the species could occur. On the other hand, whereas MAXENT scores are relative and can be sensitive to sampling bias in the locality data (or variation in population density), we expect estimates of niche centrality to be less sensitive to any issues with sampling. Thus the two measures of suitability are complementary and the differences observed in the present study caution against reliance on either metric in isolation.

More generally a growing number of studies have questioned the relationship between genetic variation, range limits and the environmental suitability of sites. Eckert *et al.* (2008) reviewed the evidence for a genetic signature an abundant centre and found mixed support for this hypothesis. Recently, Lira-Noriega & Manthey (2014) analyzed genetic variation in relation to both geographic range position and niche centrality. They found that genetic diversity was consistently negatively associated with niche centrality but not with geography and used these results to argue that an abundant-centre may only arise when environmental suitability and geographic peripherality are positively correlated (Lira-Noriega & Manthey 2014). However, a number of alternative explanations (reviewed by Sagarin & Gaines 2002; Gaston 2003) can explain an abundant-centre distribution where it exists and several recent studies have called into question the importance of the environmental suitability of sites *per se* in generating genetic patterns across the range. For instance, Pironon *et al.* (2015) and Duncan *et al.* (2015) examined genetic variation in relation to geography, the contemporary environment, and the position of historical glacial refugia in plants and frogs respectively. Both studies found declining genetic variation towards the edge of the range but concluded that range dynamics associated with the last glacial cycle had a more profound effect on patterns of genetic variation than the contemporary environment (Pironon *et al.* 2015; Duncan *et al.* 2015). *A. lyrata* similarly occupies previously glaciated regions. The clear support for an association between genomic diversity and range position and limited evidence of an association

between genomic diversity and environmental suitability suggests that founder effects associated with colonization may have had a more pronounced impact on patterns of genomic diversity than environmental conditions in this system (although additional phylogeographic analyses are required to confirm the location of historical refugia and colonization routes in this system). Of note in light of the above studies is that where the direction of historical colonization routes and changes in environmental conditions covary it can be hard to disentangle the relative effects of the environment versus genetic drift on genetic diversity.

Limitations and future directions

Our study sheds light on environmental conditions and the challenges facing adaptation at the edge of the range in a widely distributed plant species in North America. Although our niche models point to specific abiotic variables that may be particularly limiting for *A. lyrata* at the edge of its range, we caution that it is not possible to definitively attribute range limits to these variables. Niche model predictions may be influenced by correlated but unmodeled variables (both biotic and abiotic) that systematically exclude the species from regions of environmental space that are otherwise suitable (Peterson *et al.* 2011). Thus direct experimentation is necessary to evaluate the causal link between the variables examined and the species' range limits.

The interpretation of our results also depends on the accuracy of designated range limits. *A. lyrata* sub. *lyrata* is part of a larger species' complex (Schmickl *et al.* 2010) and is replaced to the north by *A. arenicola*. Recent molecular work has called into question the distinctiveness of these two subspecies (Schmickl *et al.* 2010; Hohmann *et al.* 2014; Willi unpublished) and thus taxonomic designations and range limits in this system may need revising. At the same time, *A. arenicola* and *A. lyrata* sub. *lyrata* differ in morphology and mating system, with all evidence suggesting that the former is selfing and whereas the latter is mainly outcrossing. Thus, at the very least, our results are relevant to understanding the limits of outcrossing populations in this system (see Banta *et al.* 2012

for discussion of benefits of modeling the niche of distinct trait groups within species). Notably, selfing is restricted to the edge of the range in *A. lyrata* (Griffin & Willi 2014)—areas that are predicted to be less suitable for the species based on our models. That the replacement of *A. lyrata* with a selfing congener coincides with one of the sharpest declines in the suitability of sites for *A. lyrata* (e.g. the northern range limit) further establishes a link between selfing and the environment. The link between environmental conditions and the factors that promote selfing thus warrants further investigation in this and other systems.

Finally, our study contributes one of the first tests of genome-wide patterns of genetic diversity in relation to range position and the environmental suitability of sites. That the patterns we observed were consistent with results from previous work based on microsatellites (Griffin & Willi 2014) suggests that accurate conclusions about overall patterns of diversity may be reached without whole genome data, which is reassuring given that most studies investigating genetic diversity in relation to geography and the environment to date have relied on a limited number of markers (reviewed by Eckert *et al.* 2008; Lira-Noriega & Manthey 2014). At the same time, much more work is required to understand how these patterns relate to the adaptive potential of peripheral populations. Identifying the genes and traits involved in local adaption and examining variation in these genes and in the regulatory processes that control gene expression with respect to range position and the environment are all important avenues of future investigation that are currently underway in our group. In this regard genomic data pave the way for a more complete understanding of adaptation at the edge of the range. The coupling of such data with the types of spatial analyses presented here represents a powerful approach for advancing our understanding of species' geographic range limits.

References

Achim Z, Christian K, Walter K, Kurt H (2003) Testing and dating of structural changes in practice. *Computational Statistics and Data Analysis*, **44**, 109–123.

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Banta JA, Ehrenreich IM, Gerard S *et al.* (2012) Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in *Arabidopsis thaliana*. *Ecology letters*, **15**, 769–77.

Beatty GE, Provan J (2011) Phylogeographic analysis of North American populations of the parasitic herbaceous plant *Monotropa hypopitys L.* reveals a complex history of range expansion from multiple late glacial refugia. *Journal of Biogeography*, **38**, 1585–1599.

Blows MW, Hoffmann AA (2005) A reassessment of genetic limits to evolutionary change. *Ecology*, 1371–1384.

Boria RA., Olson LE, Goodman SM, Anderson RP (2014) Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, **275**, 73–77.

Bridle JR, Polechová J, Kawata M, Butlin RK (2010) Why is adaptation prevented at ecological margins? New insights from individual-based simulations. *Ecology letters*, **13**, 485–94.

Brown JH (1984) On the relationship between abundance and distribution of species. *The American Naturalist*, **124**, 255–279.

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195–205.

Commission for Environmental Cooperation (Montréal, Québec) and Secretariat (1997) Ecological regions of North America: Toward a Common Perspective.

Duncan SI, Crespi EJ, Mattheus NM, Rissler LJ (2015) History matters more when explaining genetic diversity within the context of the core-periphery hypothesis. *Molecular Ecology*, 4323–4336.

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Molecular Ecology*, **17**, 1170–88.

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Fracassetti M, Griffin PC, Willi Y (2015) Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata* (U Melcher, Ed,). *Plos One*, **10**, e0140462.

García-Ramos G, Huang Y (2012) Competition and evolution along environmental gradients: patterns, boundaries and sympatric divergence. *Evolutionary Ecology*, **27**, 489–504.

García-Ramos G, Kirkpatrick M (1997) Genetic models of adaptation and gene flow in peripheral populations. *Evolution*, **51**, 21–28.

Gaston KJ (2003) The structure and dynamics of geographic ranges. *Oxford University Press*, Oxford.

Gomulkiewicz R, Holt RD, Barfield M, Nuismer SL (2010) Genetics, adaptation, and invasion in harsh environments. *Evolutionary Applications*, **3**, 97–108.

Griffin PC, Willi Y (2014) Evolutionary shifts to self-fertilisation restricted to geographic range margins in North American *Arabidopsis lyrata*. *Ecology Letters*, **17**, 484–90.

Hargreaves AL, Samis KE, Eckert CG (2014) Are species' range limits simply niche limits writ large? A review of transplant experiments beyond the range. *The American Naturalist*, **183**, 157–73.

Hewitt G (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58**, 247–276.

Hijmans R, Phillips S, Leathwick J, Elith J (2013) dismo: species distribution modeling. *R package version 1.0-12*.

Hoffmann A, Blows MW (1994) Perspectives, species borders: ecological and evolutionary. *Trends in Ecology and Evolution*, **9**, 223–227.

Hohmann N, Schmickl R, Chiang T-Y *et al.* (2014) Taming the wild: resolving the gene pools of non-model *Arabidopsis* lineages. *BMC Evolutionary Biology*, **14**, 224.

Holt RD (2009) Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences of the United States of America*, **106 Suppl 2**, 19659–65.

Holt RD, Barfield M (2011) Theoretical perspectives on the statics and dynamics of species' borders in patchy environments. *The American Naturalist*, **178 Suppl**, S6–25.

Hu TT, Pattyn P, Bakker EG *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, **43**, 476–81.

Hutchinson G (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology, 22,*, 415–427.

Hutchinson G (1978) An introduction to population ecology. Yale University Press, New Haven.

Jackson S, Overpeck J (2000) Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology*, **26**, 194–220.

Jackson ST, Webb RS, Anderson KH *et al.* (2000) Vegetation and environment in Eastern North America during the Last Glacial Maximum. *Quaternary Science Reviews*, **19**, 489–508.

Kawecki T (2008) Adaptation to marginal habitats. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 321–342.

Keller LF, Waller DM (2002) Inbreeding effects in wild populations. *Trends in Ecology and Evolution*, **17**, 230–241.

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kirkpatrick M, Barton N (1997) Evolution of a species' range. *The American Naturalist*, **150**, 1–23.

Koboldt DC, Zhang Q, Larson DE *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**, 568–76.

Lee-Yaw JA, Irwin JT, Green DM (2008) Postglacial range expansion from northern refugia by the wood frog, *Rana sylvatica*. *Molecular Ecology*, **17**, 867–84.

Lee-Yaw JA, Kharouba HM, Bontrager M *et al.* (2016) A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits (JM Gomez, Ed,). *Ecology Letters*.

Li P, Li M, Shi Y *et al.* (2013) Phylogeography of North American herbaceous *Smilax* (*Smilacaceae*): Combined AFLP and cpDNA data support a northern refugium in the Driftless Area. *American Journal of Botany*, **100**, 801–814.

Lira-Noriega A, Manthey JD (2014) Relationship of genetic diversity and niche centrality: a survey and analysis. *Evolution*, **68**, 1082–93.

Liu C, White M, Newell G (2013) Selecting thresholds for the prediction of species occurrence with presence-only data (R Pearson, Ed,). *Journal of Biogeography*, **40**, 778–789.

Micheletti SJ, Storfer A (2015) A test of the central-marginal hypothesis using population genetics and ecological niche modelling in an endemic salamander (*Ambystoma barbouri*). *Molecular Ecology*, **24**, 967–979.

Paccard A, Fruleux A, Willi Y (2014) Latitudinal trait variation and responses to drought in *Arabidopsis lyrata*. *Oecologia*, **175**, 577–87.

Peterson A, Soberón J, Pearson RG *et al.* (2011) Ecological niches and geographic distributions. Princeton University Press.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips S, Dudík M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 161–175.

Pironon S, Villellas J, Morris WF, Doak DF, García MB (2015) Do geographic, climatic or historical ranges differentiate the performance of central versus peripheral populations? *Global Ecology and Biogeography*, **24**, 611–620.

Sagarin RD, Gaines SD (2002) The "abundant centre" distribution: to what extent is it a biogeographical rule? *Ecology Letters*, **5**, 137–147.

Schiffers K, Schurr FM, Travis JMJ *et al.* (2014) Landscape structure and genetic architecture jointly impact rates of niche evolution. *Ecography*, **37**, 1218–1229.

Schmickl R, Jørgensen MH, Brysting AK, Koch M a (2010) The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evolutionary Biology*, **10**, 98.

Sexton JP, McIntyre PJ, Angert AL, Rice KJ (2009) Evolution and ecology of species range limits. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 415–436.

Soberón J (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, **10**, 1115–23.

Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–93.

Soule M (1973) The epistasis cycle: a theory of marginal populations. *Annual Review of Ecology and Systematics*, **4**, 165–187.

Varela S, Anderson RP, García-Valdés R, Fernández-González F (2014) Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 1084–1091.

Warren D, Seifert S (2011) Ecological niche modeling in Maxent : the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, **21**, 335–342.

Warren DL, Wright AN, Seifert SN, Shaffer HB (2014) Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern (J Franklin, Ed,). *Diversity and Distributions*, **20**, 334–343.

Willi Y, Määttänen K (2010) Evolutionary dynamics of mating system shifts in *Arabidopsis lyrata*. *Journal of Evolutionary Biology*, **23**, 2123–31.

Wos G, Willi Y (2015) Temperature-stress resistance and tolerance along a latitudinal cline in North American *Arabidopsis lyrata*. *PloS One*, **10**, e0131808.

Zeileis A, Leisch F, Hornik K, Kleiber C (2002) strucchange: an R package for testing for structural change in linear regression models. *Journal of Statistical Software*, **7**, 1–38.

Figures

Figure 1. The distribution of *Arabidopsis lyrata* sub. *lyrata* in North America. Black circles in all panels indicate known localities (thinned dataset), with the species' range limits represented by the minimum convex polygon around these points (outer polygon). Panel a shows the centroid of the MCP of the species' entire range (red star), genomic sampling sites (blue circles), the boundaries (inner polygons) and centroids (red triangles) of the western and eastern genetic groups described by Griffin & Willi (2014) and the location and orientation of the range-edge transects (numbered grey lines) used in Fig. 3. Panels b-d show the predicted distribution of suitable conditions for *A. lyrata* across the study region based on niche models built using MAXENT. Continuous estimates of suitability are shown in panel b and represent the average prediction from ten rounds of model calibration. Binary maps of suitable habitat were generated from the continuous prediction surface using c) the minimum and d) the 5th percentile of the average suitability scores of the locality data as cutoffs for defining suitable habitat.

Figure 2. The relationship between geographic peripherality (distance to geographic range centre/ [distance to geographic range centre + distance to hull of minimum convex polygon around all known localities]) and environmental marginality in North American populations of *Arabidopsis lyrata*. a) The average predicted suitability of sites declines with increasing peripherality in all major directions away from the range centre. The open circle in this plot represents a peripheral site with particularly high suitability at the eastern edge of the species' range. The inclusion or exclusion of this site did not change the results. b) The distance of localities to the estimated centre of the species' niche in environmental space increased with increasing peripherality towards the north and south in particular and actually decreased towards the eastern edge of the species' range (outlier not shown for clarity of trends). The niche centroid used here was calculated from a PCA of the variables used in the niche models based on the locality data, plus 5000 points sampled randomly

from the extent of suitable conditions for the species as determine by the minimum suitability score of any presence. Using a different number points in the PCA or a different threshold for defining suitable habitat did not qualitatively change the results.

Figure 3. Changes in a) mean early spring temperature and b) precipitation of the wettest quarter towards and across range limits in *Arabidopsis lyrata* based on transects centered on different peripheral populations. Plots in each panel correspond to the transects depicted in Fig. 1 and are ordered from west to east for northern (top row) and southern (bottom row) sections of the species' range. Transects spanned the range edge (100 km in either direction) with the focal periperphal population located at position 0. Negative values along the x-axis represent locations within the species range, positive values represent locations over the edge of the range. For clarity, the y-axis varies across plots and colour is used instead to represent the environmental value at a given location.

Figure 4. Genomic diversity of 42 outcrossing populations of *Arabidopsis lyrata* in relation to a) (arcsin) distance to putative refugial areas and b) the average predicted suitability of sites based on niche model predictions and c) the distance of sites from the centre of the species' niche in environmental space. For the latter, the minimum suitability score of the localities was used as a cutoff when calculating the niche centroid (see main text). Results are qualitatively similar for the alternative cutoff.

Tables

Table 1. Final environmental variables used to generate niche models for *Arabidopsis lyrata* sub. *lyrata* and their effects on model performance.

Type	Variable	Abb	Source*	Basis of inclusion§	Percent contribution†	Permutation importance†
Climate	Mean diurnal range (temperature)	Bio2	BioClim	RF	8.65	2.87
	Maximum temperature of warmest month	Bio5	BioClim	BI	9.32	0.45
	Precipitation Seasonality	Bio15	BioClim	RF	4.04	6.74
	Precipitation of Wettest Quarter	Bio16	BioClim	BI / RF**	9.48	22.3
	Priestley-Taylor coefficient	alpha	CGAIR	BI	4.51	18.47
	Average minimum temperature of early spring (March, April)	Tmin_sp	Derived from WorldClim	BI	61.63	46.71
Topography	Compound topographic index	CTI	USGA (Hydro1k)	BI	0.21	0.43
Vegetation	Variability in Maximum Green Vegetation Fraction	mgvf_sd	Derived from USGS-LCI	BI	2.15	2.03

* Sources: WorldClim and BioClim: WorldClim database (http://www.worldclim.org); CGAIR: CGIAR Consortium for Spatial Information (http://www.cgiar-csi.org/data); HWSD: Harmonized World Soil Database (http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/); USGS LCI: U.S. Geological Survey Land Cover Institute (http://landcover.usgs.gov/green_veg.php); USGA Hydro1k: U.S. Geological Survey (EROS Center) topographically derived datasets (https://lta.cr.usgs.gov/HYDRO1K)

§ Variables were either chosen based on information about the biology of the species (BI) or because important for classifying presence and background sites in a random forest analysis (RF); see Appendix S1

† Based on MAXENT output and averaged across those k-fold models passing modeling evaluation

** Bio 16 was also highly correlated with the derived mean spring and mean summer precipitation variables that we had *a priori* singled out as being important for the species but that were subsequently dropped

Fig. 1

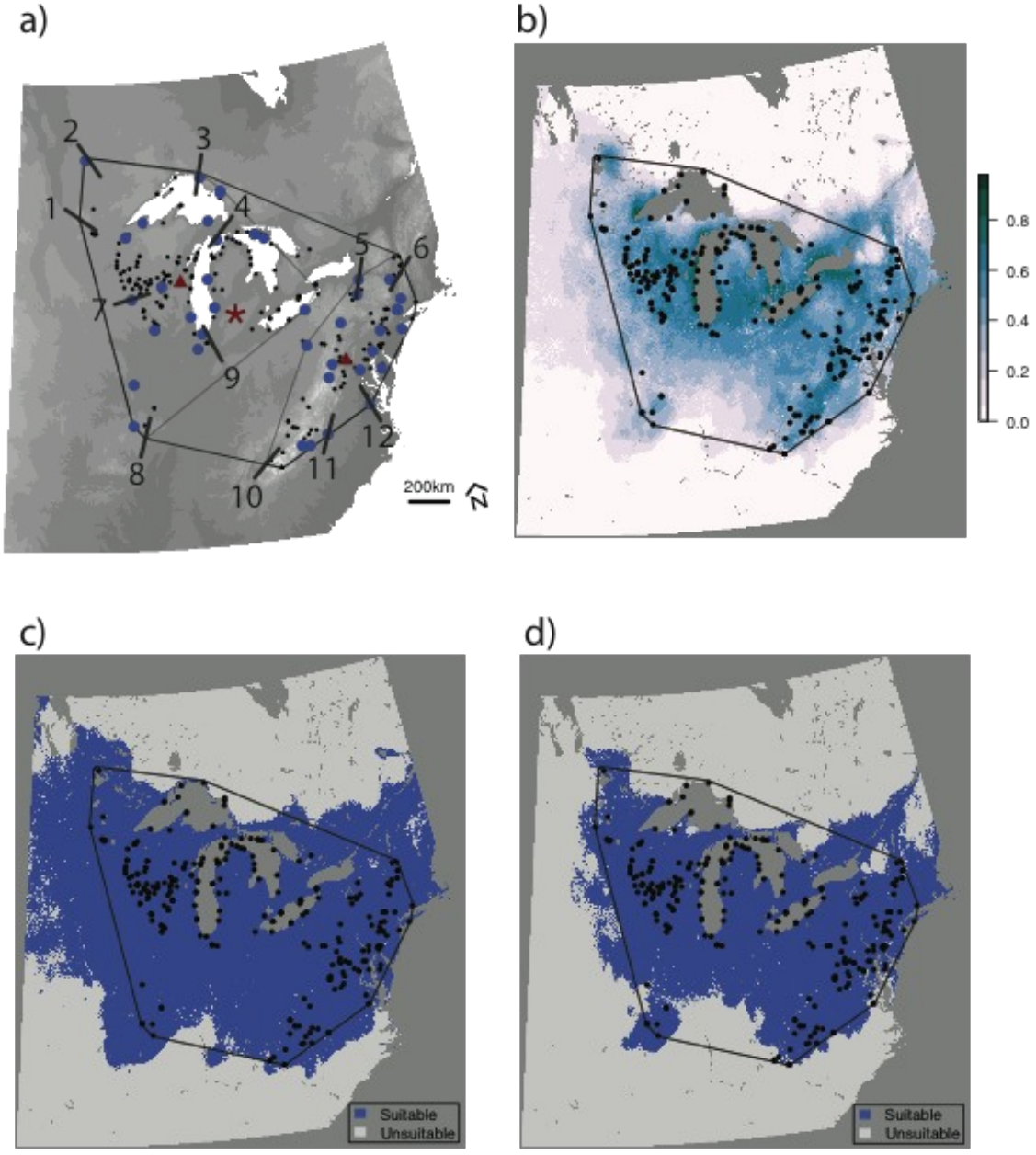

a)
b)
c)
d)
1
2
3
4
5
6
7
8
9
10
11
12
200km
0.8
0.6
0.4
0.2
0.0
Suitable
Unsuitable
Suitable
Unsuitable

Fig. 2

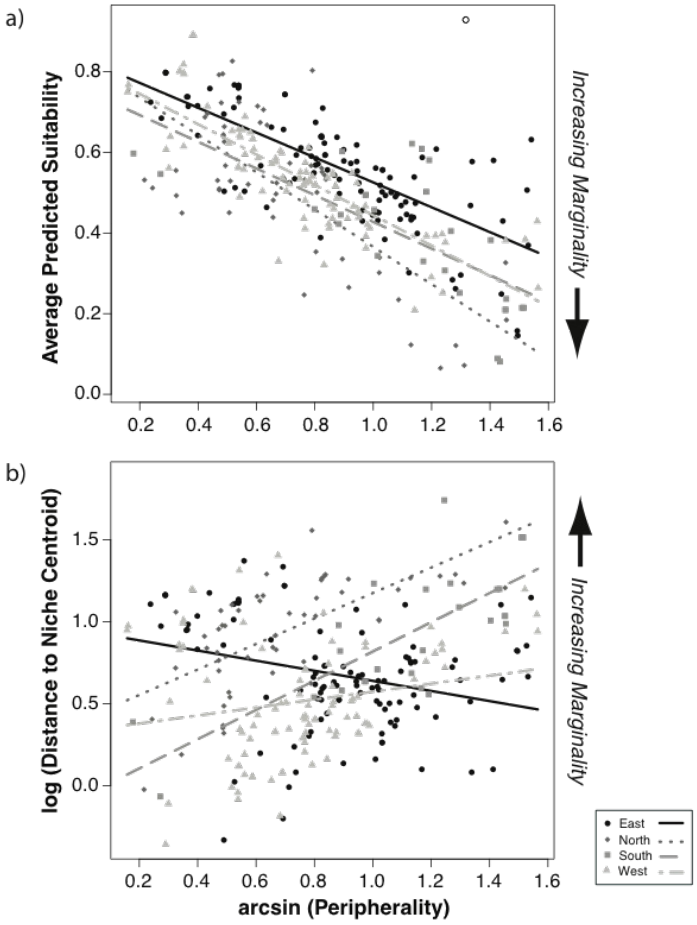

a)
Average Predicted Suitability
0.0
0.2
0.4
0.6
0.8
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Increasing Marginality
b)
log (Distance to Niche Centroid)
0.0
0.5
1.0
1.5
arcsin (Peripherality)
Increasing Marginality
East
North
South
West

Fig. 3

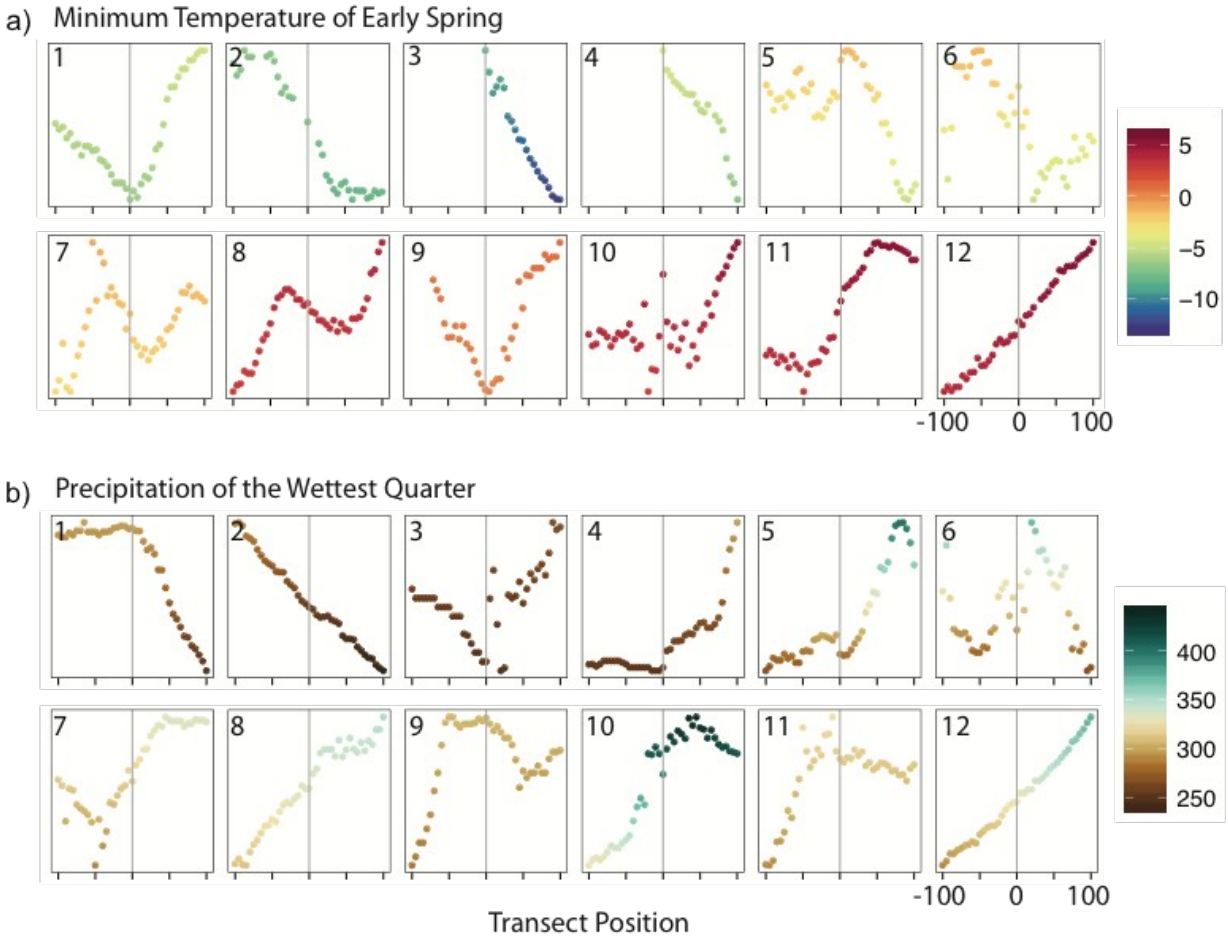
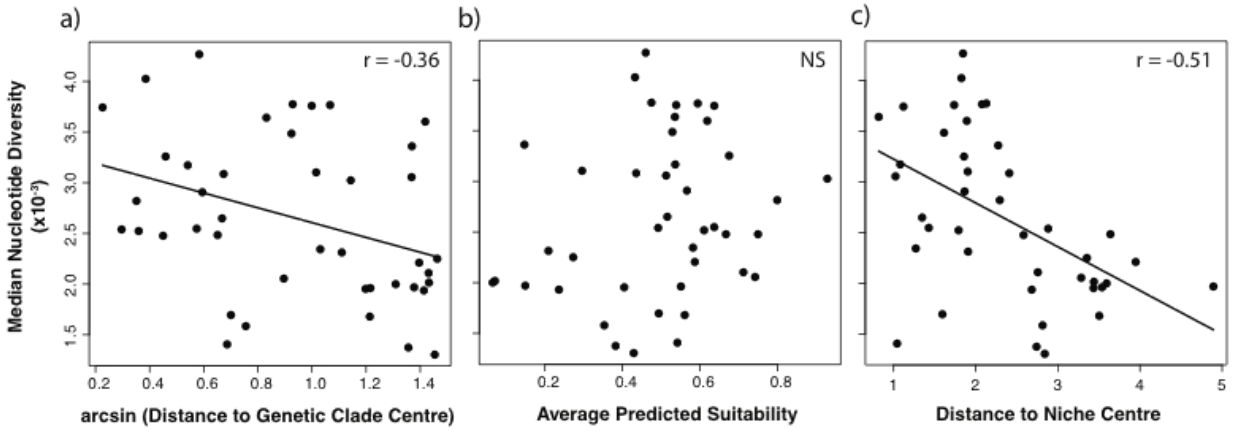

a) Minimum Temperature of Early Spring
1
2
3
4
5
6
7
8
9
10
11
12
5
0
−5
−10
-100
0
100
b) Precipitation of the Wettest Quarter
400
350
300
250
Transect Position

Fig. 4


a)
b)
c)
r = -0.36
NS
r = -0.51
Median Nucleotide Diversity (x10⁻³)
arcsin (Distance to Genetic Clade Centre)
Average Predicted Suitability
Distance to Niche Centre

CHAPTER 4: Genes linked to climate and substrate in *Arabidopsis lyrata*

Authors: M. Fracassetti[1,2]*, Y. Willi[1,2]

Affiliations:

[1] Institute of Biology, University of Neuchâtel, Rue Emile Argand 11, CH-2000 Neuchâtel, Switzerland

[2] Department of Environmental Sciences, University of Basel, Schönbeinstrasse 6, CH-4056 Basel, Switzerland

*Correspondence to: marco.fracassetti@unibas.ch

Abstract

Adaption to climate and edaphic conditions are ubiquitous in plants, but little is known about the traits and genes involved in adaptation. We performed an environmental association analysis (EAA), linking genetic variation with climatic variables known to determine the climatic niche of our study organism and substrate type (sandy and rocky sites). The study species was the North American *Arabidopsis lyrata* spp. *lyrata*. We re-sequenced pools of DNA of 42 outcrossing populations from the entire geographic range and revealed population-level single-nucleotide-polymorphism (SNP) frequency data. SNP data was regressed against the environmental variables. Subsequently we performed a gene ontology analysis on the correlated genes which suggested the top candidate genes. The highest number of associated SNPs and candidate genes were linked to precipitation during the wettest quarter of the year, followed by minimum temperature in early spring, substrate type and Priestley-Taylor alpha coefficient. The associated SNPs clustered in particular genomic regions for all environmental variables. The strength of the associations was higher for populations of the Western genetic cluster, which had more standing genetic variation that seemed to allow them to evolve improved local adaptation.

Keywords: adaptation to edaphic conditions; Brassicaceae; climate adaptation; gene ontology; genome-wide association.

Introduction

Climate and soil have been reported as main drivers of adaptive evolution in plants, but little is known about the genetic basis of such adaptation (Alvarez *et al.* 2009; Toledo *et al.* 2011). Knowing the genetic basis is important from a fundamental science point of view, but also in the face of global change. In the last decades evidence has further accumulated that global temperature is increasing and that other climatic conditions are changing (Pachauri *et al.* 2014). In the face of anthropogenic climate change the persistence of species may be strongly promoted if populations and species can adapt to changing conditions (Lynch & Lande 1993; Anderson *et al.* 2011). In turn, the ability to adapt depends on the genetic variation in traits selected by climate. Hence, to understand adaptation and to predict adaptive potential, we first have to understand what genes, gene networks and phenotypic traits are under selection.

If phenotypic change to climate or any other environmental factor is genetically based (Hoffmann & Sgrò 2011), it involves a shift in the allele frequencies; the change is heritable and evolution occurs. In a recent review, Franks *et al.* (Franks *et al.* 2014) found that evolutionary response to climate change are widespread among plants. Out of 38 analyzed studies they found that 26 studies showed an evolutionary response. However, in the case of evolutionary response the allelic change does not act in a simple manner but in a gene-network context. These networks include various genes, transcription factors and proteins that interact together to produce a phenotype suitable for the environment (Franks & Hoffmann 2012). One well-known example is flowering time in *Arabidopsis thaliana*; it is controlled by several pathways which respond to different environmental cues including climatic ones (Wilczek *et al.* 2010). In the case of evolutionary response, local adaption of populations takes place (Williams 1966). Local adaption is favored under conditions of low gene flow, moderate selection on intermediate genotypes, little temporal variation in the direction of selection, under costs of plasticity – the other way to cope

with environmental change, and large population size (Kawecki & Ebert 2004; Leimu & Fischer 2008).

The study of environmental factors that shape adaptive genetic variation and the investigation of gene variants that drive local adaptation is central to the field of landscape genomics (Turner *et al.* 2001; Manel *et al.* 2003). The core of landscape genomics is the Environmental Association Analysis (EAA), where alleles or genotypes are associated with an environmental factor, while controlling for neutral genetic structure. In this study we used a pool-sequencing approach (Pool-seq), which was shown to be a cost-effective method to obtain genome-wide allele frequency data (Schlötterer *et al.* 2014). Pool-seq has been used in different organism from bacteria (Holt *et al.* 2009) to humans (Bansal *et al.* 2011). Particularly, it has been used to perform Genome-Wide Association Studies (GWAS) in the genus *Arabidopsis* (Turner *et al.* 2010; Fischer *et al.* 2013). Several methods are available for environmental association analysis (reviewed in Rellstab *et al.* 2013), but only the Bayesian hierarchical model proposed by Coop et al. (Coop *et al.* 2010), implemented in Bayenv2 (Günther & Coop 2013), can handle Pool-seq data. This method has been improved in the program Baypass (Gautier 2015), where a binary auxiliary variable was introduced to classify each locus as associated or not.

Our study organism was *A. lyrata* spp. *lyrata*, a short-lived perennial plant with a predominantly outcrossing reproductive mode. Here we analyzed 42 outcrossing populations which covered the known specie range (Schmickl *et al.* 2010; Paccard *et al.* 2016). These populations belong both to the Eastern and Western genetic cluster of the species (Hoebe *et al.* 2009; Willi & Määttänen 2010). Common garden analysis of some of these populations had shown that phenotypic traits were linked to latitude and therefore mainly to climate. Plants from the North – compared to the South – grew to larger size, flowered earlier, were more frost resistant and less heat resistant and less heat tolerant (Paccard *et al.* 2014; Wos & Willi 2015).

Furthermore, extensive niche model analyses revealed that the latitudinal range limit of *A. lyrata* coincides with the niche limit, and the niche limit was determined strongest by temperature – minimum temperature in early spring – and to a lesser extent by water availability and precipitation (Lee-Yaw *et al.* submitted). During the early spring *A. lyrata* increases the rosette size. By April and May the plants start bolting and flower until about June and July. After fertilization of ovules, a relatively long period of 4 weeks is needed until the fruits are ripe, at a time when summer dryness becomes stronger. Another feature of the species is that populations either occur in predominantly sandy places or rocky sites. Sandy places where the species occur are close to the shore of the Atlantic Ocean, on Lake Erie, Lake Michigan and parts of Lake Superior. Rocky sites where the species occurs can be found in the Appalachians, on parts of the shore of Lake Superior and along large rivers. For both climatic and edaphic conditions, it is likely that populations adapted to them.

In this study, we addressed the following specific questions: 1) What are the SNPs associated with the different environmental conditions – climate and substrate type? Is there consistency in associated SNPs between the two ancestral groups of populations, in the Appalachians and in the Mid-West, indicating convergent evolution? 2) What are the overrepresented gene ontology terms? And 3) what is the overlap between the genes found here and in studies of other *Arabidopsis* species, suggesting convergence in pathways involved in adaptation across species?

Materials and Methods

Genomic data

In this study we analyzed 42 outcrossing populations of *Arabidopsis lyrata* ssp. *lyrata* (figure 1, table S1). The genomic data used in this study was taken from the data set analyzed in Fracassetti

& Willi (in prep), where pools of 25 individual DNA samples were sequenced according the pool-sequencing protocol described in Fracassetti et al. (2015). The SNPs were called by the program VarScan (Koboldt *et al.* 2012) for each population individually with a minimum count of the variant allele of 3, a minimum frequency of the variant allele of 0.015, a P-value lower than 0.15, minimum mapping quality of 20 and a strand bias less than 90%. We retained 1,692,676 biallelic SNPs, with a MAF (minimum allele frequency) across the populations of > 0.05 and when the specific site had been sequenced at least in half of the populations.

Selection of climatic variables

Four environmental variables were used to test for an association with SNPs. The first variable was the minimum temperature early spring (March, April; **Tmin_ESp**). During this period *A. lyrata* is very susceptible to cold and frost since plants start growing. The second variable was the precipitation during the wettest quarter (**bio16**). The third variable was the Priestley-Taylor coefficient that indicates the general water availability of the environment (**alpha**). Tmin_ESp and bio16 were estimated from worldclim data (http://www.worldclim.org). Alpha was estimated from the data of the Consortium for Spatial Information (http://www.cgiar-csi.org/data). These three climatic variables had been shown previously to predict well habitat suitability and niche limits of the species (Lee-Yaw *et al.* submitted). The fourth variable was the substrate on which the plants grew: rocky or sandy sites (**Sub**). The values of environmental variables for each population are listed in table S1. We performed a principal component analysis (PCA) on the values of these variables of each population with the R package ggbiplot (https://github.com/vqv/ggbiplot). The PCA showed the environmental variation between the sampled populations (figure S1).

Testing for an association between genomic variation and the environment

We performed an environmental association analysis with the program Baypass (Gautier 2015). The bayesian method controls for (co-)variance in population SNP frequencies due to sampling (individuals within populations and sequencing of pooled DNA) and population history while testing for a correlation between SNP frequency and environmental variable. We used the auxiliary covariate model with the default parameters. To detect the posterior probability of each auxiliary variable associated with each SNP, we ran 25,000 MCMC (Markov chain Monte Carlo) simulations with a burn-in time of 5,000 simulations. The population variance matrix was calculated based on 50,000 randomly picked, intergenic SNPs. We compared two different variance matrices calculated based on different random sets of intergenic SNPs. The Forstner-Moonen Distance (FMD; Förstner & Moonen 2003) turned out to be low (FMD = 0.48), which indicates that 50,000 SNPs were enough to detect populations structure in the data set. The program was run independently three times to avoid high run-to-run variability detected in Bayenv-like methods (Blair *et al.* 2014). We considered a SNP to be significant ("associated") when the Bayes Factor (BF) was greater than 20 decibians in all the three independent runs. Subsequently, the Spearman correlation coefficient (rho) between the frequencies of the associated SNPs revealed by the EAA and environmental variables was calculated for each genetic cluster separately (table S2), and a t-test was performed on the absolute values of rho between the two genetic cluster. We performed a PCA on the all SNPs frequencies of populations grouped by the two genetic clusters. Finally, we performed a gene ontology (GO) analysis on the associated SNPs with the R package snp2go (Szkiba *et al.* 2014) using the most recent annotation of *A. lyrata* (Rawat *et al.* 2015) and only the biological process subcategory. The program performed a candidate SNP enrichment analysis based on the number of candidate SNPs and non-

candidate SNPs in a GO term. The overrepresented GO terms with a false discovery rate (FDR) lower than 0.05 were reduced with revigo (Supek *et al.* 2011) based on semantic similarity.

Results and Discussion

SNPs associated with environmental variables

Environmental Association Analysis (EAA) revealed 9,327 SNPs associated with the four environmental variables of minimum temperature early spring (Tmin_ESp), precipitation during the wettest quarter (bio16), the Priestley-Taylor coefficient of general water availability (alpha) and substrate type (Sub) – rock versus sand (table S2). Only 45 SNPs were associated with more than one environmental variable. Precipitation of the wettest quarter was the environmental variable with the highest number of associated SNPs (5005 SNPs), followed by minimum temperature early spring (2637 SNPs), substrate type (1308 SNPs) and alpha (422 SNPs) (Table 1). For each environmental variable more of the associated SNPs were in genic regions compared to intergenic regions (figure 2). The difference was higher for precipitation during the wettest quarter and alpha (77% and 66% of SNPs in genic regions, respectively) compared to minimum temperature early spring and substrate type (58% and 53% of SNPs in genic regions, respectively). The associated SNPs clustered in distinct genomic regions (figure 3, table 2). Particularly, 40.46% of the SNPs associated with bio16 were located in chromosome 2 between the positions 17,247,881 and 17,630,257.

In a next step, we investigated the strength of associations for the Eastern and Western ancestral genetic clusters separately. The absolute values of Spearman correlation coefficients were significant higher in the Western genetic cluster for all environmental variables (p value of t-test < 0.001) (figure 4). The PCA done on the SNP frequencies (figure S2) indicated that the

SNP frequencies in the Western genetic cluster were more differentiated, which may have contributed to strengthen associations.

Gene ontology and genes associated with environmental variables

The number of associated genes and significant gene ontology (GO) terms were higher for precipitation during the wettest quarter, followed by minimum temperature early spring, substrate type and alpha. The trend was similar to that found for the number of associated SNPs (table 1). In the following paragraphs, results of the gene ontology analysis are described for each of the four environmental variables, followed by a discussion of candidate genes. The candidate genes were selected picking the gene with the highest number of SNPs associated for each genomic regions that had more than 50 SNPs associated (table 2). The description of these genes is written in table 3.

The environmental variable of minimum temperature during the months of March and April had 557 associated genes. Most of the SNPs that were found to be associated – 57.32% – were located in 67 genes, each of which with more than 5 associated SNPs (the latter was part of the settings for finding associated genes). The GO enrichment analysis revealed 35 GO terms overrepresented (table S3) and they mainly belonged to the GO terms "branched-chain amino acid biosynthesis" (GO:0009082) and "regulation of hydrolase activity" (GO:0051336) (figure 6). One candidate gene (AL2G25500) was found to be involved in the response to both frost and heat in *A. lyrata* (Wos & Willi submitted). This gene encoded a midasin-like protein involved in seed development (Chen *et al.* 2014). Other two candidate genes (AL3G32800, AL2G31960) had a homologous gene in *A. thaliana*. The first encoded a protease involved in programmed cell death (Ondzighi *et al.* 2008). The second encoded a water-soluble chlorophyll protein involved in herbivore resistance activation (Boex-Fontvieille *et al.* 2015).

Precipitation during the wettest quarter had 832 genes associated. Most of the SNPs associated (69.66%) were in 127 genes that had more than 5 SNPs. The GO enrichment analysis revealed 97 overrepresented GO terms (table S3), and after a reduction based on semantic similarity in GO terms, the most significant GO terms were: translational initiation (GO:0006413), response to UV-B (GO:0010224), sepal development (GO:0048442) and exocytosis (GO:0006887) (figure 5). In line with SNP numbers, the highest number of candidate genes were found for precipitation during the wettest quarter, and they were located on chromosome 2. Three of the candidate genes (AL2G15270, AL2G15460, AL2G28070) had previously been found in association with the response to salt stress (Sottosanto *et al.* 2004; Reis 2014; Zhang *et al.* 2016). Another candidate, AL2G36920, encodes a heat shock protein involved in multiple stress response pathways (Swindell *et al.* 2007) and the AL2G24570 gene is involved in the response to abscisic acid and drought (Ren *et al.* 2010). Other genes that are candidates based on this study, AL2G27910 and AL5G24920, had been shown to be involved in defense response (Reumann 2013) and response to fungus (Oelmuller *et al.* 2005). The remaining candidate genes had no counterparts in *A. thaliana*.

The Priestley-Taylor coefficient, reflecting general water availability in the environment, had 93 genes associated. Most of the SNPs associated (62.99%) were in 11 genes that had more than 5 SNPs. The GO enrichment analysis revealed 15 GO terms overrepresented (table S3), which could be grouped in "protein acetylation" (GO:0006473) and "tryptophan biosynthesis" (GO:0000162) (figure 8). One of the candidate genes of alpha had no homologue in *A. thaliana*, the other gene (AL4G36300) encoded a acyltransferases involved in the response to drought (Trijatmiko 2005).

Substrate type had 321 genes associated. The percentage of SNPs associated in 33 genes with more than 5 SNPs (42.45%) was lower compared to the former two variables, minimum

temperature in spring and precipitation during the wettest quarter. The GO enrichment analysis revealed 23 GO terms overrepresented (table S3), and after a reduction of terms based on semantic similarity, they grouped in "response to UV-B" (GO:0010224), "vesicle docking" (GO:0048278) and "glutamine metabolism" (GO:0006541) (figure 7). The candidate gene, AL2G14660, encodes a kinase protein, which is involved in a signal transduction pathway (Nemoto *et al.* 2011).

Finally, we also tested for an overlapped in the candidate genes found in other studies on *Arabidopsis*: one was done on *A. thaliana* (Hancock *et al.* 2011) and the other in *A. halleri* (Fischer *et al.* 2013). Few genes overlapped between the different studies (figure 5).

Conclusions

This study has highlighted the genomic regions that are associated with environmental variables that determine the specie distribution (Lee-Yaw *et al.* submitted) and substrate type. Most of the associated SNPs clustered in particular genomic regions. The pattern was most noticeable for precipitation during the wettest quarter, for which 58.6% of SNPs clustered together. For the other variables, the pattern was weaker: for Priestley-Taylor alpha 35.55%, minimum temperature early spring 21.8% and for substrate type 4.91%. The strength of the associations was higher in the Western genetic cluster compared to the Eastern genetic cluster. This could be due to the fact that the SNP frequencies in the Western cluster were more differentiated, therefore the Western populations had more standing genetic variation that allowed them to adapt better to different environments. The overlap between the candidate genes of this study and other studies on *Arabidopsis* (Hancock *et al.* 2011; Fischer *et al.* 2013) was found to be low but the overrepresented GO terms were similar, such as defense response, response to UV and translational initiation.

REFERENCES

Alvarez N, Thiel-Egenter C, Tribsch A *et al.* (2009) History or ecology? Substrate type as a major driver of patial genetic structure in alpine plants. *Ecology Letters*, **12**, 632–640.

Anderson JT, Willis JH, Mitchell-Olds T (2011) Evolutionary genetics of plant adaptation. *Trends in Genetics*, **27**, 258–266.

Bansal V, Tewhey R, Leproust EM, Schork NJ (2011) Efficient and cost effective population resequencing by pooling and in-solution hybridization. *PloS one*, **6**, e18353.

Blair LM, Granka JM, Feldman MW (2014) On the stability of the Bayenv method in assessing human SNP-environment associations. *Human genomics*, **8**, 1.

Boex-Fontvieille E, Rustgi S, von Wettstein D, Reinbothe S, Reinbothe C (2015) Water-soluble chlorophyll protein is involved in herbivore resistance activation during greening of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 7303–8.

Chen C, Wu C, Miao J *et al*. (2014) *Arabidopsis* SAG protein containing the MDN1 domain participates in seed germination and seedling development by negatively regulating ABI3 and ABI5. *Journal of experimental botany*, **65**, 35–45.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Fischer MC, Rellstab C, Tedder A *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Molecular Ecology*, **22**, 5594–5607.

Förstner W, Moonen B (2003) A metric for covariance matrices. In: *Geodesy-the challenge of the 3rd millennium* , pp. 299–309. Springer Berlin Heidelberg, Berlin, Heidelberg.

Fracassetti M, Griffin PC, Willi Y (2015) Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata* (U Melcher, Ed,). *PLOS ONE*, **10**, e0140462.

Franks SJ, Hoffmann AA (2012) Genetics of climate change adaptation. *Annual Review of Genetics*, **46**, 185–208.

Franks SJ, Weber JJ, Aitken SN (2014) Evolutionary and plastic responses to climate change in terrestrial plant populations. *Evolutionary Applications*, **7**, 123–139.

Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–79.

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–20.

Hancock AM, Brachi B, Faure N (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, **334**, 83–86.

Hoebe PN, Stift M, Tedder A, Mable BK (2009) Multiple losses of self-incompatibility in North-American *Arabidopsis lyrata* □?: Phylogeographic context and population genetic consequences. *Molecular Ecology*, **18**, 4924–4939.

Hoffmann AA, Sgrò CM (2011) Climate change and evolutionary adaptation. *Nature*, **470**, 479–485.

Holt KE, Teo YY, Li H *et al.* (2009) Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. *Bioinformatics (Oxford, England)*, **25**, 2074–2075.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters*, **7**, 1225–1241.

Koboldt DC, Zhang Q, Larson DE *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**, 568–576.

Leimu R, Fischer M (2008) A meta-analysis of local adaptation in plants. (A Buckling, Ed,). *PloS one*, **3**, e4010.

Lynch M, Lande R (1993) Evolution and extinction in response to environmental change. In: *Biotic Interaction and Global Change* (eds Kareiva PM, Kingsolver JG, Huey RB), pp. 234–250. Sinanev Associate.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Nemoto K, Seto T, Takahashi H *et al.* (2011) Autophosphorylation profiling of *Arabidopsis* protein kinases using the cell-free system. *Phytochemistry*, **72**, 1136–44.

Oelmuller R, Peskan-Berghofer T, Shahollari B *et al.* (2005) MATH domain proteins represent a novel protein family in *Arabidopsis thaliana*, and at least one member is modified in roots during the course of a plant-microbe interaction. *Physiologia Plantarum*, **124**, 152–166.

Ondzighi AC, Christopher DA, Cho EJ, Chang S-C, Staehelin LA (2008) *Arabidopsis* protein disulfide isomerase-5 inhibits cysteine proteases during trafficking to vacuoles before programmed cell death of the endothelium in developing seeds. *The Plant cell*, **20**, 2205–20.

Paccard A, Van Buskirk J, Willi Y (2016) Quantitative genetic architecture at latitudinal range boundaries: reduced variation but higher trait independence (CG Eckert, JL Bronstein, Eds,). *The American Naturalist*, **187**, 667-677.

Paccard A, Fruleux A, Willi Y (2014) Latitudinal trait variation and responses to drought in *Arabidopsis lyrata*. *Oecologia*.

Pachauri RK, Allen MR, Barros VR *et al.* (2014) Climate change 2014: synthesis report. contribution of working groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change. *EPIC3Geneva, Switzerland, IPCC, 151 p., pp. 151, ISBN: 978-92-9169-143-2*.

Rawat V, Abdelsamad A, Pietzenuk B *et al.* (2015) Improving the annotation of *Arabidopsis lyrata* using RNA-seq data. *PloS one*, **10**, e0137391.

Reis MV dos (2014) Gene expression profiles in roses under stress conditions. Universidade federal del lavras.

Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC (2013) Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS ONE*, **8**, e80422.

Ren X, Chen Z, Liu Y *et al.* (2010) ABO3, a WRKY transcription factor, mediates plant responses to abscisic acid and drought tolerance in *Arabidopsis*. *The Plant Journal*, **63**, 417–429.

Reumann S (2013) Biosynthesis of vitamin K1 (phylloquinone) by plant peroxisomes and its integration into signaling molecule synthesis pathways. *Sub-cellular biochemistry*, **69**, 213–29.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**, 749–763.

Schmickl R, Jørgensen MH, Brysting AK, Koch MA (2010) The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evolutionary Biology*, **10**, 1–18.

Sottosanto JB, Gelli A, Blumwald E (2004) DNA array analyses of *Arabidopsis thaliana* lacking a vacuolar Na+/H+ antiporter: impact of AtNHX1 on gene expression. *The Plant Journal*, **40**, 752–771.

Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms (C Gibas, Ed,). *PLoS ONE*, **6**, e21800.

Swindell WR, Huebner M, Weber AP *et al.* (2007) Transcriptional profiling of *Arabidopsis* heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. *BMC Genomics*, **8**, 125.

Szkiba D, Kapun M, von Haeseler A, Gallach M (2014) SNP2GO: Functional analysis of genome-wide association studies. *Genetics*, **197**, 285–9.

Toledo M, Poorter L, Peña-Claros M *et al.* (2011) Climate is a stronger driver of tree and forest growth rates than soil and disturbance. *Journal of Ecology*, **99**, 254–264.

Trijatmiko KR (2005) Comparative analysis of drought resistance genes in *Arabidopsis* and rice. article Thesis. Wageningen University.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin S V (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics*, **42**, 260–263.

Turner MG, Gardner RH, O'neill R V, others (2001) *Landscape ecology in theory and practice*. Springer.

Wilczek AM, Burghardt LT, Cobb AR *et al.* (2010) Genetic and physiological bases for phenological responses to current and predicted climates. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 3129–47.

Willi Y, Määttänen K (2010) Evolutionary dynamics of mating system shifts in *Arabidopsis lyrata*. *Journal of Evolutionary Biology*, **23**, 2123–31.

Williams GC (1966) *Adaptation and natural selection*. Princeton (NJ): Princeton University Press.

Wos G, Willi Y (2015) Temperature-stress resistance and tolerance along a latitudinal cline in North American *Arabidopsis lyrata* (Z Chan, Ed,). *PLOS ONE*, **10**, e0131808.

Zhang Y, Liu Z, Khan AA *et al.* (2016) Expression partitioning of homeologs and tandem duplications contribute to salt tolerance in wheat (*Triticum aestivum L.*). *Scientific reports*, **6**, 21476.

Table 1. Number of SNPs, genes and GO terms for each environmental variable.

Env	SNPs tot	SNPs gen	Genes	Genes 5SNPs	Perc 5SNPs	GO term
bio16	5005	3425	831	127	69.66%	97
Tmin_ESp	2637	1591	556	67	57.32%	35
Sub	1308	742	321	33	42.45%	23
alpha	422	335	93	11	62.99%	15

Env environmental variables**.**

SNPs tot number of SNPs associated.

SNPs gen number of SNPs associated in genetic region.

Genes number of genes associated.

Genes 5SNPs number of genes associated that have more than 5 SNPs associated.

Perc 5SNPs percentage of associated SNPs in genes that have more than 5 SNPs associated.

GO number of GO terms associated.

Table 2. Genomic regions with more than 50 SNPs associated with environmental variables.

Env	Chr	Start	End	SNPs	Perc	Gene
bio16	1	3435415	3518282	63	1.26%	AL1G19550
bio16	2	2660327	2719097	202	4.04%	AL2G15270
bio16	2	2762362	2816602	214	4.28%	AL2G15460
bio16	2	11530672	11601244	64	1.28%	AL2G24570
bio16	2	13424791	13480687	69	1.38%	AL2G27910
bio16	2	13497672	13535329	68	1.36%	AL2G28070
bio16	2	17247881	17630257	2026	40.48%	AL2G36920
bio16	5	12420610	12447861	52	1.04%	AL5G24920
bio16	5	19083072	19121791	102	2.04%	AL5G40790
bio16	8	12045356	12064070	73	1.46%	AL8G21750
Tmin_ESp	2	12158010	12187596	66	2.50%	AL2G25500
Tmin_ESp	2	15328669	15372557	95	3.60%	AL2G31960
Tmin_ESp	3	8358784	8378481	74	2.81%	AL3G32800
Tmin_ESp	3	17608920	17622585	120	4.55%	AL3G44910
Tmin_ESp	8	11539355	11576290	220	8.34%	AL8G21140
Sub	2	2397658	2405223	51	4.91%	AL2G14660
alpha	4	18900248	18925846	55	13.03%	AL4G36300
alpha	8	2129257	2145511	95	22.51%	AL8G13590

Env environmental variables**.**

Chr chromosome.

Start start of the genomic regions.

End end of the genomic regions.

SNPs number of associated SNPs in the genomic regions.

Perc percentage of associated SNPs in the genomic regions relative to the total number of SNPs associated.

Gene gene present in the region with the highest number of associated SNPs.

Table 3. Candidate genes in regions with the highest number of associated SNPs.

Env	ID lyrata	Chr	Start	End	SNPs	ID thaliana	Desc
bio16	AL1G19550	1	3479497	3482448	17	NA	NA
bio16	AL2G15270	2	2704310	2706198	42	AT1G61110	NAC domain containing protein 25
bio16	AL2G15460	2	2788175	2793670	78	AT1G60995	S3 self-incompatibility locus-linked pollen protein
bio16	AL2G24570	2	11559302	11560652	11	AT1G66600	ABA overly sensitive mutant 3
bio16	AL2G27910	2	13443303	13454344	15	AT1G68890	2-oxoglutarate decarboxylase/hydro-lyase/magnesium ion-binding protein
bio16	AL2G28070	2	13517749	13521362	40	AT1G69020	Prolyl oligopeptidase family protein
bio16	AL2G36920	2	17322232	17329664	328	AT1G76780	HSP20-like chaperones superfamily protein
bio16	AL5G24920	5	12440001	12442900	36	AT3G46190	TRAF-like family protein
bio16	AL5G40790	5	19112107	19114395	28	NA	NA
bio16	AL8G21750	8	12053095	12057696	18	NA	NA
Tmin_ESp	AL2G25500	2	12161888	12188505	65	AT1G67120	midasin-like protein
Tmin_ESp	AL2G31960	2	15331522	15332472	39	AT1G72290	Kunitz-protease inhibitor
Tmin_ESp	AL3G32800	3	8378106	8380315	3	AT3G19390	Granulin repeat cysteine protease family protein
Tmin_ESp	AL3G44910	3	17619295	17624192	43	NA	NA
Tmin_ESp	AL8G21140	8	11565959	11568937	72	NA	NA
Sub	AL2G14660	2	2403991	2406585	14	AT1G61590	Protein kinase superfamily protein
alpha	AL4G36300	4	18915610	18917714	27	AT2G39000	Acyl-CoA N-acyltransferases (NAT) superfamily protein
alpha	AL8G13590	8	2132609	2142052	82	NA	NA

Env environmental variables**.**

ID lyrata gene identifier of *A. lyrata*

Chr chromosome.

Start start of the genomic regions.

End end of the genomic regions.

SNPs number of associated SNPs in the gene.

ID thaliana gene identifier of the homologous gene in *A. thaliana.*

Desc description of the *A. thaliana* gene.

Table S1. *A. lyrata* populations with coordinates and values of climatic variable.

Table S2. SNPs associated with the environmental variables.

Table S3. Overrepresented GO terms for each environmental variables.

Table S4. Genes associated with the climatic variables.

The supplementary tables are available at https://github.com/fraca/Thesis_data.

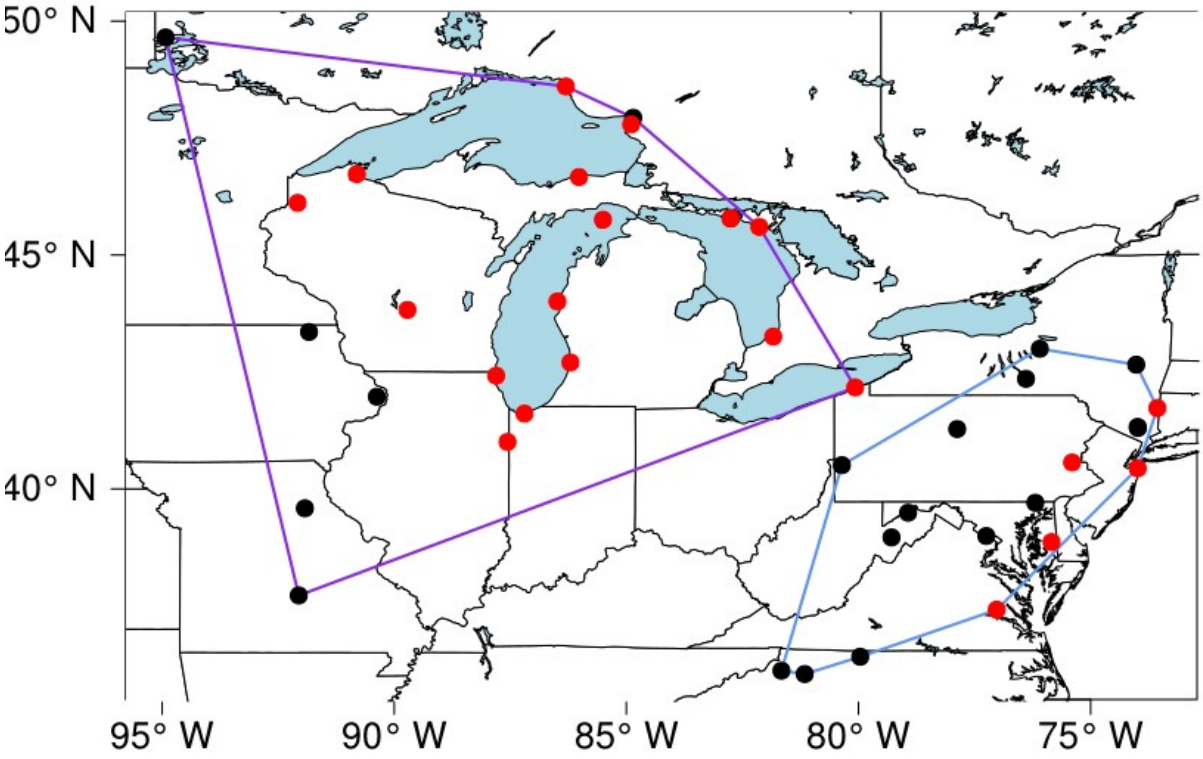


Figure 1. Map of the outcrossing *A. lyrata* populations included in this study. The minimum convex polygon hull of the Eastern cluster is indicated in blue, the one of the Western cluster is indicated in purple. The black circles represent the populations at rocky sites and the red circles those at sandy sites.

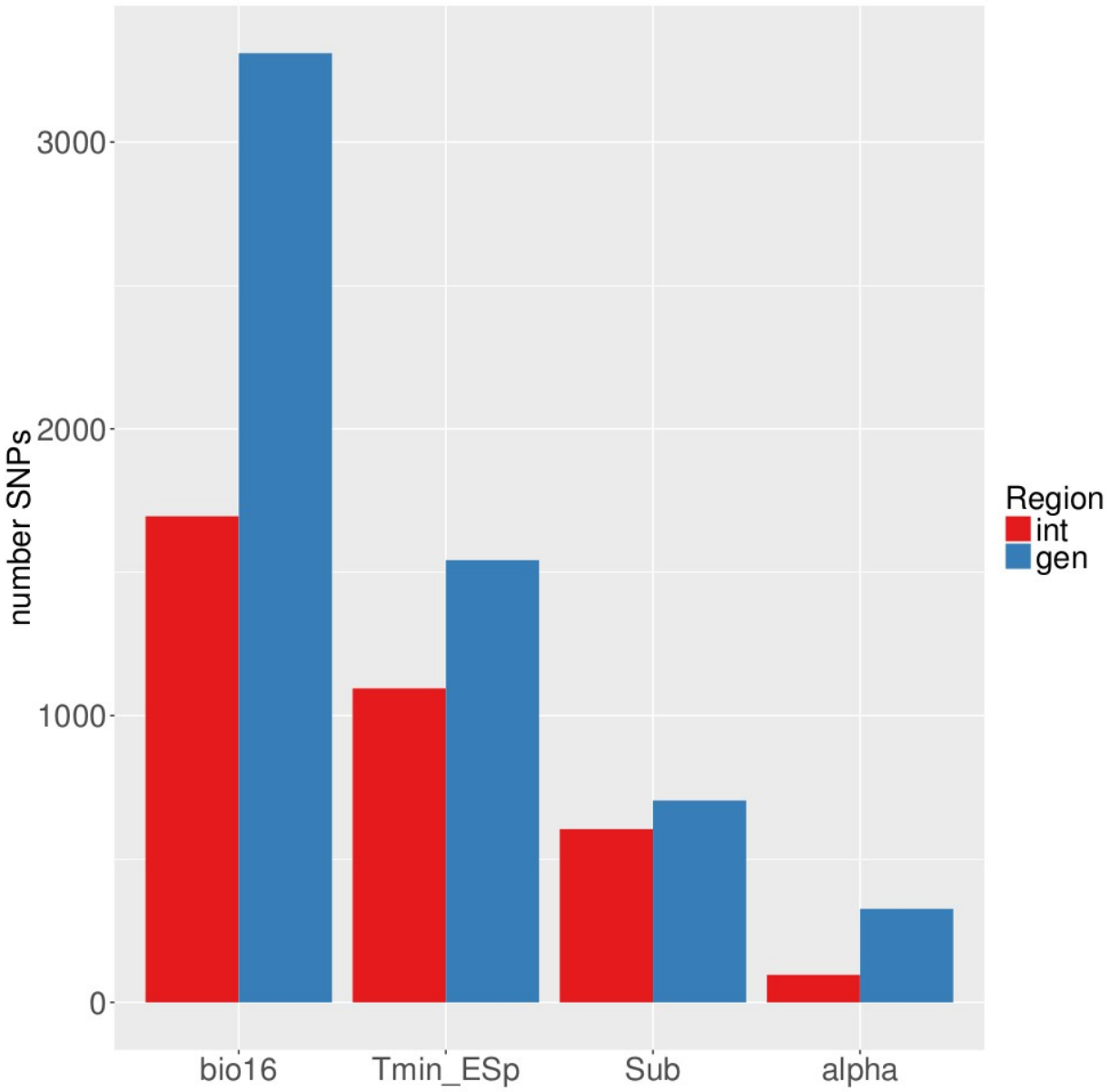


Figure 2. Associated SNPs and their distribution across intergenic and genic regions of the nuclear genome. The red bars are the SNPs in intergenic regions, the blue bars are the SNPs in genic region, for the four environmental variables of precipitation wettest quarter, bio16; minimum temperature early spring, Tmin_ESp; substrate type, Sub; and the Priestley-Taylor coefficient of water availability, alpha. The environmental variables are sorted by the number of SNPs that were found to be associated with them.

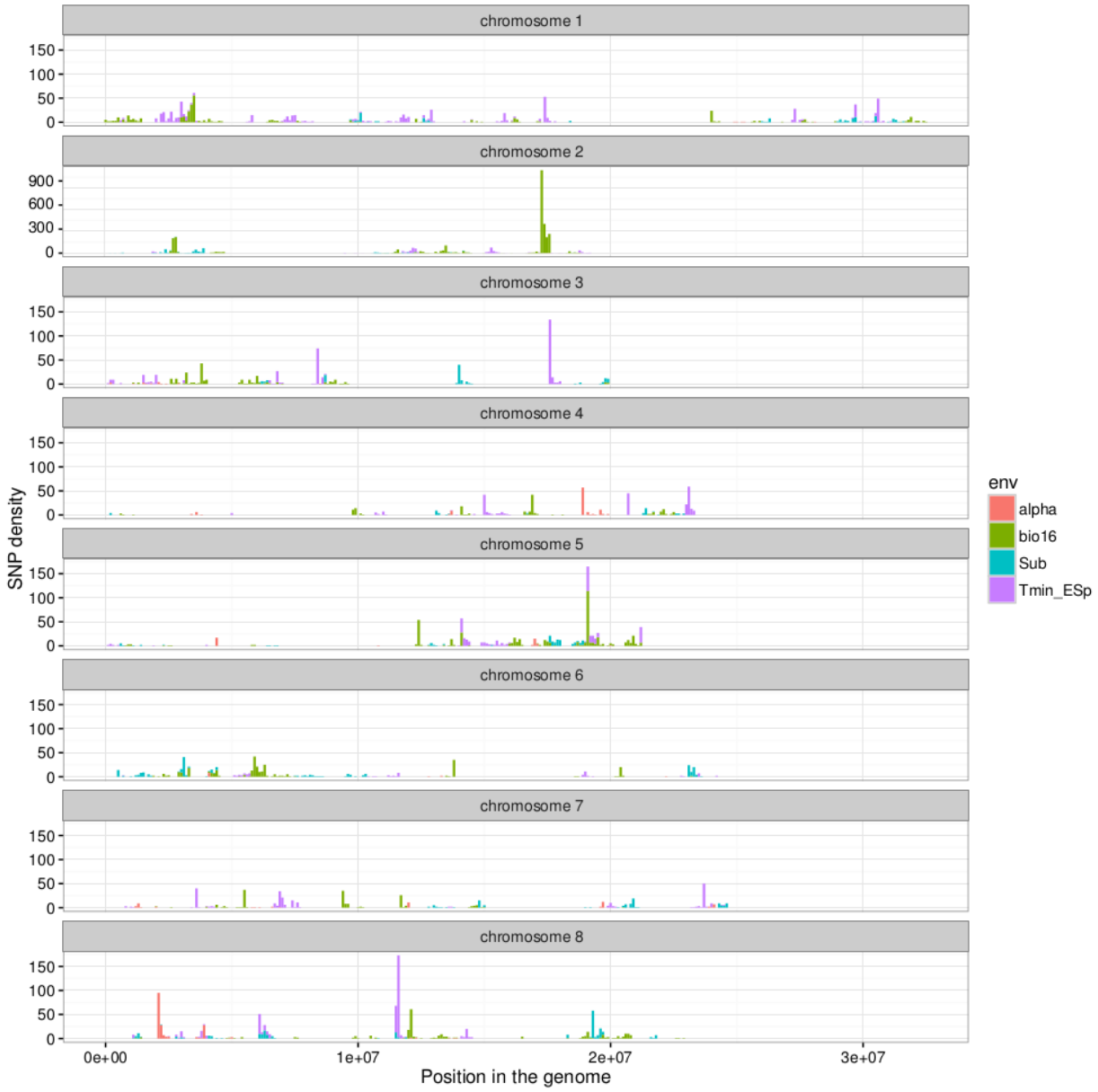


Figure 3. Histogram of the significant SNPs of the association study distributed across the chromosomes. In pink: SNPs associated with the Priestley-Taylor coefficient, alpha. In green: SNPs associated with precipitation during the wettest quarter, bio16. In blue: SNPs associated with substrate type, Sub. In purple: SNPs associated with the minimum temperature early spring, Tmin_ESp. The vertical bars represent the SNP density in windows of 100,000 base pair.

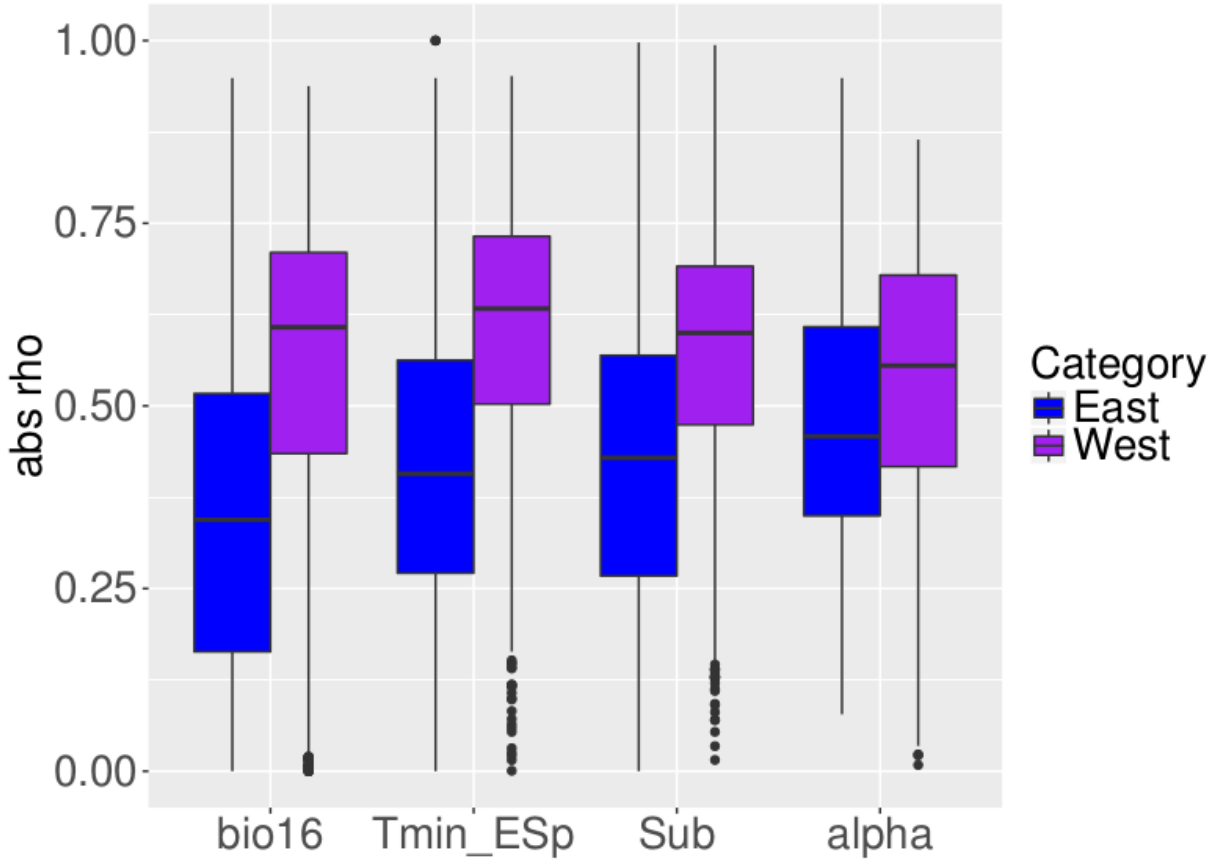


Figure 4. Boxplot of the absolute Spearman correlation coefficients. Values of the Eastern genetic cluster are indicated in blue, those for the Western genetic cluster in purple.

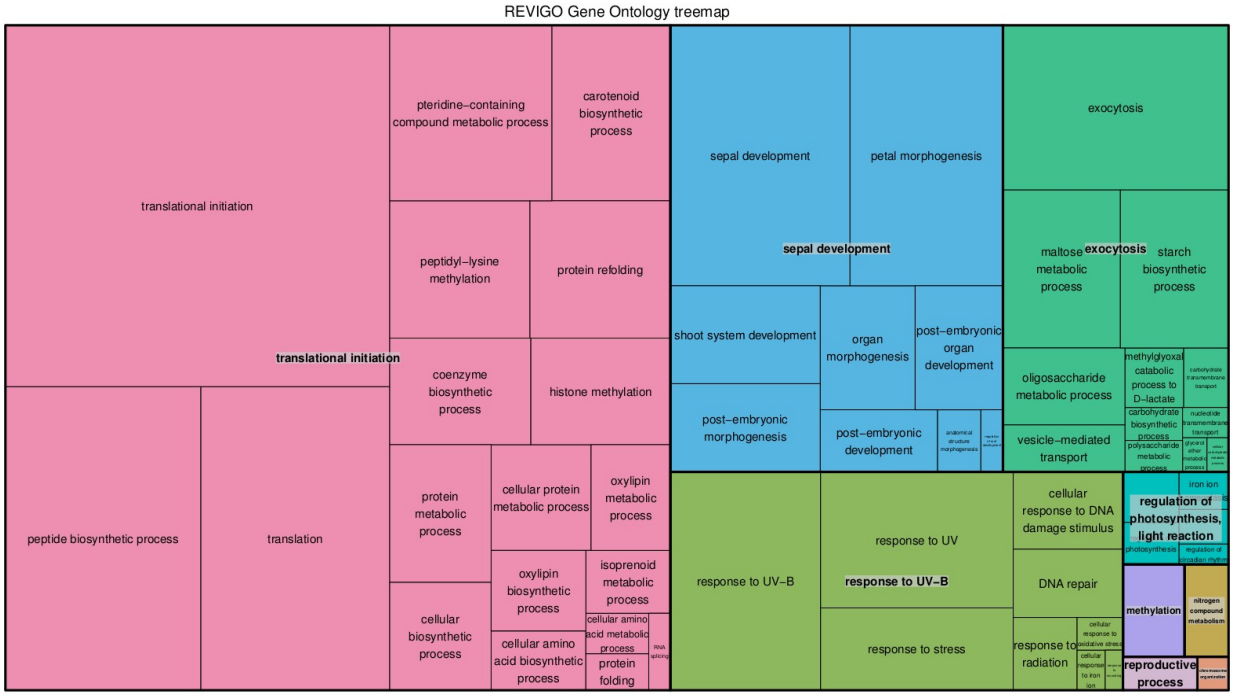


Figure 5. Tree map of GO terms linked to the environmental variable of precipitation during the wettest quarter of the year (bio16). Each rectangle represents an enriched GO term. The size of the rectangles reflects the FDR value of the snp2go analysis.

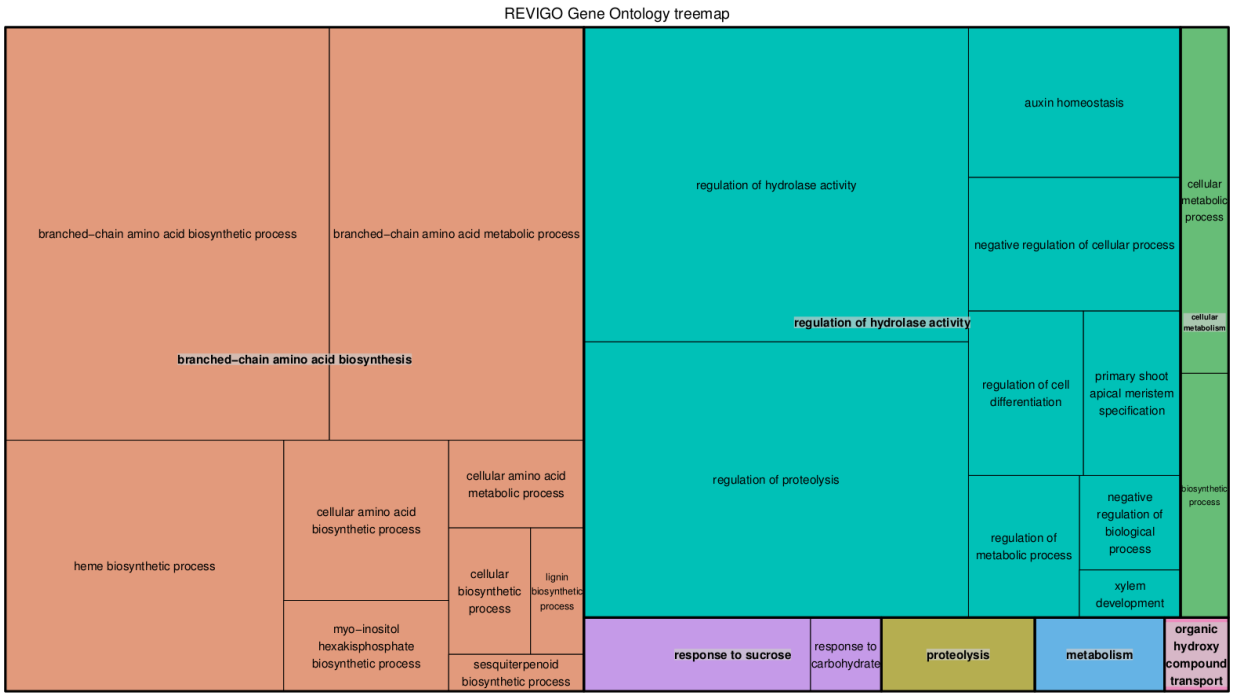


Figure 6. Tree map of GO terms linked to the environmental variable of minimum temperature early spring (Tmin_ESp). Each rectangle represents an enriched GO term. Size of the rectangles reflects the FDR value of the snp2go analysis.

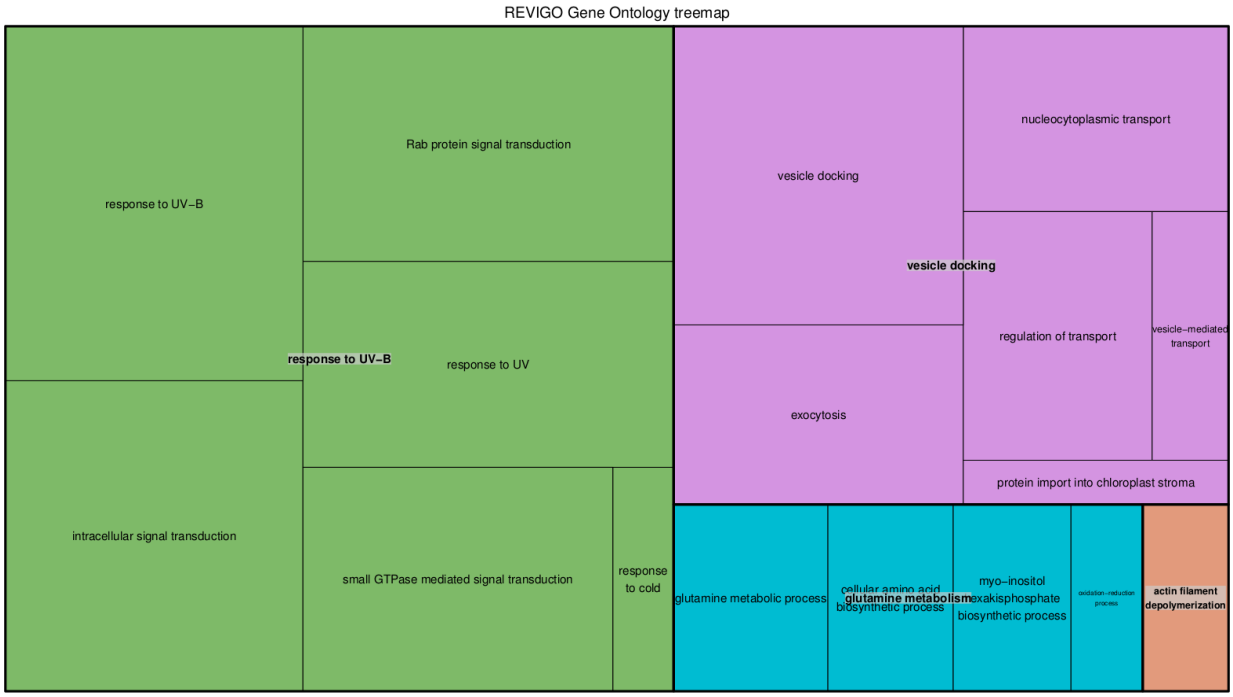


Figure 7. Tree map of GO terms linked to the environmental variable of substrate type (Sub). Each rectangle represents an enriched GO term. Each rectangle represents an enriched GO term. The size of the rectangles reflects the FDR value of the snp2go analysis.

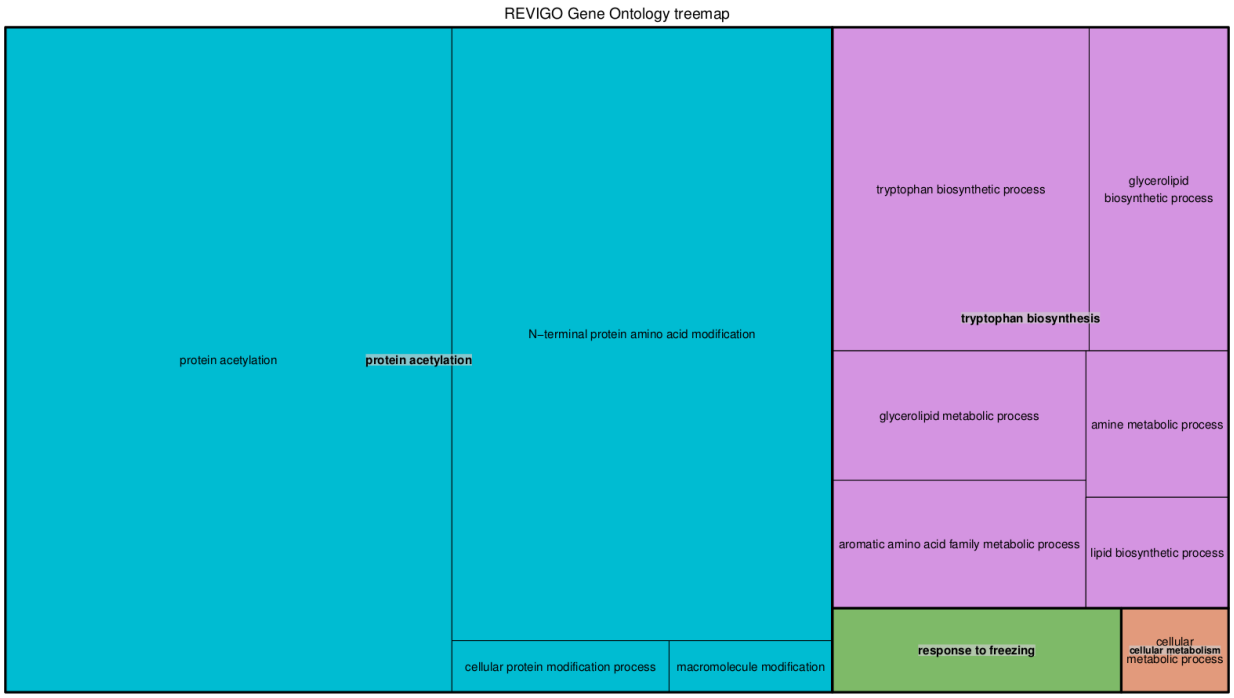


Figure 8. Tree map of GO terms linked to the environmental variable of the Priestley-Taylor coefficient for general water availability (alpha). Each rectangle represents an enriched GO term. The size of the rectangles reflects the FDR value of the snp2go analysis.

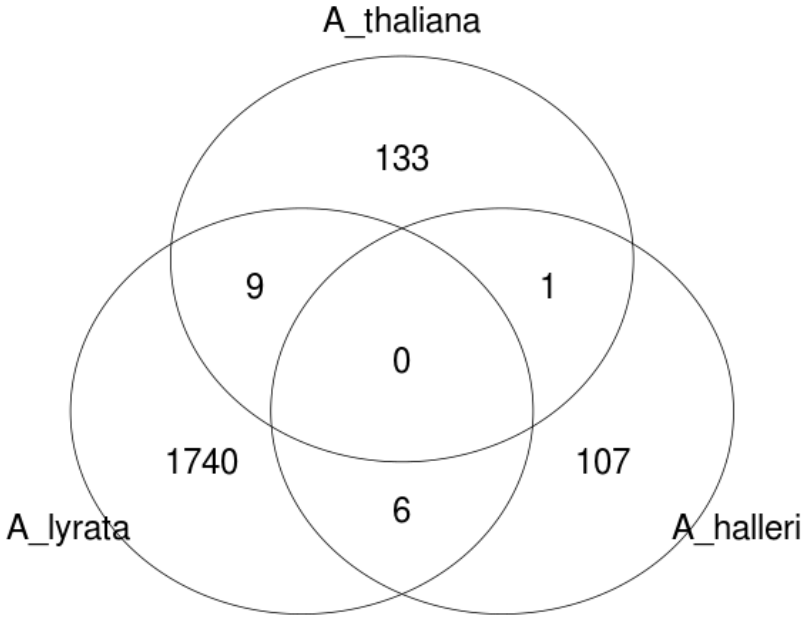


Figure 5 Venn diagram of the candidate genes of the EEA executed in three *Arabidopsis* species. *A. lyrata*, of this study, *A. thaliana* (Hancock et al. 2011) and *A. halleri* (Fischer et al. 2013). Few candidate genes overlapped between pairs of studies/species, and none overlapped across the three species.

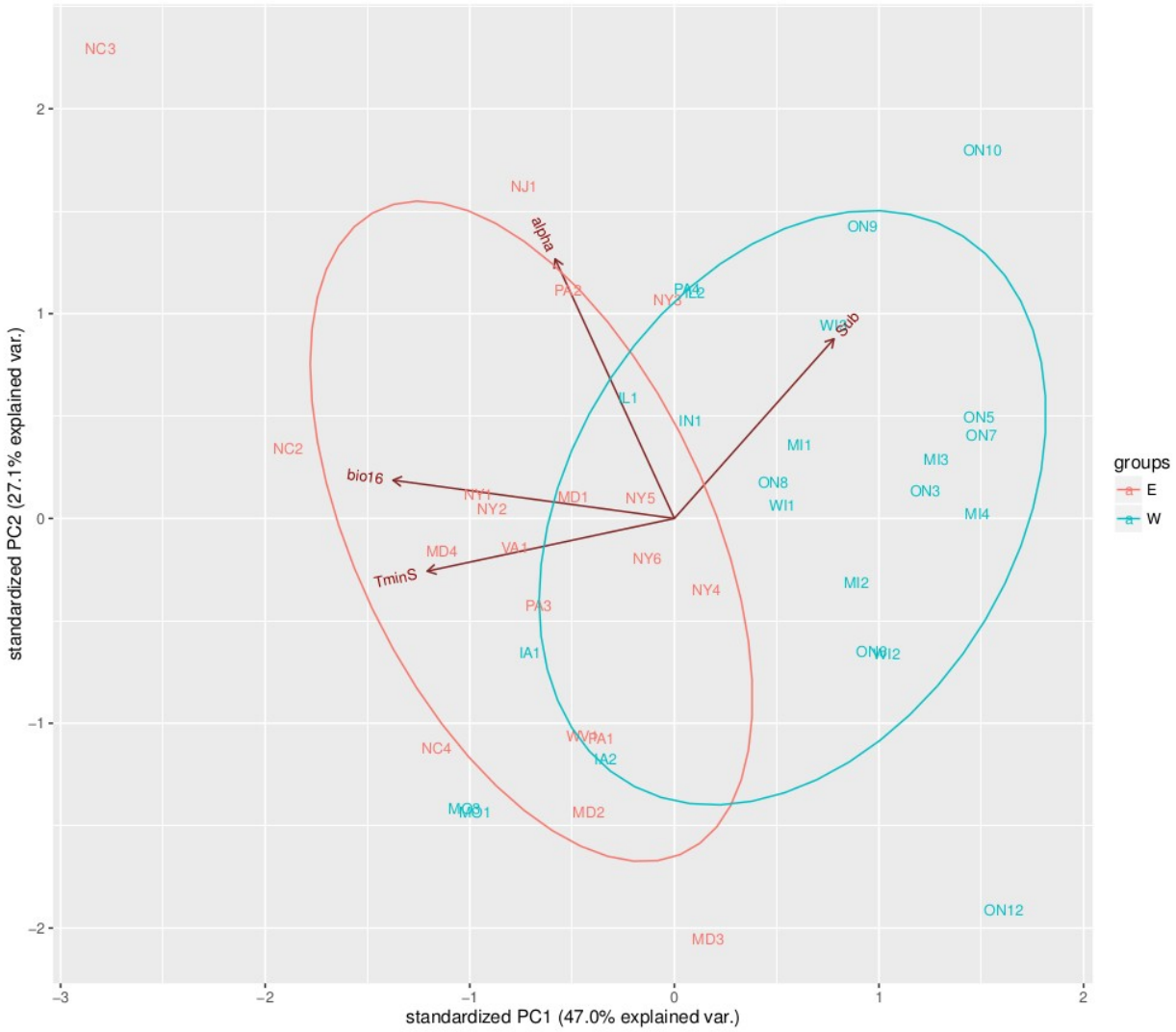


Figure S1. Principal component analysis on the environmental variables grouped by genetic cluster, East (E) and West (W).

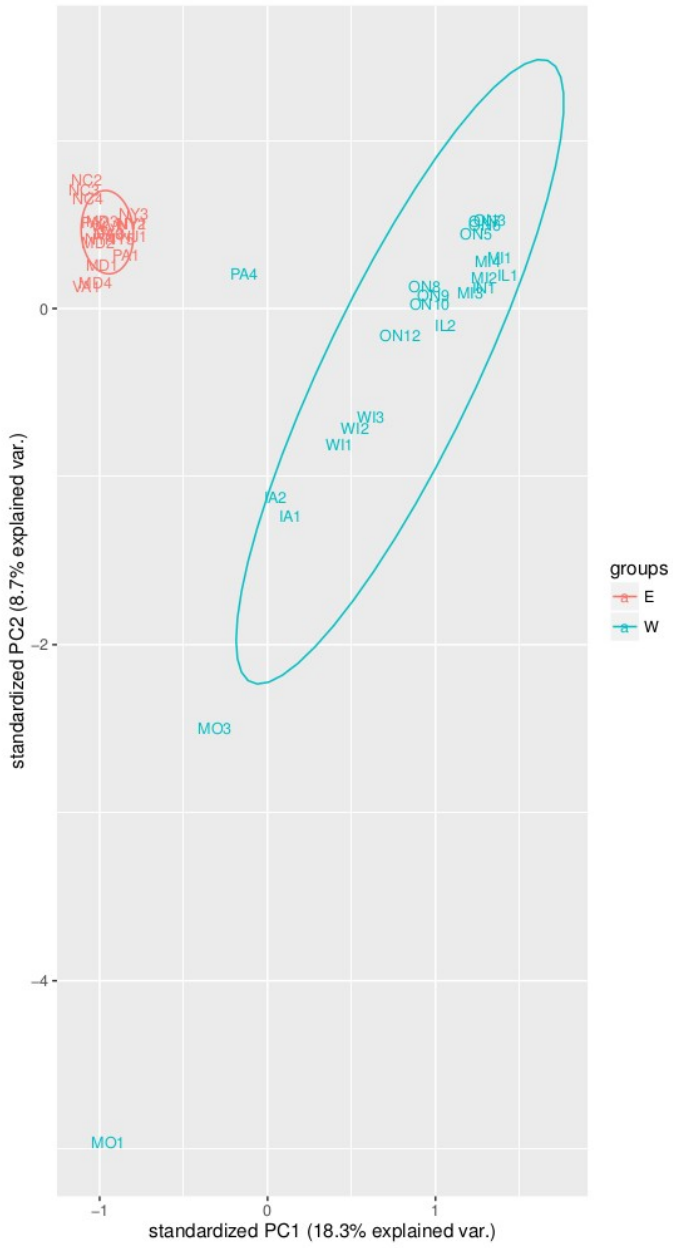


Figure S2. Principal component analysis on the SNP frequencies grouped by genetic cluster, East (E) and West (W).

DISCUSSION

In this thesis I analyzed the genetic diversity and the genetic basis of adaptation to different environmental conditions in the plant *Arabidopsis lyrata* spp. *lyrata*. I analyzed 52 populations of *A. lyrata* that covered the whole species distribution range (Schmickl *et al.* 2010; Paccard *et al.* 2016). The estimation of genetic data was done with a Pool-seq approach (Schlötterer *et al.* 2014), in which pools of 25 individuals for each population were sequenced.

In the first chapter, I compared the accuracy of the estimation of the SNP frequencies detected by Pool-seq with individual-based Genotyping-By-Sequencing (GBS) (Elshire *et al.* 2011). The key parameters I tested for impacting accuracy of Pool-seq were the pool size and the sequencing depth. I demonstrated that pools of 25 individuals with a sequencing depth of 100× produce accurate allele frequency estimates for common SNPs with a minor allelic frequency above 0.05 (Fracassetti *et al.* 2015).

In the second chapter, I reconstructed a relatedness tree based on the SNPs frequencies with the program Treemix (Pickrell & Pritchard 2012), that confirmed the presence of two genetic clusters: an Eastern cluster along the Appalachian Mountains and a Western cluster west of Lake Erie with signature of past gene flow between them. These genetic clusters had already been detected in previous studies (Hoebe *et al.* 2009; Willi & Määttänen 2010). Based on a population relatedness tree, I could reconstruct the colonization history of the species at the end of the last glacial maximum. In a next step, I analyzed the predictors of within-population genomic diversity for different genomic regions (intergenic and coding). The main drivers of genomic diversity at the whole-genome level were mating system (selfing compared to outcrossing) and historic range dynamics after the last glaciation maximum (LGM). Historic demographic processes before LGM and past gene flow between clusters had a minor impact on genomic diversity in intergenic regions.

As these drivers are predicted to predominantly determine neutral evolution and are unlikely to directly impose selection, it means that genomic diversity is strongly affected by genetic drift.

In the third chapter, genomic diversity was linked also to the species distribution. In *A. lyrata* the geographic range limits reflect the ecological niche limits at the latitudinal range margins (Lee-Yaw et al. Submitted). In other word the suitability of sites declines towards the latitudinal edge of the species distribution. The variables that mainly contributed to predicting the niche limits were average minimum temperature during the early spring and precipitation during the wettest quarter of the year, and marginally by the general moisture availability. Extreme temperatures and levels of precipitation are harmful to species persistence during this period.

In my last chapter, I performed an environmental association analysis (EAA) to reveal SNPs and genes associated with environmental variables that determine the niche limits and substrate type (sandy and rocky sites). Most of the SNPs and genes clustered in particular genomic regions. Particularly, 40% of the SNPs associated with the precipitation during the wettest quarter of the year are located on chromosome 2 between the position 17247881 and 17630257. The association between the SNPs frequencies and the environmental variables was higher in the Western genetic cluster populations, which were more genetically differentiated.

Overall, my thesis contributed to our understanding of the drivers of within-population genomic diversity necessary for adaptive evolution – the role of the species history versus local factors, the investigation of important niche variables that determine a species distribution and the genes linked to adaptation to these variables within the general distribution of a species. For all these analysis, Pool-seq turned out to be an effective approach to perform population genomic analysis.

REFERENCES

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Ferretti L, Ramos-Onsins SE, Pérez-Enciso M (2013) Population genomics from pool sequencing. *Molecular Ecology*, **22**, 5561–5576.

Fracassetti M, Griffin PC, Willi Y (2015) Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata* (U Melcher, Ed,). *PLOS ONE*, **10**, e0140462.

Hoebe PN, Stift M, Tedder A, Mable BK (2009) Multiple losses of self-incompatibility in North-American *Arabidopsis lyrata*?: Phylogeographic context and population genetic consequences. *Molecular Ecology*, **18**, 4924–4939.

Paccard A, Van Buskirk J, Willi Y (2016) Quantitative genetic architecture at latitudinal range boundaries: reduced variation but higher trait independence (CG Eckert, JL Bronstein, Eds,). *The American Naturalist*, 000–000.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. (H Tang, Ed,). *PLoS Genetics*, **8**, e1002967.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**, 749–763.

Schmickl R, Jørgensen MH, Brysting AK, Koch MA (2010) The evolutionary history of the Arabidopsis lyrata complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evolutionary Biology*, **10**, 1–18.

Willi Y, Määttänen K (2010) Evolutionary dynamics of mating system shifts in *Arabidopsis lyrata*. *Journal of Evolutionary Biology*, **23**, 2123–31.