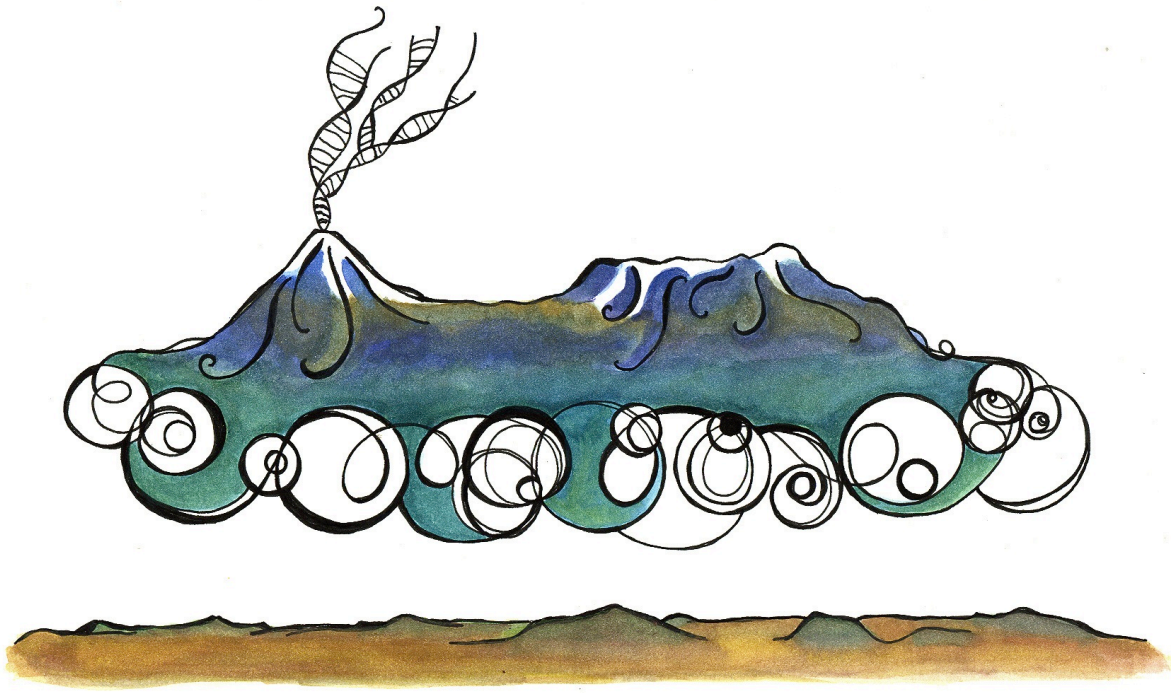
Landscape genomics of tropical high altitude plant species



Transmexican sky-islands, ticatla 2014

Alicia Mastretta-Yanes

A thesis submitted for the degree of Doctor of Philosophy
School of Biological Sciences
University of East Anglia, UK

September 2014

In memory of Arturo Maldonado, who first inspired me to become a scientist, and of Godfrey Hewitt, whose work lured me into studying biodiversity distribution in space and time

Dedico esta tesis a Julia Carabias, a su pasión por el mundo natural y a su integridad como ser humano.

Abstract

Changes to species distributions involve demographic processes that occur over generations and affect allele frequencies within populations, leading to patterns of genetic structuring. The specific genetic structuring patterns that will be observed as a consequence depend on explicit geographic features, such as topography and latitude. Over the first decades of phylogeography, the effect of climate history and geography on species genomes was examined at low resolution with DNA sequences and other traditional molecular markers. However, during the last five years it has become feasible to obtain genomic data for non-model organisms and large sample sizes.

The present thesis spans the transition years between phylogeographic studies being restricted to low resolution molecular markers, and new methods facilitating the generation of genomic data for non-model species. As such, this thesis focuses on two main points. First, on the methodological aspects of utilising double digest RAD-seq (ddRAD) for individual-based population genetics and phylogeography of plant species. Second, on applying the obtained data to examine one of the classic, but as yet not fully explained, biodiversity patterns: the biodiversity excess within tropical mountains.

The main contributions of this thesis at the methodological level are: (1) demonstrating the utility of DNA replicates for the estimation of genotyping error and optimisation of *de novo* assembly; (2) proposing a method for identifying paralogous loci resulting from recent gene duplications; and (3) showing that such loci provide a measure of population differentiation.

Regarding the drivers of biodiversity excess within tropical mountains, I used landscape genomic analyses and ddRAD data to examine two plant species from the alpine grasslands of the Transmexican Volcanic Belt. As a main result, this thesis supports from a population-level perspective that tropical mountains: (1) allow for long-term *in situ* population persistence; and (2) promote population differentiation as a function of topographic isolation.

Supervisor details

Main supervisor:
Dr Brent Emerson
Island Ecology and Evolution Research Group,
Instituto de Productos Naturales y Agrobiología (IPNA-CSIC),
C/Astrofísico Francisco Sánchez 3,
La Laguna,
Tenerife (Canary Islands, 38206),
Spain

Secondary supervisor:
Dr Tove H Jørgensen
Department of Bioscience
Aarhus University
8000 Aarhus C
Denmark

Acknowledgements

Esta tesis lleva mi nombre como yo llevo el mío cuando a la vez soy la genética y las enseñanzas de tantos y tantas otras. Un profundo gracias a quienes hicieron posible mi doctorado:

Brent Emerson, mi asesor principal, por responder el primer correo electrónico con el que empezó este proyecto y demostrar a lo largo de cuatro años que es un científico y ser humano fuera de serie. Yo quería trabajar con Brent y en UEA, pero fue fundamental para mi vida, personal y académica, que Brent sugiriera que realizáramos un proyecto diseñado por mí y que saliera de la sombra inmediata de su investigación. Sin esa apertura (que no es regla en el mundo académico) esta tesis no se trataría de la Faja Volcánica Transmexicana y mi vida, creo, hubiera derivado por caminos distintos. El proyecto también le debe a Brent el componente RAD. "I think we should do this for your thesis"- dijo el día que discutimos el artículo de Kevin Emerson et al. (2010), *uno de los primeros estudios que utilizaron RAD en un contexto filogeográfico. Él no sabía cómo utilizar el método, ni yo, ni teníamos el dinero, pero sabíamos para qué queríamos generar ese tipo de datos. Aprendí que ese salto al agua fría y oscura es el mejor modo de hacer ciencia y que una vez que comienzas a moverte, el agua se siente fresca y la oscuridad no puede sostenerse ante el escudriño de nuestra mirada. Brent es un gran asesor porque enseña sin decirte que lo está haciendo y porque concilia en una vida posible a un científico admirable y a un hombre feliz y cálido. Ojalá, de algún modo, todos aprendamos de él.*

Daniel Piñero, asesor durante mis estudios de Biología y gran amigo, por convertirse en un colaborador esencial para la existencia de esta tesis. Años después de haber dejado su laboratorio siguió (sigue) velando mi desarrollo académico y es gracias a su iniciativa que conseguimos financiar la secuenciación RAD. El Doctor π es otro de esos científicos de altísima calidad humana que muchos creen inexistentes pero con quienes yo he tenido la fortuna de trabajar. Aquí agradezco su contribución inmediata a la tesis, pero sus aportaciones a mi vida van mucho más allá.

Mis dos progenitores por heredarme su entusiasmo y ser la mejor inspiración. Me dedico a la biología, pero mis cimientos son el arte, la historia, el periodismo y el compromiso social que aprendí de ellos.

Mis dos sisternitas por iluminar mi vida y perseguir sus propios horizontes. Todo lo que hacen y su forma de ser me llena de orgullo y me impulsa.

Las niñas, Migue y Eo, por mantener el ánimo de todos. Al resto de mi farmacia, sobretodo a mis primates Dani y Cris, por ser la antología de personajes en quién me refugio.

El pequeño ejército de amigos, colaboradores y personal de la CONANP por el que fue posible mi trabajo de campo: José Ramón Pérez Pría Kasusky, Sergio Mastretta Guzmán, Juan José Ramírez Lerma, Roberto Ortíz Flores, Sara

*Straffond, Rocío Aguilar Fernández, Ana Mastretta Yanes, Tania Salazar Armenta, José Armas Aramburu, Francisco David Rubio Greathouse, Felipe Quintero Bazán, Maria José Leal Fernández, Sergio Nicasio Arzeta, Mariana Hernández, Santiago Salas Fadul, Alejandra Ortíz Medrano, Laura Figueroa Corona, Josue García Amaya, Genaro Mondragón, Epifanio Hernández Juárez, Katia Lemus Ramírez, Rubén Marroquín, Lorenzo Otlica Reyes, Eliud Vergara Sánchez, Jorge Osorno Doroteo, Zanón Cabrera Ibarra, Adolfo Sánchez Pérez, Agustín Aguirre, Martha Andraca, Guadalupe Andraca, Ricardo Flores, Alejandra Moreno Letelier, Ana Wegier Brioulo, Carlos Nieto Irigoyen***,*** *Maritriny Serrano Cuatláyotl, Rodrigo León Pérez* y *Miguelángela.*

Las Montessoris por la locura y pasión de su amistad. Mención honorífica a Marcela Nieto por sus palabras en el momento justo.

Los Naturos y ex-Naturos por mantener el ambiente de oficina a través de los océanos y los objetivos comunes a través de especializaciones divergentes.

Eleonor Cortés y Alonso Zamora por ser mis nahuales.

Libertad Arredondo por compartir su objetividad y fuerza, por hacer posibles los últimos meses y por convertirse en una gran colaboradora y amiga.

Camille Pitteloud, Nancy Galvéz y Adriana Uscanga por ser víctimas de mis primeros intentos en la enseñanza. Me han hecho aprender mucho y es un gusto trabajar con ustedes.

*Al Equipo Naranja por darle sazón (picante) a la recta final, sobretodo al Guayito por los tiriti*tititi.*

This thesis has my name as my name as I have mine while being the genes and the teachings of so many others. A deep thank you to all those who made possible my Ph.D:

My supervisors Brent Emerson (hey Alicia you mentioned him already! I know, can't help it… he is such a great guy) and Tove Jørgensen. Brent coached me first of all by being a burning source of scientific inspiration, hard work and good life, and then by devoting many time to in-depth discussion and correction of my work, and lastly by keeping an open spirit to the not few unusual ideas and situations I ended up in. Tove, of all the people involved in this project, was who endured me with endless patience and who provided a very detailed review of my work. *Tak*. Meeting her was another great surprise, and I hope (to her dismay) life will let me stick to her for long.

My collaborators in Switzerland and Mexico: Nadir Alvarez, Nils Arrigo, Daniel Piñero, Alejandra Moreno-Letelier and Sergio Zamudio.

Nadir Alvarez and his lab for the time I expended in Lausanne. Tomasz Suchand, Nils Arrigo and Camille Pitteloud, plus Tania Wyss and Catherine

Berney by parapatric effect. With a special *gracias* to Nadir for receiving me in his lab and informally helping so much to my development as a scientist and a person. Alas, he is one of the other great guys I have the luck to end up working with. Also *grand merci* to Nils for coaching my first steps on the bioinformatics learning hike.

The Friday Poetry group at Unthank road for sharing their art and keeping the scientist alive. Special thanks to Philip Wilson for the long walks and The Drawings and Poems Book, and to Gareth Jones for cooking.

My officemates. Emma Barrett and Claudia Fricke whose analytical discipline inspired me, and Lewis Spurgin and Damian Smith, who survived me and helped to get done my never ending long-distance paperwork.

The housemates with whom sharing a ceiling became a crucial stage of my life. Hiroko Furukawa, I still feel you are only one corridor apart. Becky Laidlaw, keep looking for the lettuce, I will always be hiding it. The Holmans (Rachel, Paul & Harry), I'm forever grateful for the time under your adoption. Chris, thanks for all the cositas buenas.

James Kitson, Christiana Faria, Gerardo Hernández, Conrad Gillett and Heriberto Hernández for accepting a little bit of photosynthesis in the insect world of the Emerson's Lab.

All of BIO research students and friends, who proved true the hypothesis that Norwich is inhabited by hobbits. Special mention to Alyson Lumley, Matt Dickerson, Lewis Spurgin, Karl Philips and Manuel Shärr for corridor dances and life chats; to Sarah Marburger for paralogs wondering; and to Dave Wright and Karl Philips for de-spanishing my writing and providing helpful last minute advice.

BIO faculty for their advice and inspiring work. Specially to Jenny Gill, Matt Gage, Douglas Yu, David Richardson, Tracey Chapman and Richard Davies.

Francisco Pina for becoming my personal Stacksoverflow.

Graeme Wigg, Richard Evans-Gowing, Emma Hall, Jennifer Phillips, Darren Spauls and the rest of the staff that kept BIO running flawlessly (except for certain lift and certain automatic doors).

Finally, I would also like to thank my two examiners, Martin Taylor and Roger Butlin, for accepting to review and discuss this work.

This thesis was also possible thanks to the kind funding and professional help of:

Consejo Nacional de Ciencia y Tecnología (CONACYT) for my PhD scholarship and further funding for the ddRAD project.

Company of Biologists and John and Pamela Salter Trust for partially funding fieldwork.

Rosemary Grant Student Research Award from SSE for funding cpDNA sequencing.

Biól. Verónica Juárez, Dr. Hilda and Dr. David Gernandt for facilitating MEXU Herbarium specimens and Dr. Socórro González Elizondo for providing samples of juniper outgroups.

Biól. Rafael Torres for taxonomical help with *Eryngium.*

Oscar Trejo from DGFS and the Areas Naturales Protegidas staff for help with sampling permissions paperwork.

Dylan Edwards for facilitating UEA support for visas and internships paperwork.

Paul Wright for never failing IT help.

Grace - the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at UEA. Specially Leo Earl and Chris Collins for their wiliness to help and efficiency.

Contents

Author Contributions

At the time of submission, two of four data chapters presented in this thesis are published and one has been submitted for review. I am lead author on all manuscripts and I have made by far the largest contribution to the work presented in this thesis. Below, I provide a citation for each data chapter specifying my contributions.

Chapter 2. Mastretta-Yanes, A., A. Moreno-Letelier, G. M. Hewitt, T. H. Jorgensen and Brent C. Emerson. Biodiversity in the Mexican highlands and the complex interaction of geology, geography and climate at a tropical latitude. (submitted to *Journal of Biogeography*).

- AMY role in conceiving the review, summarizing most of the reviewed references, drawing figures and writing the manuscript.

Chapter 3. Mastretta-Yanes, A., N. Arrigo, N. Alvarez, T. H. Jorgensen, D. Piñero, and B. C. Emerson. 2014. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources.* doi: 10.1111/1755-0998.12291.

-AMY role in conceiving and designing the study, performing fieldwork, performing labwork, leading data analyses and drafting manuscript.

Chapter 4. Mastretta-Yanes, A., S. Zamudio, T. H. Jorgensen, N. Arrigo, N. Alvarez, D. Piñero, and B. C. Emerson. 2014. Gene duplication, population genomics and species-level differentiation within a tropical mountain shrub. *Genome Biology and Evolution*. doi: 10.1093/gbe/evu205

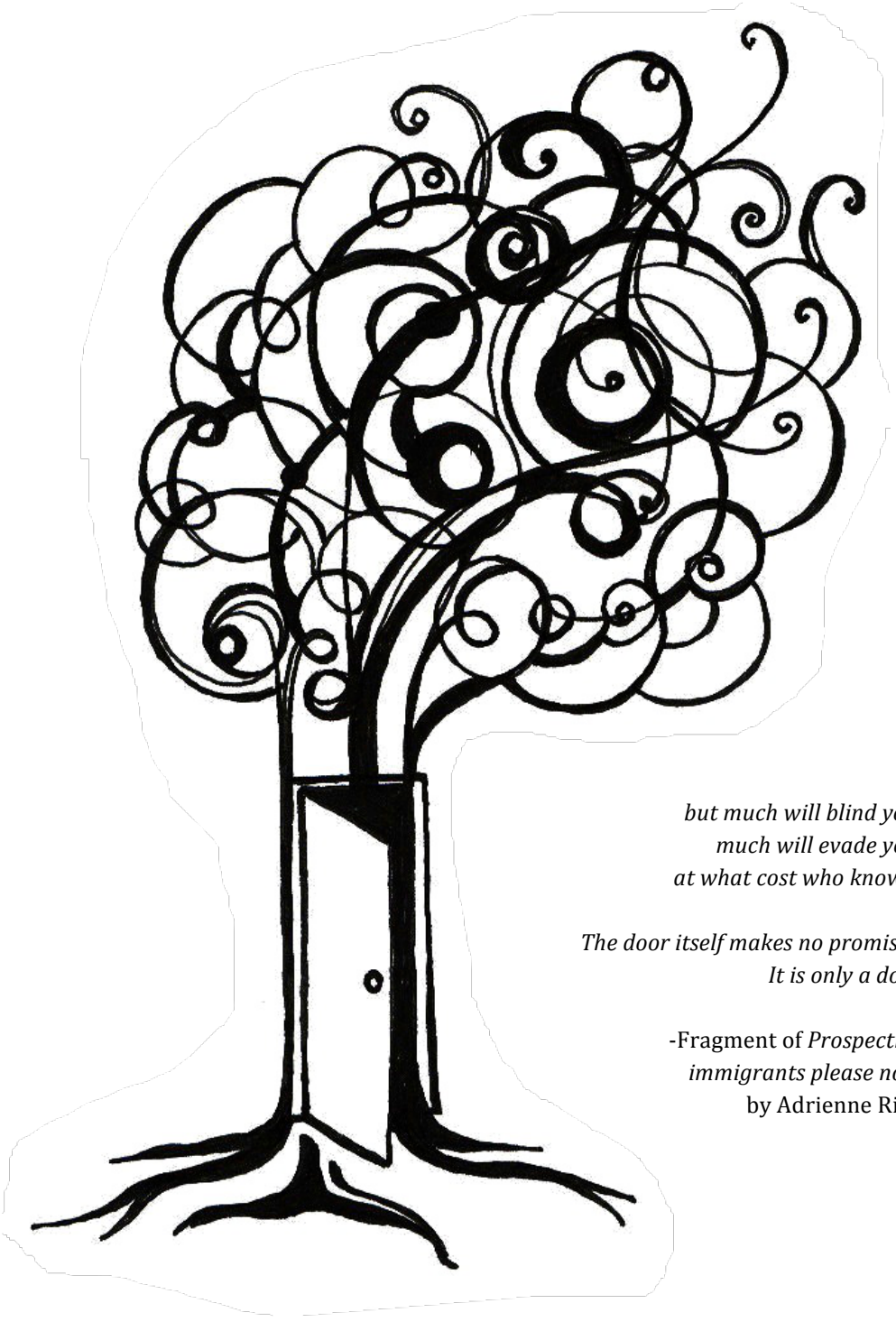
-AMY role in conceiving and designing the study, performing data analyses and writing the manuscript.

Chapter 5. Mastretta-Yanes, A., A. Moreno-Letelier, T. H. Jorgensen, N. Arrigo, N. Alvarez, D. Piñero, and B. C. Emerson. (in prep.).

-AMY role in conceiving and designing the study, leading data analyses and writing the manuscript.

CHAPTER 1

General Introduction



but much will blind you,
much will evade you,
at what cost who knows?

The door itself makes no promises.
It is only a door

-Fragment of *Prospective immigrants please note*
by Adrienne Rich

Idea, ticatla 2014

Changes to species distributions involve demographic processes that occur over generations and affect allele frequencies within populations, leading to patterns of genetic structuring (Avise *et al.* 1987; Hewitt 1996). Climate and geological history are the geophysical phenomena that drive species distribution changes, but the specific phylogeographic patterns that will be observed as a consequence depend on explicit landscape features, such as topography and latitude. Understanding how genetic variation is structured as a function of landscape history is relevant for the broader understanding of how diversity is distributed at the species and community levels (Emerson & Hewitt 2005; Vellend & Geber 2005; Papadopoulou *et al.* 2011). It may also help to define geographic areas relevant for conservation, not only due to the biodiversity that they currently hold, but also due to certain spatial-historical characteristics that make them important for the promotion and persistence of biodiversity in the long term (Carnaval *et al.* 2009). Such areas have been globally identified to be comprised, to a great extent, by tropical mountains (Fjeldså *et al.* 1999, 2012; Sandel *et al.* 2011). However, studies examining diversification and long-term persistence of biodiversity in tropical mountains have focused mostly on species distributions and coarse continental data, leaving a knowledge gap at the level of landscape and population differentiation. This thesis aims to contribute to closing this knowledge gap by examining the role of topography and climate history on shaping the genetic structuring of timberline-alpine grassland plants from a set of high altitude tropical mountains: the Transmexican Volcanic Belt (TMVB, Fig. 1.1).

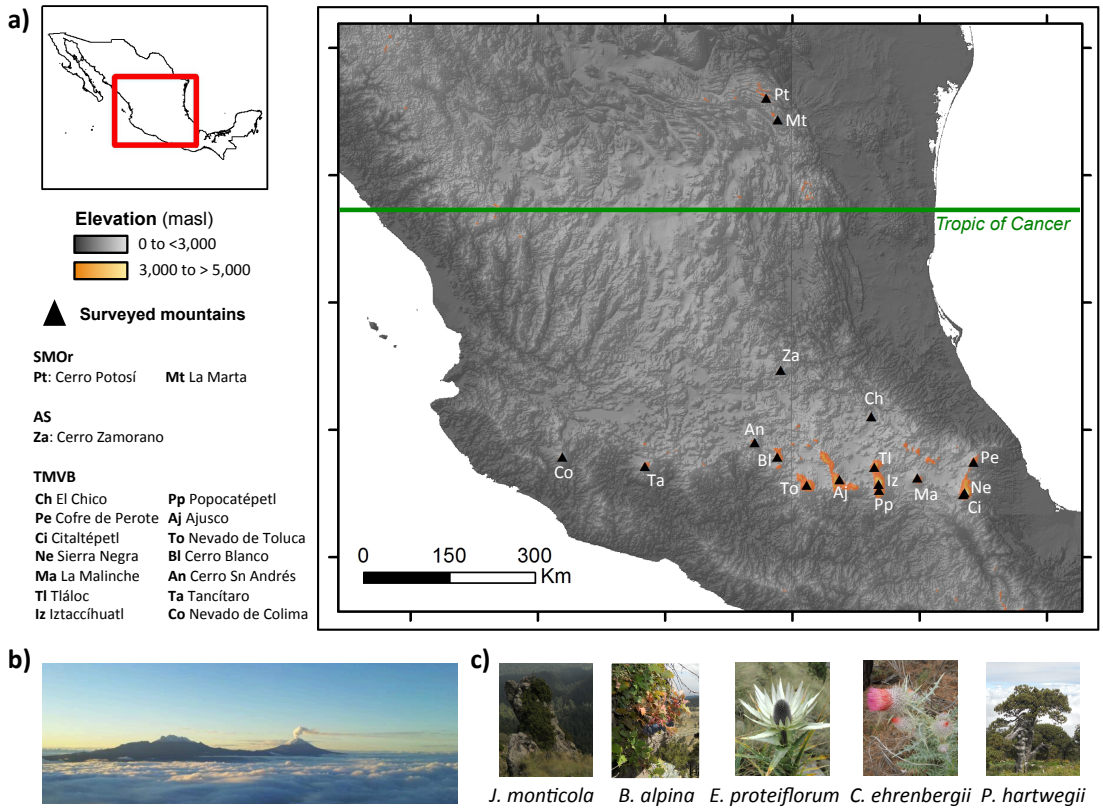


Figure 1.1. Study system. *a)* High elevation mountains of the Sierra Madre Oriental (SMOr) Altiplano Sur (AS) and Transmexican Volcanic Belt (TMVB) that were surveyed for *c)* five timberline–alpine plant species (*Juniperus monticola, Berberis alpina, Eryngium proteiflorum, Cirsium ehrenbergii* and *Pinus hartwegii,* left to right). Timberline is around 3,700-3,900 masl, just above clouds level of *b*. Photo credits: A. Mastretta-Yanes for species pictures on *c*, and Pezetaroi for Iztaccíhualt and Popocatéleptl volcanoes view of *b*.

The study of tropical mountain biodiversity, and in particular of the Mexican highlands, has historically mostly relied on species occurrence data for three main reasons. First, the need to include historical variables (as opposed to only macroecological features) for explaining the biodiversity patterns of tropical mountains has only recently been realised in full (Kessler & Kluge 2008; Fjeldså & Bowie 2008). Second, phylogeography is a relatively recent field (Avise 2000), with much focus until recently on European landscapes and the northern most latitudes of North America (Emerson & Hewitt 2005; Avise 2009). Within these regions, the Pleistocene climate fluctuations had dramatic effects on

species distributions, leading to relatively clear phylogeographic patterns that were possible to elucidate from animal mtDNA sequences (Hewitt 2000). However, by the time phylogeographic methods started to be applied within more southern latitudes (for instance for the Mexican region 85% of papers were published within the last decade, see Chapter 2) it was becoming clear that to distinguish pattern and infer process, and to obtain more accurate divergence times, a multilocus approach would be necessary (Zhang & Hewitt 2003; Brito & Edwards 2009; McCormack *et al.* 2011). This made phylogeographic studies focusing on plants particularly problematic because (1) cpDNA sequences have been found to be more informative for phylogenies than to examine infra-species level variation (Avise 2009); and (2) plant nuclear genes were poorly explored and relatively difficult to target for each particular species (Schaal *et al.* 1998). For instance, in a pilot study for the realization of this thesis, five plant species from high altitude Mexican mountains (*Juniperus monticola, Berberis alpina, Eryngium proteiflorum, Cirsium ehrenbergii* and *Pinus hartwegii,* Fig. 1.1c) were sampled and screened for variation at the most variable cpDNA regions among plants (Shaw *et al.* 2005, 2007). After sequencing >1,000 bp per species in 5 – 15 individuals per sampling locality of distantly located mountains, we found only 10 substitutions or indels in the most variable species down to two haplotypes or no variation at all (Fig. 1.2). Finally, even though there has been a sustained interest in the biogeography of the Mexican highlands since Humboldt & Bonpland's (1805) altitudinal regionalization (Espinosa *et al.* 2008), the natural history knowledge of the region is still incomplete. To move forward two important methodological challenges associated with the biological and physical complexity of the TMVB need to be addressed: (1) the TMVB occurs in the

transition zone between the Neartic and Neotropical biogeographic realms (Halffter 1987; Morrone & Márquez 2001) and; (2) it is a topographically heterogeneous landscape whose Quaternary volcanic origin has just recently been geographically mapped with detail (Gómez-Tuena *et al.* 2007; Ferrari *et al.* 2012), and for which paleoclimatic data was relatively scarce until recent years.

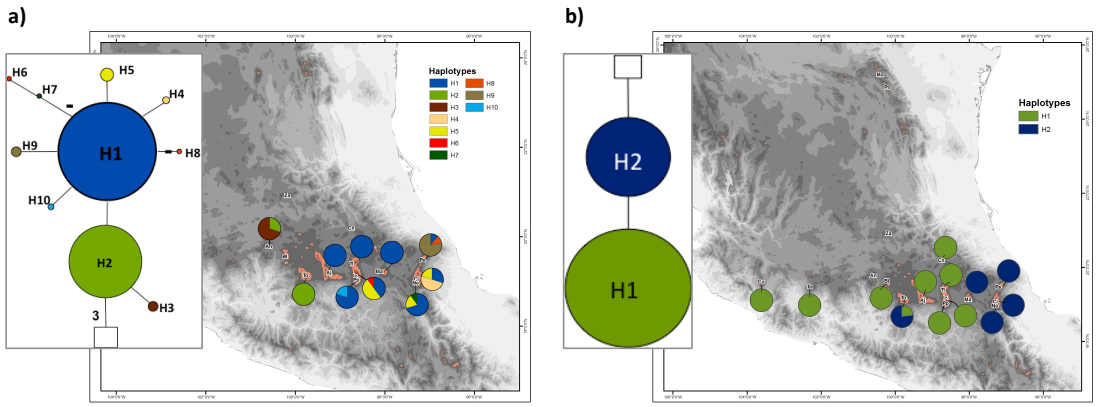


Figure 1.2. Haplotype network and population frequencies for *a) E. proteiflorum,* the species where more haplotypes were found, and *b) J. monticola* the species where only two haplotypes were found. The species where less variation was found was *B. alpina* (data not shown). The circle size in the network corresponds to the haplotype frequency. Outgroups are represented as squares. The colours on the pie charts represent the frequency of each haplotype in each sampling locality.

To address these methodological challenges, in Chapter 2 I gather and review the information on the biogeographic and physical history of the Mexican highlands, and then propose a set of phylogeographic hypotheses that can be tested with landscape explicit analyses. I then test some of these hypotheses in Chapter 5 using thousands of genomic loci from two species: *Berberis alpina* and *Juniperus monticola*. These species were chosen among five timberline-alpine grasslands species (Fig. 1.1c) for being diploid (*E. proteiflorum* and *C. ehrenbergii* are not), being restricted almost exclusively to the TMVB (*P. hartwegii* extends to Northern Mexico and Central America) and presenting two important

differences: (1) being insect pollinated (*B. alpina*) vs wind pollinated (*J. monticola*), and (2) having a more restricted distribution limited to few of the highest mountain peaks (*B. alpina*) vs occurring throughout the TMVB at the highest peaks, but also at slightly lower elevations (*J. monticola*). But before proceeding to landscape analyses, I first improve existing molecular and bioinformatics protocols for double digest restriction-site associated DNA sequencing (ddRAD; Peterson *et al.* 2012), the method used in this thesis to generate genomic data.

Double digest restriction-site associated DNA sequencing forms part of the family of genotyping-by-sequencing methods (reviewed by Davey et al. 2011; Poland and Rife 2012) that allow subsampling of a genome at putatively homologous locations across many individuals to identify and type single nucleotide polymorphisms (SNPs) in short DNA sequences, rapidly and at low cost, regardless of genome size and previous genomic knowledge. ddRAD is a modification of the RAD sequencing (RADseq) protocol (Miller *et al.* 2007; Baird *et al.* 2008). Briefly, it consists of digesting a genome with two restriction enzymes and using a parallel sequencing platform, such as Illumina, to sequence the fragments (Fig. 1.3a and Fig. 1.4). Sequencing reads are then processed by bioinformatic tools to either map them to a reference genome or to *de novo* assemble them into anonymous loci (Fig. 1.3b). The main difference of ddRAD over RAD-seq is that the former allows to subsample fewer loci of a genome, thus also allowing us to target species with large genomes (such as conifers, with genome sizes typically larger than 10 Gb), or increase the number of individuals to be sequenced in the same lane (Peterson *et al.* 2012), therefore becoming more useful for the objectives of this thesis.

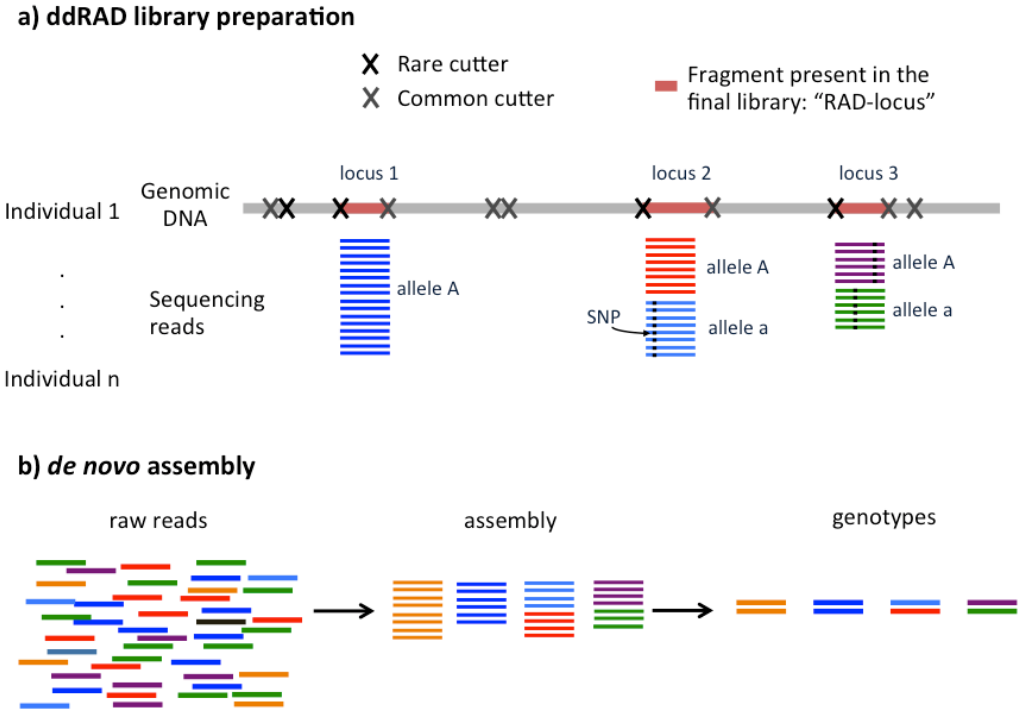


Figure 1.3. Generation of ddRAD data. (a) During library preparation genomic DNA is digested with two restriction enzymes (a rare and a common cutter) and processed to create sequencing competent fragments (details on Fig. 1.4). The RAD-loci present in the final library are the fragments kept after the size selection, and a RAD-locus is thus a short DNA sequence. Each locus can have one or more alleles, which differ from each other by a small number of SNPs (black squares). Sequencing produces a number of reads per allele, which is referred to as coverage. The same procedure is repeated with several individuals, where the same loci are expected to be recovered. (b) In the absence of a reference genome, loci are *de novo* assembled by matching together similar sequences and considering them either different loci, or alleles from the same locus. This is done based on a given number of mismatches (which are defined by researchers). Once loci and alleles are assembled, genotypes are scored for each individual of the dataset.

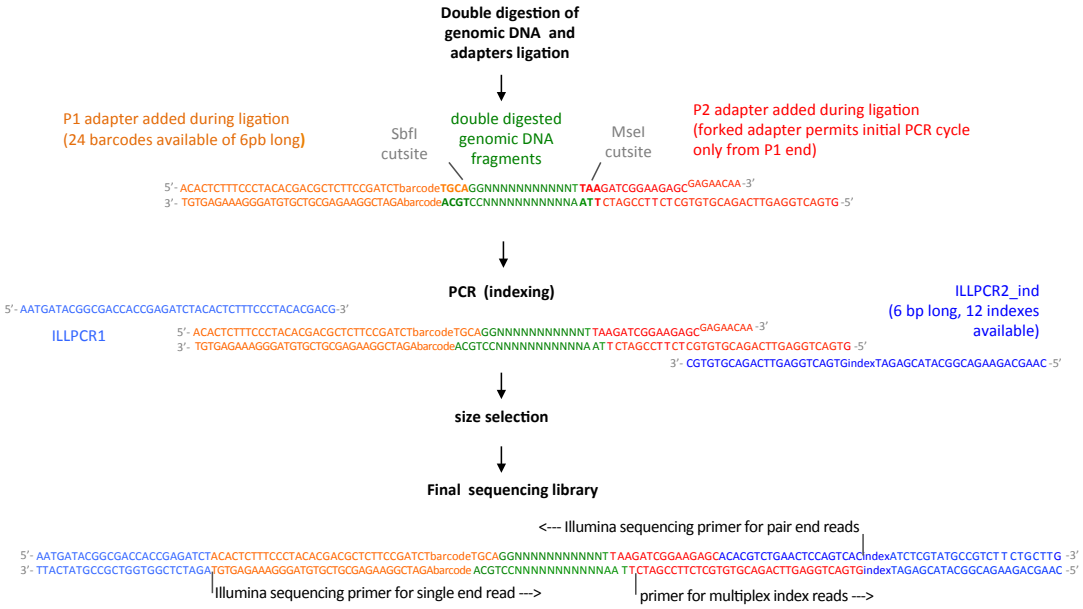


Figure 1.4. Preparation of sequencing competent fragments with a ddRAD double indexing protocol. Genomic DNA of each individual sample is digested with two restriction enzymes (SbfI and MseI in this example). The fragments of double digested genomic DNA (green) are then ligated to the adapters (red), which contain the reverse complement of each restriction enzyme cutsite. Adapters P1 include barcodes (a short unique sequence, 24 available in this example) that are sample specific (in combination with an index, see next). Individual samples with different barcodes can then be pooled together and amplified using Illumina PCR primers. A different index (short unique sequence added by primer ILLPCR2, 12 available in this example) can be used for each group of unique barcodes, thus allowing to "re-use" the barcodes and to include more individuals per sequencing lane. After the amplification, a desired fragment range is size selected. The library (if only one index was used) or libraries (if several) can be sequenced either in a single sequencing lane or pooling together libraries with different indexes. After sequencing, samples are demultiplexed first by unique indexes and then by barcodes, to end with a set of raw reads distinguishable per individual sample.

Although the validity of RADseq data has been demonstrated and it is increasingly starting to be used on molecular ecology studies of non-model species (Narum *et al.* 2013), it is still a novel approach with caveats and with taxon-specific issues that need attention. In this thesis I address three of these needs: (1) quantifying genotyping error rates, (2) providing an objective way of optimising *de novo* assembly, and (3) dealing with paralogous loci, which are

particularly common in plant genomes and complicate *de novo* assembly. In Chapter 3 I focus on error rates and *de novo* assembly by introducing a novel approach: using DNA replicates to both (i) quantify error rates at the locus, allele and SNP level, and (ii) optimise *de novo* assembly parameters by minimizing error and maximizing the retrieval of informative loci. In Chapter 4, I use population level data to identify paralogs, and then, rather than simply filtering them out as genotyping-by-sequencing studies have typically done previously, I use them to explore recent gene duplication as a source of population divergence.

In Chapter 5 I use the climatic data discussed in Chapter 2 along with models of climate conditions during the Last Glacial Maximum (~20 kyr ago) to demonstrate that the TMVB may have provided long-term environmentally stable conditions for timberline-alpine grasslands to occur throughout glacial/interglacial cycles. I then propose that genetic differentiation among populations and private genetic variation within populations can be explained as a function of historical environmental isolation. That is to say, that montane taxa from the TMVB are under a sky-island dynamic, such that they are forced to high-elevation refugia during the interglacial periods, where divergence would be promoted by restricted gene flow, and to lower elevations during glacial periods, where the probability of admixture between previously isolated populations would increase. Population differentiation would therefore not only depend on simple geographic distance, but also on the topography of lower elevations, such that some presently disjunct populations may have experienced higher genetic connectivity during periods of glacial maxima, while other would may have remained isolated and as genetically disconnected as they are during

glacial minima. To test this hypothesis, I perform landscape analyses by: (1) using the ddRAD data processed for quality and orthologous loci as in Chapters 3 and 4; and (2) explicitly quantifying spatial isolation under different scenarios of population connectivity based on topography and environmental conditions for glacial and interglacial stages.

1. 1. References

Avise JC (2000) *Phylogeography the history and formation of species*. Harvard University Press Cambridge, MA.

Avise JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography*, **36**, 3–15.

Avise JC, Arnold J, Ball RM *et al.* (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP Discovery and genetic mapping using sequenced rad markers. *PLoS ONE*, **3**, e3376.

Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009) Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785 –789.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Emerson BC, Hewitt GM (2005) Phylogeography. *Current Biology*, **15**, R367–R371.

Espinosa DS, Ocegueda S, Aguilar Zuñiga C, Flores-Villela Ó, Llorente-Bousquets J (2008) El conocimiento biogeográfico de las especies y su regionalización natural. In: *Capital natural de México* , pp. 33–65. CONABIO, México.

Ferrari L, Orozco-Esquivel T, Manea V, Manea M (2012) The dynamic history of the Trans-Mexican Volcanic Belt and the Mexico subduction zone. *Tectonophysics*, **522–523**, 122–149.

Fjeldså J, Bowie RCK (2008) New perspectives on the origin and diversification of Africa's forest avifauna. *African Journal of Ecology*, **46**, 235–247.

Fjeldså J, Bowie RCK, Rahbek C (2012) The role of mountain ranges in the diversification of birds. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 249–265.

Fjeldså J, Lambin E, Mertens B (1999) Correlation between endemism and local ecoclimatic stability documented by comparing Andean bird distributions and remotely sensed land surface data. *Ecography*, **22**, 63–78.

Gómez-Tuena A, Orozco-Esquivel MT, Ferrari L (2007) Igneous petrogenesis of the Trans-Mexican Volcanic Belt. *Geological Society of America Special Papers*, **422**, 129 –181.

Halffter G (1987) Biogeography of the montane entomofauna of Mexico and Central America. *Annual Review of Entomology*, **32**, 95–114.

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

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

Humboldt A von, Bonpland A (1805) *Essai sur la géographie des plantes : accompagné d'un tableau physique des régions équinoxiales, fondé sur des mesures exécutées, depuis le dixième degré de latitude boréale jusqu'au dixième degré de latitude australe, pendant les années 1799, 1800, 1801, 1802 et 1803 /*. Chez Levrault, Schoell et compagnie, libraires. Paris, France.

Kessler M, Kluge J (2008) Diversity and endemism in tropical montane forests-from patterns to processes. *Biodiversity and Ecology Series*, **2**, 35–50.

McCormack JE, Heled J, Delaney KS, Peterson AT, Knowles LL (2011) Calibrating divergence times on species trees versus gene trees: implications for speciation history of *Aphelocoma* jays. *Evolution*, **65**, 184–202.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240 –248.

Morrone JJ, Márquez J (2001) Halffter's Mexican Transition Zone, beetle generalized tracks, and geographical homology. *Journal of Biogeography*, **28**, 635–650.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

Papadopoulou A, Anastasiou I, Spagopoulou F *et al.* (2011) Testing the species–genetic diversity correlation in the Aegean archipelago: toward a haplotype-based macroecology? *The American Naturalist*, **178**, 241–255.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome Journal*, **5**, 92.

Sandel B, Arge L, Dalsgaard B *et al.* (2011) The influence of late Quaternary climate-change velocity on species endemism. *Science*, **334**, 660 –664.

Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA (1998) Phylogeographic studies in plants: problems and prospects. *Molecular Ecology*, **7**, 465–474.

Shaw J, Lickey EB, Beck JT *et al.* (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, **92**, 142–166.

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

Vellend M, Geber MA (2005) Connections between species diversity and genetic diversity. *Ecology Letters*, **8**, 767–781.

Zhang D, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563–584.

CHAPTER 2

Biodiversity in the Mexican highlands and the complex interaction of geology, geography and climate at a tropical latitude

A version of this chapter has been submitted to *Journal of Biogeography*

No amo mi patria.
Su fulgor abstracto
es inasible.
Pero (aunque suene mal)
daría la vida
por diez lugares suyos,
cierta gente,
puertos, bosques de pinos,
fortalezas,
una ciudad deshecha,
gris, monstruosa,
varias figuras de su historia,
montañas
-y tres o cuatro ríos.

- *Alta traición* by José Emilio Pacheco

Dormant volcano, ticatla 2013

2.1. Abstract

Aim To (i) synthesise the currently dispersed data on the physical and phylogeographic history of the Mexican highlands, and (ii) review approaches that can be used for explicit hypothesis testing regarding the complex interactions of topography, recent volcanism and climate fluctuations at a tropical latitude.

Location Mexico

Methods We perform a literature and data survey of the climatic, geological and phylogeographic history of the Mexican highlands. We then assess how the expected effects of topographic isolation, co-occurring climate fluctuations and volcanism can be tested against the distribution of genetic diversity of high altitude taxa.

Results The Mexican highlands present a complex biogeographic, climatic and geological history. Montane taxa have been exposed to a sky-islands dynamic through climate fluctuations, allowing for long-term *in situ* population persistence, while also promoting recent divergence and speciation events. Volcanic activity transformed part of the Mexican highlands during the Pleistocene, leading to co-occurring climate and topographical changes. The Mexican highlands provide the conditions to examine how low-latitude mountains can allow both the long-term persistence of biodiversity as well as allopatric and parapatric speciation driven by climatic and geological events.

Main conclusions Climate fluctuations and recent volcanism have driven the diversification and local persistence of Mexican highlands biodiversity. The climate-volcanism interaction is challenging to study, however this can be

overcome by coupling genomic data with landscape analyses that integrate the geological and climatic history of the region.

2.2. Introduction

High altitude biotas are attracting increasing attention from macroecologists and evolutionary biologists in attempts to understand the relative importance of history and ecology in shaping the distribution of biodiversity (Graham *et al.* 2014). Much of this interest has been focussed upon tropical mountains like the Mexican highlands. Of the 70 phylogeographic studies that deal with montane taxa within Mexico, 86% were published in the last decade, and than 50% in the last four years (Table 2.1). It is therefore timely to review this research in an attempt to obtain a synthetic understanding of the origin and maintenance of biodiversity in these highlands and identify any knowledge gaps.

Mexico is located in a transition zone between tropical and subtropical latitudes within North America. It is characterised by mostly warm climates and arid ecosystems; however, biomes of temperate and cold affinity also exist in highland areas which are colder and moister than the lower elevations (Fig. 2.1a: Rzedowski 1978; García 1998; Challenger & Soberón 2008). These highlands are covered with oak-conifer forests that transition to alpine grasslands and represent a hot spot of temperate biodiversity (Mittermeier 2004). For example, Mexican highland forests contain approximately 50% of all *Pinus* species, 30% of all *Quercus* and 25% of all *Juniperus* species (Farjon & Styles 1997; Valencia 2004; Adams 2008), and in terms of endemic richness, 59 endemic plant species

have been recorded within a subset of mountain tops totalling no more than 6 km^2 in area (McDonald 1993).

The Mexican highlands have proven useful for understanding the diversification of North American temperate biodiversity within the framework of the Pleistocene climate fluctuations (e.g. Gugger *et al.* 2011; Wood *et al.* 2011; Aguirre-Planter *et al.* 2012), and are highlighting the role of low-latitude mountains as areas of long-term population persistence (e.g. Bryson & Riddle 2011; Moreno-Letelier *et al.* 2014). Although it has not been the specific focus or research efforts to date, the recent volcanic origin of some of the Mexican highlands (Ferrari *et al.* 2012) may facilitate the study of inland volcanism and its biological consequences. As an example, the area provides a system to compare the importance of either colonisation and evolution *in situ* within newly formed sky-islands. However, as we discuss below, studies of biodiversity and diversification in the area will have to consider carefully the complex biological and physical history of the Mexican Highlands.

Most of the Mexican highlands occur in the Mexican Transition Zone between the Neartic and Neotropical biogeographic realms (Halffter 1987; Morrone & Márquez 2001). They represent a heterogeneous topography of different geologic ages (Ferrusquía-Villafranca 1993) where species distributions would have been subject to altitudinal shifts during the Pleistocene climate fluctuations (Toledo 1982; Metcalfe *et al.* 2000). Thus, what makes the region interesting also makes it challenging to study: in some geographic areas volcanic activity modified the landscape during the Pliocene-Pleistocene (Ferrari *et al.* 2012) and climate variation was also important during the Neogene (Graham 1999; Salzmann *et al.* 2011). Therefore, to formulate and test specific

phylogeographic hypotheses it is necessary to integrate both climate and geological data with geographically explicit frameworks.

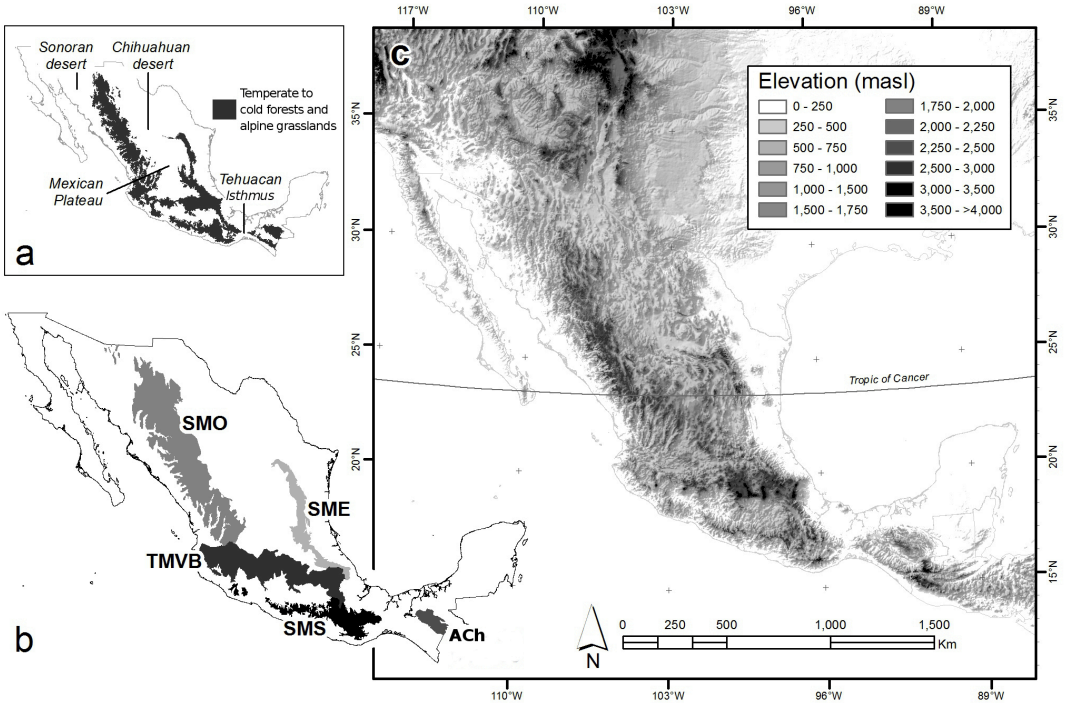


Figure 2.1. Distribution and altitude of Mexican highlands. a) distribution of temperate -cold ecosystems (oak-conifer forests to alpine grasslands) and main geographic barriers. b) Mexican mountain ranges: Sierra Madre Occidental (SMO), Sierra Madre Oriental (SME), Transmexican Volcanic Belt (TMVB), Sierra Madre del Sur (SMS) and Altos de Chiapas and Guatemala (ACh). c) Altitudinal range in meters above sea level (masl) for the Mexican region.

Spatial analyses including climate and geological data seldom feature in phylogeographic studies of the Mexican highlands for two main reasons. First, until recently phylogeographic methods have lacked analytical techniques for such integrative analyses. Second, information on the geological and climatic history of the region was little or did not include spatial data. Advances in phylogeographic methods (Richards *et al.* 2007; Knowles & Carstens 2007; Chan *et al.* 2011) now allow the integration of geographic information systems (GIS) and species distribution modelling (SDM) to test spatially explicit hypotheses

(e.g. Hugall et al., 2002; Carnaval et al., 2009; Knowles and Alvarado-Serrano, 2010). Also, the study of the geological history of the Mexican region is now yielding spatial data that can be used to test landscape explicit scenarios (e.g. Ferrari *et al.* 2005, 2012).

Here we review the climate, geological and phylogeographic history of the Mexican highlands. We then discuss the expected effect of topographic isolation and co-occurring climate fluctuations and orographic processes on the distribution of genetic diversity of high altitude taxa. Finally, we suggest how climate and geological data may be used to test geographically explicit hypotheses. We focus these models on the timberline and alpine grasslands of the Transmexican Volcanic Belt (TMVB, Fig. 2.1b) because they are distributed at the same tropical latitude (~19-20 ºN) on highly isolated volcanic peaks with a Neogene to present volcanic origin (Ferrari *et al.* 2012), thus becoming a unique but challenging system to test the role of climate fluctuations and volcanism on shaping the distribution and diversification of montane biodiversity.

2.3. Geographic setting

The Mexican highlands extend from the south of the Rocky Mountains in the United States down to the northern limits of the Central America mountain systems. Within Mexico, the montane areas can be divided in the Sierra Madre Occidental (SMO), Sierra Madre Oriental (SME), Trans-Mexican Volcanic Belt (TMVB), Sierra Madre del Sur (SMS) and Altos de Chiapas and Guatemala (ACh) (Ferrusquía-Villafranca, 1990; Fig. 2.1b). Together, these mountain ranges present altitudes from a minimum of 1,800 meters above sea level (masl) up to

more than 5,000 masl (Fig. 2.1c). The SMO and SME are North-South mountain ranges in the West and East of Mexico respectively, and are separated by the Chihuahuan Desert. The TMVB is constituted by hundreds of volcanic structures that extend from the Mexican Pacific coast to the Gulf of Mexico at a latitude of ~19-20º N. The SMS and the ACh are in South Mexico and are divided by the Tehuantepec Isthmus (TI, Fig. 2.1a). The SMO and SME extend above the Tropic of Cancer in their northern parts, while the remainder of the Mexican montane ranges occur south of it. The SMO and SMS are topographically complex yet represent continuous high elevation massifs, whereas the TMVB, the SME and ACh are characterized by isolated peaks surrounded by much lower elevations (Fig. 2.1c).

The vegetation types that characterize the highlands are oak, conifer and cloud forests, as well as subalpine and alpine grasslands, distributed in an altitudinal gradient from ~2,000 masl to >4,000 masl (Rzedowski 1978; Calderón de Rzedowski & Rzedowski 2005; Challenger & Soberón 2008; Socorro *et al.* 2012). The lowlands that separate the higher areas have warmer climates with deserts, dry rainforests and tropical rainforests (Challenger & Soberón 2008).

2.4. Physical history

2.4.1.Climate history

After the Mid-Miocene Climatic Optimum (ca. 15 Myr), the global climate began a cooling trend that was followed by the establishment of a major ice-sheet in Antarctica (~10 Ma) and later the onset of the Northern Hemisphere Glaciation

(~3.2 Ma) (Zachos *et al.* 2001). Despite this cooling trend, the global temperature was still 2-3ºC higher than today by the end of the Pliocene (3.6-2.5 Ma) (Salzmann *et al.* 2011). There is little specific information on Neogene climate for the Mexican highlands, but it is possible that ecosystems existed with species compositions that lack a modern analogue (Salzmann *et al.* 2011). Pliocene fossils of temperate plants genera now characteristic of the Mexican highlands have been found outside the taxa's current distributional range (Fig. 2.2; Graham, 1999), so it has been suggested that during the Pliocene: a) arid scrublands extended in most of North Mexico; b) warm-temperate evergreen conifer forests existed in the northernmost mountains of the SME, and; c) a warm-temperate mixed forest with broadleaved trees and conifers covered highland regions of South Mexico and Central America (Graham 1999; Salzmann *et al.* 2011).

The Pliocene was followed by the high magnitude glacial-interglacial climate oscillations of the Pleistocene, between 2.58 million years ago (Ma) and 11.7 thousand years ago (kya ago) (Cohen & Gibbard 2011). During the glacial periods of this epoch polar ice sheets advanced southwards across North America, but they did not penetrate into Mexico (Porter 2000; Lachniet & Vazquez-Selem 2005). Climate fluctuations in Mexico were therefore less dramatic than in higher latitudes and no ice-sheets covered large extensions of land. However, temperatures still decreased considerably, and precipitation patterns and seasonality changed. Mean temperatures were, for example, around 6ºC lower than today in some parts of the SMO during the Last Glacial Maximum (LGM, ~20kya) (Metcalfe 2006). Similarly, ice caps formed on some mountains in the TMVB that are currently unglaciated, thereby lowering the vegetation line

around 1,000 m (Metcalfe 2006; Vázquez-Selem & Heine 2011) (Table 2.2). The local effect of the glacial periods varied over time. For example, the mean altitude of the glacier terminus on Iztaccihuatl volcano (TMVB), today above 4,700 masl, was 3,390±160 during the LGM and down to 3,000 masl during a previous glacial period 200-175 kya (Vázquez-Selem and Heine, 2011). The geographic arrangement of mountains also resulted in regional variance for glacial effects because rain and humidity conditions varied across Mexico due to latitude, distance to the oceans and the level of topographic isolation (Bradbury 1997; Metcalfe *et al.* 2000).

As a consequence of the climate fluctuations of the Pleistocene, the distributions of temperate to cold-affinity taxa underwent altitudinal changes. For example, fossil records from the LGM to the Holocene show that temperate-cold affinity taxa extended to lower altitudes during the LGM and contracted to higher elevations as conditions started to become warmer (Van Devender 1990a; b; Lozano-García & Ortega-Guerrero 1994; McAuliffe & Van Devender 1998; Lozano-García *et al.* 2005; Ortega-Rosas *et al.* 2008). Fossil records and glacial deposits of other glacial stages previous to the LGM suggest that similar conditions to those in the LGM would have characterized previous glacial periods (Caballero & Guerrero 1998; Lozano-García *et al.* 2002; Ortega *et al.* 2002; Vázquez-Selem & Heine 2011). Changes in precipitation also occurred due to the Pleistocene climate fluctuations and were of particularly importance to some taxa, for instance cloud forests species (Ramírez-Barahona & Eguiarte 2013). Fossil records and geologic evidence of the glacial/interglacial fluctuations for the Mexican highlands are available in Table 2.2.

In addition to point fossil records and geologic evidence, changes in environmental conditions through time can also be examined with present climate data and simulated conditions of the past (Hijmans *et al.* 2005; Braconnot *et al.* 2007). Models of temperature and precipitation show that for the Mexican Highlands, cold and humid conditions were geographically more extensive during the LGM than today, where they are restricted to the highest mountains of the TMVB, SME and SMO (Fig. 2.3).

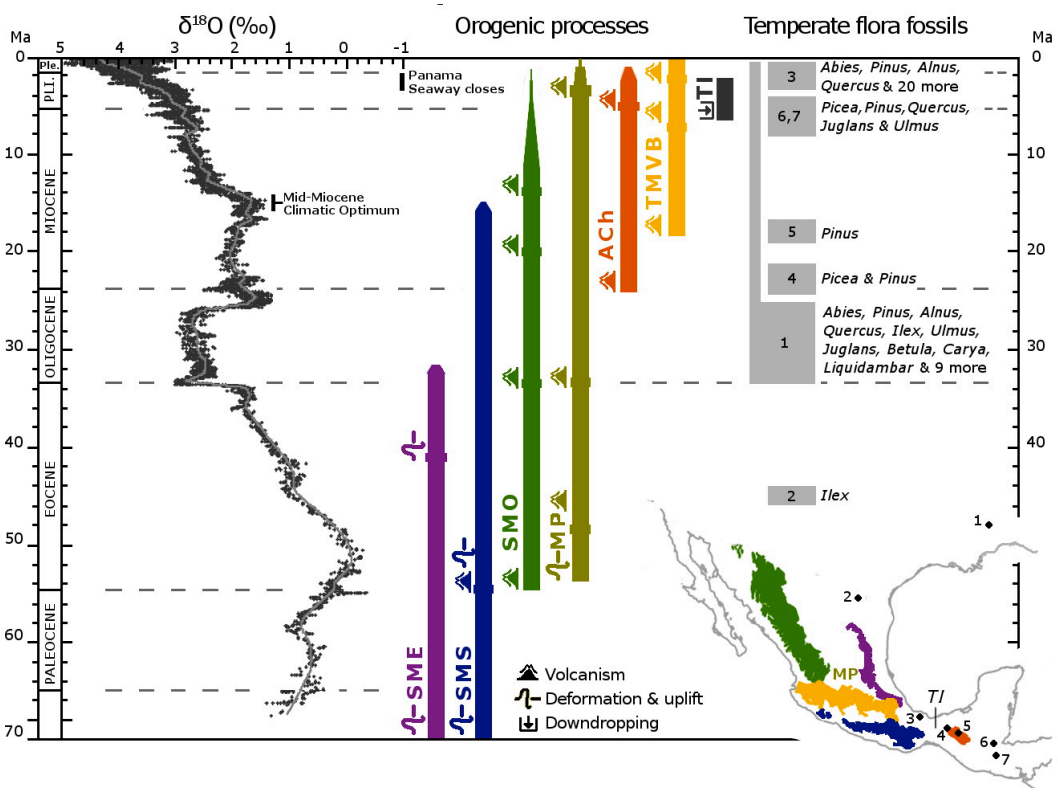


Figure 2.2. Global deep-sea oxygen isotope records ($\delta^{18}O$, higher levels mean lower temperature), major geologic events related to the Mexican highlands and their geographic location, as well as the oldest fossil records for temperate Neartic taxa for Cretaceous to present. Oxygen record modified from Zachos et al. (2001), fossil data from Graham *et al*. (1999) and geologic data from Barrier et al. (1998), Eguiluz de Antuñano, et al. (2000), Manea and Manea (2006), Ferrari et al. (2007, 2012), Morán-Zenteno et al. (2007) and Nieto-Samaniego et al. (2007). Orogenic processes are shown for Sierra Madre Oriental (SME), Sierra Madre del Sur (SMS), Sierra Madre Occidental (SMO), Mexican Central Plateau (CP), Altos the Chiapas (ACh), Transmexican Volcanic Belt (TMVB) and Tehuacan Isthmus (TI).

2.4.2. Geological history

The Mexican highlands are a complex mixture of distinct geological provinces with different ages and origins (Fig. 2.2). In the North, the SME is related to the Laramide orogeny (ca. 70 to 40 Ma) that is also implicated in the formation of the Rocky Mountains, and the SMO as a high plateau is related to the intense explosive volcanism that took place in the Oligocene and Early Miocene (Ferrari et al., 2007). In Central Mexico Late Miocene to recent volcanism produced the TMVB (Ferrari *et al.* 2012). In South Mexico, the formation of the SMS and Ach is associated with the interaction of several tectonic plates which caused the uplift of Central America and the closure of the Panama Isthmus (Ferrusquía-Villafranca 1993; Nieto-Samaniego *et al.* 2007). As a consequence the Mexican highlands have been under continuous geologic change from the Paleocene to the present (West 1964).

The SME is the oldest of the mountain ranges as its formation ceased in the Oligocene (Eguiluz de Antuñano, *et al.* 2000). The SMO and SMS are also relatively old, with the major part of their orogeny occurring during the Oligocene, although they were still active during the early Miocene and some parts during the Pleistocene (Ferrari & Luna-González in press; Ferrari *et al.* 2007; Morán-Zenteno *et al.* 2007). Tectonic activity and volcanism from ACh formed during the Late Miocene and Pliocene (Manea & Manea, 2006; Mora *et al.* 2007; Witt *et al.* 2012). The region between ACh and SMS then suffered a partial down-dropping during the latest Miocene to early Pliocene, leading to the destruction of what is thought to have been a highland corridor spanning what is now the Isthmus of Tehuantepec (Barrier *et al.* 1998). The most recent geologic changes in Mexico occurred in the TMVB, generating thousands of volcanic

structures in Central Mexico from the Miocene to the present, with the largest volcanoes (>3,500 masl) forming during the Pleistocene (Ferrari *et al.* 2012) (Fig. 2.4d).

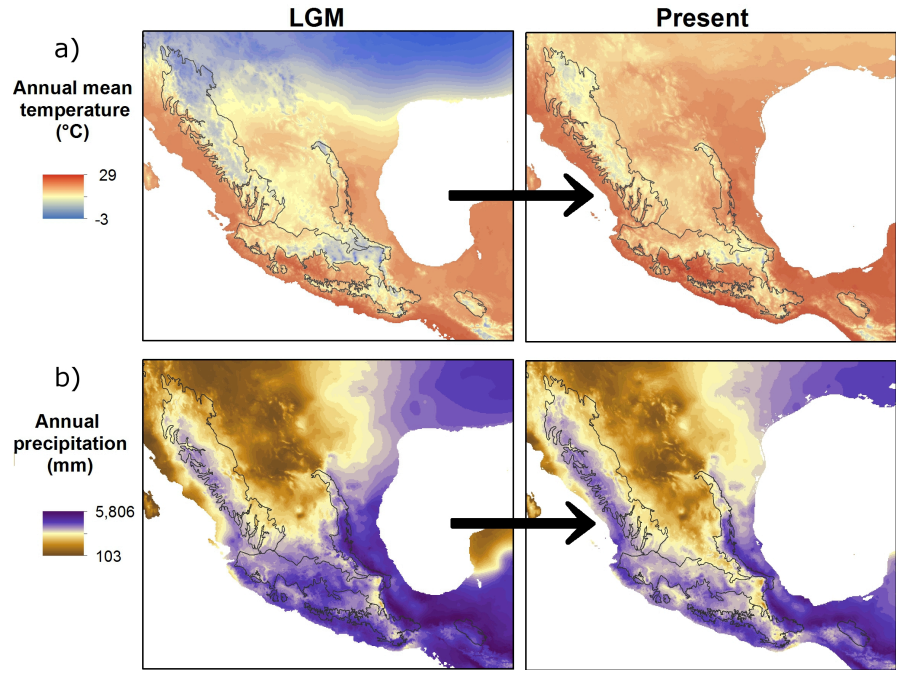


Figure 2.3. Annual mean temperature and precipitation for the Mexican highlands for the Last Glacial Maximum (LGM, ~20 kyr ago) and the Present. Data from Hijmans et al. (2005) and Braconnot et al. (2007).

Although the TMVB is very complex and our understanding of its geologic history is still incomplete, it is the most geologically studied area of Mexico. There are comprehensive summaries of the origin and ages of many volcanoes and regions (see Gómez-Tuena et al., 2007, Ferrari et al., 2012 and supplementary materials therein). Briefly, the geological evolution of the TMVB has been divided into four episodes: (1) early to mid Miocene; (2) late Miocene; (3) latest Miocene - early Pliocene, and; (4) late Pliocene and Pleistocene (Fig. 2.4, Ferrari *et al.* 2012). The final episode of the TMVB formation was characterized by the construction of large (<3,500 masl) stratovolcanoes during the last 1.5 Myr, some of which are still active (Gómez-Tuena *et al.* 2007; Ferrari

et al. 2012). Therefore the topography of the TMVB has changed considerably over the last 3 Myr, coincident with the dramatic climate fluctuations of the Pleistocene.

2.5. Phylogeographic consequences of climate fluctuations and geological changes

2.5.1. Origin and diversification of lineages

The Mexican highlands are inhabited by temperate-cold tolerant taxa that include species with different biogeographic origins: Neartic, Neotropical and Paleoamerican (Halffter 1987; Marshall & Liebherr 2000; Halffter *et al.* 2008; Morrone 2010). The Neartic taxa are those that have northern relatives within the Rocky Mountains and areas across the United States and Canada (Morrone 2010). The Neotropical taxa (corresponding to the Panamanian in Holt *et al.* (2013)'s update to biogeographic regions) are those that are related to species from Central and South America. Among the neotropical taxa, the subset that inhabits temperate to cold and humid ecosystems are the Montane Mesoamerican taxa, whose major centre of diversity is the montane habitats of Central America (Morrone 2010). The Palaeoamerican taxa are those whose closest relatives are temperate or tropical taxa from the Old World and whose presence in Mexico is suggested to be very old (Halffter 1987; Morrone 2010).

The distinction of Neartic and Neotropical biogeographic histories within Mexican highland taxa has two immediate consequences for phylogeography. First, it provides a sense of direction that could be used as a null hypothesis when testing range expansion and colonization routes. For Neartic species,

ancestral genetic diversity is expected in the northern part of their distribution, with a west-to-east colonisation of the TMVB for species (or sister species) distributed in the SMO, and east-to-west for species related to taxa from the SME (e.g. conifers and rodents; Rodríguez-Banderas *et al.* 2009; Aguirre-Planter *et al.* 2012; Mathis *et al.* 2014). For Neotropical taxa, ancestral variation would be expected in the Southern region of a given species range, as has been observed in a cloud-forest shrub (Ornelas & González 2014). The second consequence is that the division of Neartic and Neotropical biogeographic history provides a time frame and a set of diversification hypotheses that can be explored with molecular markers. Neotropical taxa are thought to be composed of groups that arrived from the south only after the closure of the Panama land bridge between 3.5 to 2.5 Ma (Fig. 2.2, Graham 1992; Coates *et al.* 1992; Webb 2006, but see Montes *et al.* 2012 for evidence of the bridge existing since late Eocene to late Miocene) and by groups that diversified in Central America during the Oligocene (33-23 Ma) (Graham 1992; Wendt 1993; Morrone 2006, 2010). It has been suggested that the Neartic species inhabiting Mexico could be the product of southwards migrations that occurred as a consequence of the cooling trend after the Mid-Miocene Climatic Optimum (15-10 Ma) (Fig. 2.2, Graham 1999) followed by diversification during the Pliocene-Pleistocene (Marshall & Liebherr 2000; Morrone 2010). Pleistocene divergence times (Table 2.1) have been found in some Mexican highlands taxa of Neartic origin, such as mice (Edwards & Bradley 2002) and snakes (Bryson *et al.* 2011b; c; Wood *et al.* 2011). However, in a number of other neartic reptiles (Devitt 2006; Bryson & Riddle 2011; Bryson *et al.* 2012a; b), birds (McCormack *et al.* 2008b, 2011) and conifers (Willyard *et al.* 2007; Aguirre-Planter *et al.* 2012; Moreno-Letelier *et al.* 2014) divergence times

are estimated to fall within the Miocene (Table 2.1, see more examples reviewed in Bryson *et al.* 2012b). This is in conflict with a strict Pliocene-Pleistocene diversification hypothesis (Marshall & Liebherr 2000; Morrone 2010) and suggests that Mexican highland taxa of Neartic origin are the product of multiple arrival events (Halffter 1987; Graham 1999; Morrone & Márquez 2001).

A more detailed understanding of the drivers of diversification within the Mexican highlands is the next step. Excitingly, given the new analytical tools and the spatial data that has recently accumulated, it is now feasible to consider how the joint, and possibly synergetic, effect of climate fluctuations and recent volcanism may have resulted in low-latitude mountains becoming 'cradles of biodiversity' (Fjeldså *et al.* 2012).

2.5.2. Pleistocene climate fluctuations

The fact that the Mexican highlands are on the limits of the tropics has had two important consequences for species distributions through the Pleistocene climate fluctuations. First, some species that today inhabit the United States had southern glacial refugia in the Northern regions of the SMO and the SME within Mexico (e.g. Masta 2000; Gugger *et al.* 2011 Table 2.1). Second, the Mexican highlands are among the areas of low-climate change velocity, meaning they are areas where biodiversity can survive relatively *in situ* through global climate fluctuations by undertaking altitudinal shifts instead of long latitudinal movements (Sandel *et al.* 2011). From this we expect that species from the Mexican highlands were led to high-elevation refugia during the interglacial periods, where divergence could be promoted by restricted gene flow. Similarly, it is expected that genetic admixture could be promoted at lower elevations

during the glacial periods. This can be considered a sky-island dynamic similar to other montane regions of the word (e.g. Knowles 2000). The translation of a sky-island dynamic into population differentiation and speciation will be a function of several factors: the age of a taxon, its environmental preferences, its dispersal ability and the particular area it inhabits.

Populations of several species seem to have been exposed to the sky-islands dynamic through millions of years (e.g. Aguirre-Planter *et al.* 2012; Bryson *et al.* 2012), which has been suggested to be the driver of some recent speciation events (e.g. Bryson *et al.* 2011). Regarding the spatial scale of the sky-island dynamic, it can occur both within and between mountain ranges (Table 2.1). For example, populations of some species may have survived relatively *in situ* through several glacial/interglacial fluctuations by undertaking down/upslope movements in "archipelagos" within the SMO (e.g. Wood *et al.* 2011). Alternatively some populations may have extended their distributions from a mountain range to lower elevations, sometimes colonising a different mountain range (e.g. from the SMO and TMVB to the SME, Moreno-Letelier & Piñero 2009; Gutiérrez-Rodríguez *et al.* 2011). In such cases, connectivity may have been restricted by wide geographic barriers, like the Chihuahuan Desert of the Tehuacan Isthmus. Whether species remained in a single mountain range or extended across the Mexican highlands would have been determined by their dispersal ability but also their environmental preferences. Species with a more temperate than cold temperature affinity tend to be distributed at lower elevations, so connectivity during glacial periods and gene flow among different mountain ranges is more likely (e.g. Anducho-Reyes *et al.* 2008; Cavender‐Bares *et al.* 2011; Moreno-Letelier *et al.* 2013). In contrast, species with colder

affinities tend to be more restricted to mountain tops and present higher population differentiation (e.g. Aguirre-Planter *et al.* 2000; Gugger *et al.* 2011). Similarly, species with a high hydrological vulnerability (like cloud forest taxa) are considerably influenced by the distribution of precipitation through climate fluctuations (Ramírez-Barahona & Eguiarte 2013; Ornelas & González 2014).

The effect of the Pleistocene fluctuations on species distributions varied across Mexico and within each mountain range. Latitudinal migrations may have been more pronounced in Northern areas, where in addition to altitudinal fluctuations, species could be expected to undertake north/southwards movements, as it has been found for Northern populations of species inhabiting the SMO (e.g. Wood *et al.* 2011 se also Mastretta-Yanes in press for a review). In contrast, both the TMVB and the SMS are formed by relatively isolated mountains at approximately the same latitude. For these mountain ranges, changes in species distributions and population differentiation are expected to be driven mainly by altitude, topographic isolation and distance to the oceans (because of its influence on precipitation). These variables have been discussed by several phylogeographic studies of taxa from the TMVB (e.g. Bryson & Riddle 2011; Bryson *et al.* 2011b, 2012b; Cavender‐Bares *et al.* 2011; Salas-Lizana *et al.* 2011, Table 2.1), but explicit landscape analyses have only recently been applied to evaluate the genetic data against competing glacial/interglacial scenarios (Bryson *et al.* 2011b; Parra-Olea *et al.* 2012; Ornelas & González 2014).

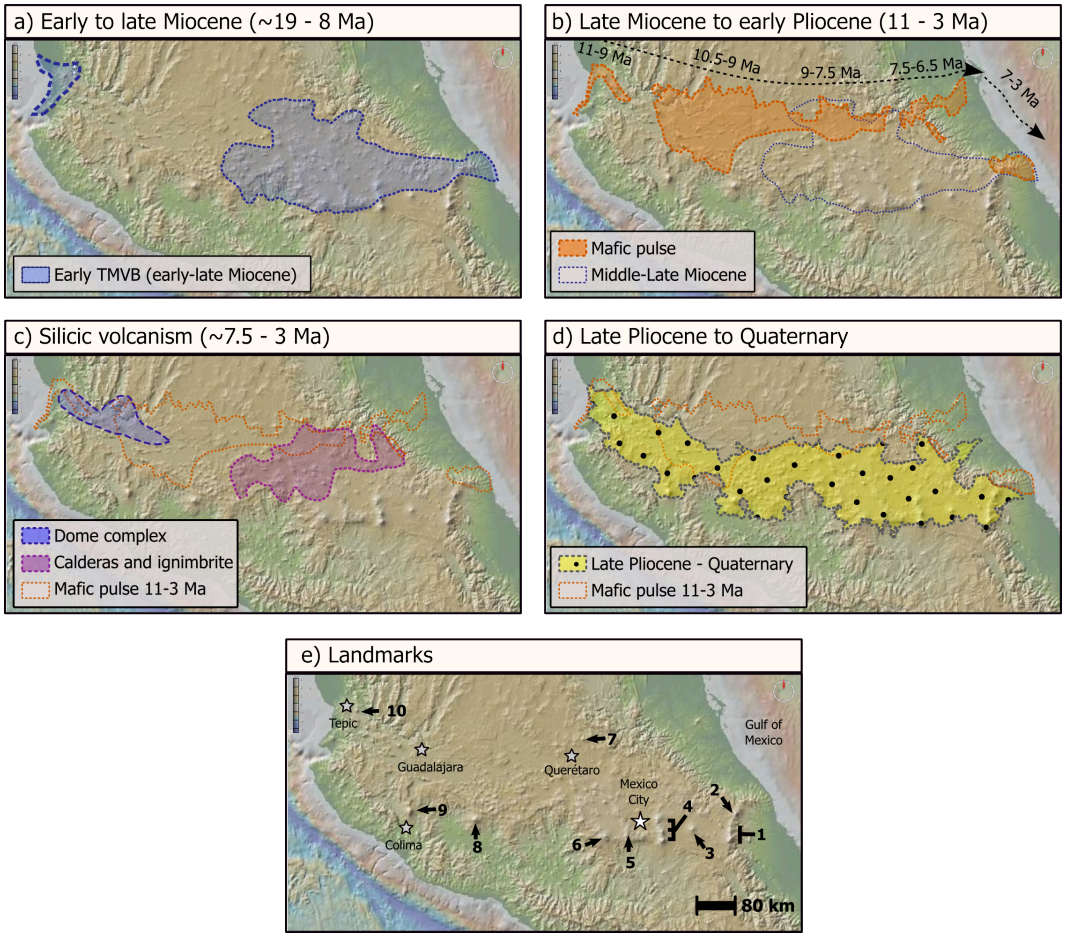


Figure 2.4. Formation episodes (a-d) of the TMVB. a) Start of the formation of a volcanic arc in the early to mid Miocene that lasted until the late Miocene. This included the formation of low altitude volcanoes (<1,000 m above plateau level) in the Central-Eastern region (Estado de Mexico and Southern part of Hidalgo states). b) Northern pulse migrating eastwards from late Miocene to early Pliocene, during which plateaus covering large regions were created and the elevation range also could have increased gradually (hundreds of meters over 1-2 Myr) due to isostatic movements (Ferrari 2004). c) From the latest Miocene to the early Pliocene calderas and ignimbrites (explosive eruptions) covered large areas of Central-Eastern region, and dome complexes (up to 500-600 m high above ground level) formed in the Western region. d) Development of a volcanic arc from the Pliocene to Pleistocene, with large stratovolcanoes (>2,000 m above plateau level, leading to >3,500 masl) forming during the last 1.5 Myr. e) Some cities (stars) and stratovolcanoes (numbers) of the TMVB. 1) Citlaltépetl (Pico de Orizaba) and Sierra Negra, 2) Cofre de Perote, 4) Tláloc, Iztaccíhuatl and Popocatépetl, 5) Sierra de las Cruces, 6) Nevado de Toluca, 7) Cerro Zamorano, 8) Tancítaro, 9) Nevado de Colima and 10) Sangangüey. Some volcanoes formed during the Miocene (e.g. Cerro Zamorano, number 7, ~11 Myr ago) while most emerged at different points during the Pleistocene (rest of the numbers) or even more recently (e.g. Paricutín, near number 8, which erupted in 1943). Modified from Ferrari et al. (2012).

2.5.3. Geological events and past topographic configurations

The geological activity of the Mexican highlands may have promoted diversification by two mechanisms: (1) generating new geographic barriers that could promote allopatric speciation; or (2) as a source of new mountains for colonization and subsequent divergence (Halffter 1987). The orogenic processes that formed most Mexican highlands started or finished by the Miocene (Fig. 2.2), so divergence times that fall within the Pleistocene are inferred as being a consequence of climate fluctuations of that epoch, whereas older dates tend to be attributed to geologic activity (Table 2.1). Such an approach has been informative when examining taxa distributed among several mountain ranges. For instance, the geographic distribution of major lineages among closely related species from the SMO, SME, TMVB and SMS tends to match each mountain range and is temporally congruent with major geological events (e.g. Bryson & Riddle 2011; McCormack *et al.* 2011; Aguirre-Planter *et al.* 2012; Bryson *et al.* 2012). Similarly, east-west population divergence across the Tehuacan Isthmus has been found to be congruent with the age of its formation, confirming its emergence as a geographic barrier for species that were previously distributed continuously from the ACh to the SMS (e.g. McCormack *et al.* 2011; Rodríguez-Gómez *et al.* 2013; Ornelas & González 2014).

However, when it comes to population or species differentiation within the TMVB, there is a caveat for interpreting divergence times: although the TMVB started to form in the Miocene, many of the volcanoes of this area are less than 1.5 Myr young (Ferrari *et al.* 2012), so genetic patterns that temporally fell in the Pleistocene could be related not only to climate fluctuations, but also to volcanism. In other words, the sky-islands dynamic likely occurred with a

particular landscape configuration during the last 1 Myr of the Pleistocene, whilst genetic signatures of older landscape configurations (Miocene-Pliocene; Fig. 2.4) may still be detectable. This is an important analytical challenge: although estimates of the timing of divergence are central for testing the underlying causes of diversification (McCormack *et al.* 2011), in this case they alone are not informative enough to distinguish between the relative roles of orogeny and climate. Below, we propose what would be the expected effect of the climate-volcanism interaction and suggest which data and analyses are needed for examining it.

2.6. Landscape hypotheses for the Transmexican Volcanic Belt

2.6.1. Scenarios and expected effects of the climate-volcanism interaction

Within the TMVB both climatic and topographic changes overlapped during the last few million years. The expected effect of the Pleistocene climate fluctuations is a sky-islands dynamic that would promote differentiation during the interglacial period and admixture during the glacials. The expected effect of the volcanism is promoting species divergence either in allopatry (among highlands created in different volcanic events) or parapatry (by colonizing newly emerged high habitats from lower altitudes). These two phenomena cannot be easily disentangled, however it is possible to construct general hypotheses of how their interaction may have affected biodiversity.

Under a sky-islands dynamic caused by the Pleistocene climate fluctuations, the geographic distance separating the mountain tops of the TMVB is not the only variable determining isolation. Depending on the altitude of the

lowlands, continuous suitable habitat between two or more mountains may have existed during glacial periods, leading to differences between the geographic and the effective distance separating two mountains (Fig. 2.5a). If a sky-islands dynamic occurred within a fixed topography (Fig. 2.5b), a stepping stone colonisation of new mountains would lead to a gradient of decreasing genetic diversity from the source population, and the level of differentiation of each mountain would be expected to conform to a model of isolation by distance from the source. This can be used as null hypothesis against a more complex scenario where recent volcanism modifies the landscape. For instance, a newly emerged sky-island could be colonised from multiple sources, leading to admixture (Fig. 2.5c). Under such a model isolation by distance from the source is not expected, but the age of the stratovolcanoes would correlate positively with the relative contributions of ancestral and derived genetic variation within the gene pool (Fig. 2.5c).

Scenarios from Fig. 2.5 are of relevance to cold-adapted taxa that arrived to the TMVB during the Pliocene-Pleistocene, when the early stages of the TMVB formation had already finished and most of the high stratovolcanoes were forming. However, several species inhabiting the highest mountain peaks of the TMVB are closely related to species that inhabit nearby lowlands (but that are still high, relative to sea level), and that presumably existed in the TMVB during the early stages of its formation (Halffter 1987; Graham 1999; Morrone & Márquez 2001). If such species do have a longer history within the TMVB, they may have diverged in different highlands that gradually became less isolated from one another due to continuous geological activity (Fig. 2.6a). This should be particularly true for the Eastern part of the TMVB, where some stratovolcanoes

emerged during the first stage of the TMVB formation (Fig. 2.4a). There are two plausible outcomes of this process (Fig. 2.6b): (1) the evolution of reproductive isolation mechanisms that prevent genetic admixture, or (2) the formation of zones of genetic admixture. The subsequent emergence of large stratovolcanoes in the Pleistocene would have generated new habitat, providing a geographic template where both allopatric and parapatric speciation could occur (Fig. 2.6c). The glacial periods would be expected to result in different scenarios of range expansion and secondary contact as a function of species specific traits (e.g. niche, dispersal) and the effective distance among mountains (Fig. 2.6d). However, genetic differentiation caused by pre-Pleistocene landscape configurations could still remain detectable in species genomes. In particular for the TMVB, the first two episodes (Fig. 2.4a-b) could have created a set of highlands (corresponding to the volcanic arc of the Central-Eastern region and the eastwards migrating pulse) separated by lowlands acting as barriers. Such barriers would become less prominent as the volcanism continued during the following stages (Fig 4c-d). Thus, a set of west-east phylogeographic breaks could be expected within the TMVB despite lack of clear current geographic barriers.

2.6.2. Data and analyses needed for examining the climate-volcanism interaction

Although the scenarios described above would lead to complex patterns that may be difficult to disentangle, geographically explicit hypotheses can be plausibly constructed and tested by: (1) focusing on species for which the effect of the climate-volcanism interaction is expected to be clear; (2) using climatic and geological data to generate models for landscape genetic and comparative

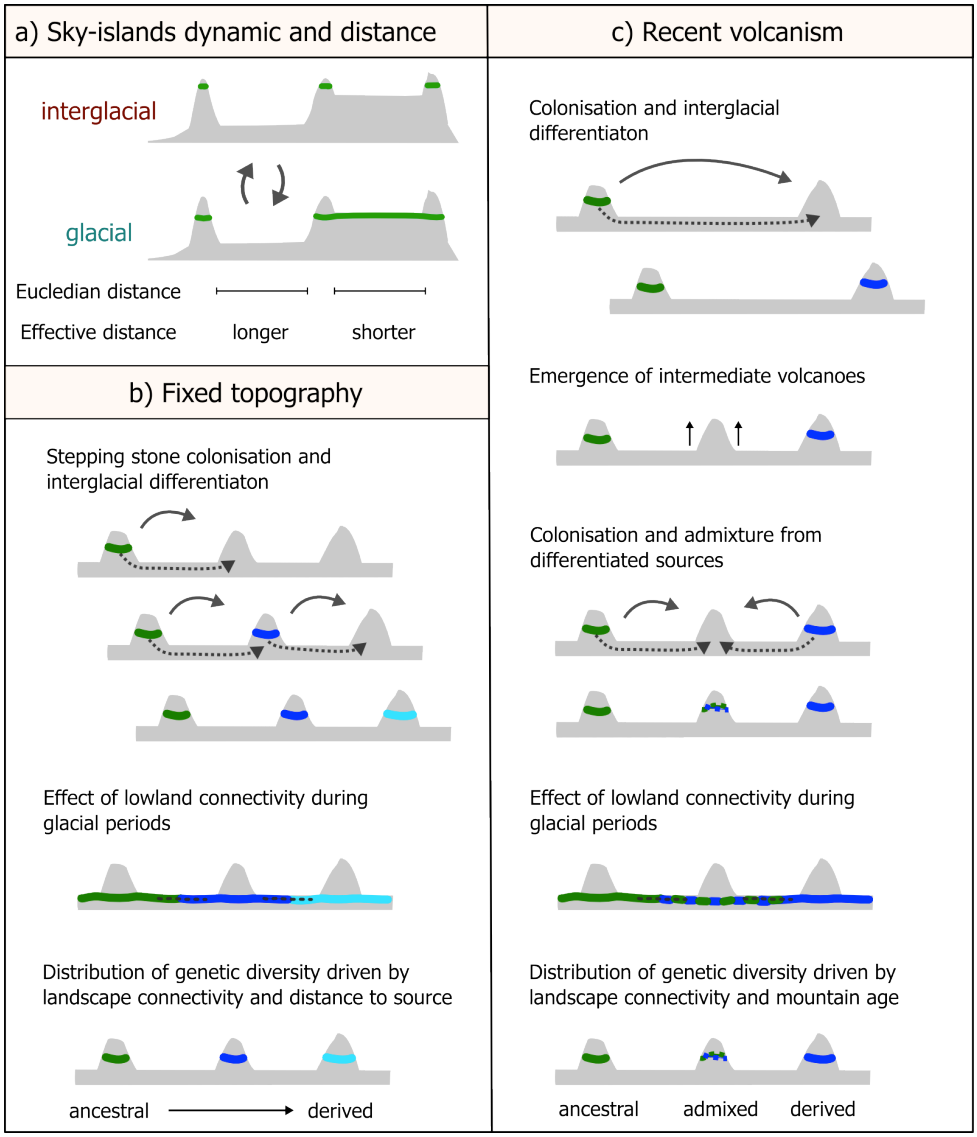


Figure 2.5. a) Sky-islands allow long term (glacial/interglacial) persistence of species within the same mountain but at different altitudes. Depending on topography, during the glacial periods there may be continuous suitable conditions connecting mountains that are isolated during the interglacials, leading to effective distance between mountains different than the Euclidean (geographic) distance. Effects of the sky-island dynamic on a fixed topography (b) and when it is being modified by recent volcanism (c). b) High altitude taxa colonises new mountains in a stepping stone fashion either by long distance colonisation (black arrows) or advancing in the lowlands during the glacial stages (dashed arrows). Populations of each mountain differentiate during the interglacials and admixes with its close neighbours during the glacials. Genetic variation is excepted to be higher in the source population and gradually decrease following the colonisation route. c) If recent volcanism modified the landscape creating a new intermediate mountain it would get colonised by multiple sources, leading to admixture. The distribution of genetic variation would thus not follow a geographic gradient, but be related also to mountain ages.

phylogeographic analyses; and (3) using multilocus genetic data with enough resolution to examine differentiation among populations, as well as hybridization and incomplete lineage sorting among populations or recently diverged species. Below we treat these three points in more detail.

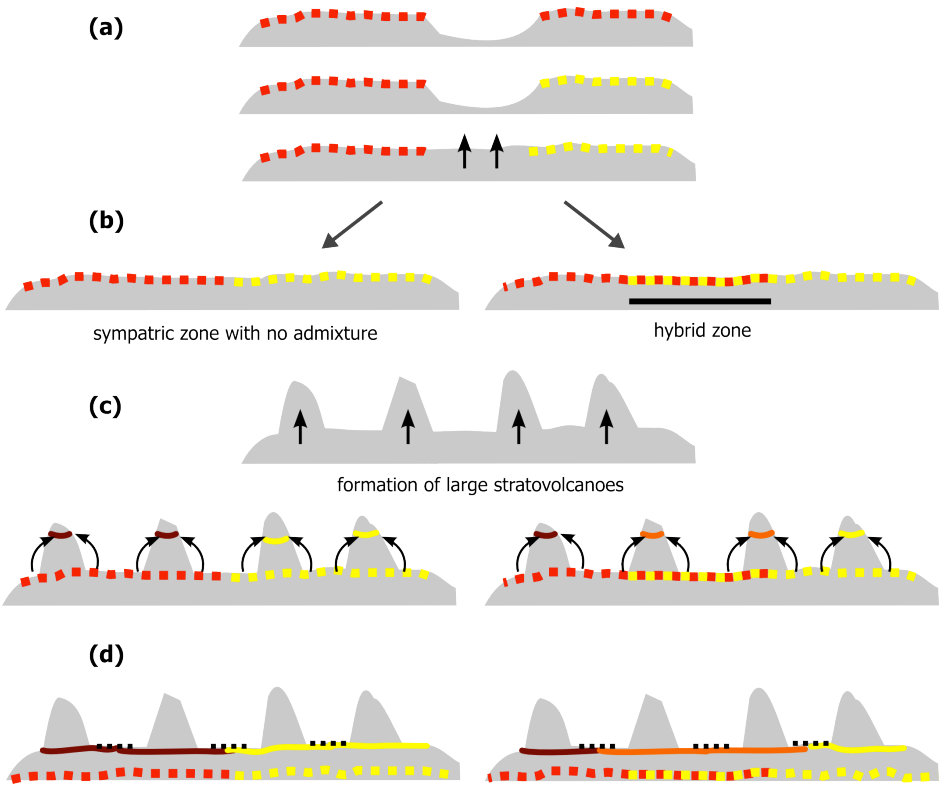


Figure 2.6. a) The first three episodes of the formation of the TMVB (Fig. 2.4a-c) formed highlands where populations of a given species (red dots) could undertake allopatric differentiation (yellow dots). Some of the volcanic episodes partially overlapped, creating continuous highlands in what used to be a geographic barrier. b) Populations in the now continuous landscape could have been sufficiently differentiated to prevent admixture (left) or generated a hybrid zone (right). c) Large stratovolcanoes emerged during the last episode of the formation of the TMVB (Fig. 2.4d). The new available habitat could have been colonised by populations of the lower land species, leading to parapatric speciation. d) Low land admixture (black dots) could increase gene flow during the glacial periods, but it may not completely erase structuring of genetic variation caused by previous landscape configurations.

2.6.3. Study species

Although the scenarios described above should hold for all montane taxa, they are expected be more pronounced for species with a higher coldness affinity. This is because (1) such species are distributed at the highest altitudes, thus enhancing the effect of topographic isolation, and (2) the high elevation of stratovolcanoes (which emerged during the most recent episode of the TMVB formation) would have served as refugia to these cold tolerant taxa during interglacial periods. For the TMVB, the uppermost vegetation is composed of grasses, shrubs and a herbaceous stratum growing above 3,900 masl, where the annual mean temperature is 3-5ºC (Rzedowski 1978; Almeida-Leñero, L. *et al.* 2007), and the timberline occurs around 3,500 masl, where open and monospecific forests of *Pinus hartwegii* gradually transition to the alpine grasslands (Calderón de Rzedowski & Rzedowski 2005; Almeida-Leñero, L. *et al.* 2007). We suggest species from these ecosystems are suitable to examine the scenarios proposed in Fig. 2.5 and Fig. 2.6. To explore signatures of recent parapatric speciation among taxa growing at different elevations (as suggested in Fig. 2.6c), the chosen species should come from neighbouring communities within the altitudinal gradient. Examples could be one of the ~25 alpine-subalpine species that are morphologically similar to species growing in ecosystems immediately below (estimated based on Calderón de Rzedowski & Rzedowski, 2005, data not shown). It is important to noting that, regardless of the scenarios to be explored, the phylogeographic signal would also depend on the dispersal characteristics and generation time of the study species.

2.6.4. Landscape and comparative phylogeographic analyses

The sky-island dynamic predictions from the scenarios of Fig. 2.5 can be tested with the aid of comprehensive volcanic age data (summarized by Gómez-Tuena et al. 2007 and Ferrari et al. 2012) and a matrix of effective distances among the TMVB mountains under the current topography. Effective distance can be estimated with spatial methods and SDM (see Alvarado-Serrano and Knowles 2013 for a review). For example, landscapes can be represented as conductive surfaces, with low resistances assigned to areas that best promote gene flow (Adriaensen 2003; McRae 2006; McRae *et al.* 2008), allowing for the incorporation of spatially explicit analyses into hypotheses testing (Chan *et al.* 2011). Areas that best promote gene flow can be modelled based on present climate and soil type variables, but also based on where suitable conditions were distributed during glacial periods.

Scenarios from Fig. 2.6 and the west-east split are more difficult to test with spatial data for two reasons. Firstly, the stages of the TMVB formation partially overlapped in the same geographic area (Fig. 2.4), thus the topography generated by the early stages may not be part of the current landscape. And secondly, it is likely that species distributions would have changed considerably since then. However, Escalante and Ocegueda (2007) and Gámez et al. (2012) suggested that western and eastern biogeographic districts may exist, and that they may be related to episodes of the TMVB formation. To further examine this under a comparative phylogeographic framework, it is necessary to look for common phylogeographic breaks and Neogene divergence times among several taxa distributed across regions of the TMVB that emerged during different episodes. This has already been partially achieved: several TMVB taxa show a

west-east division of sister lineages, or structuring of genetic variation, that are neither congruent with geographic distances among sampling sites, nor current topographic connectivity (Bryson *et al.* 2011b, 2012a; b; Bryson & Riddle 2011; Parra-Olea *et al.* 2012). However, these studies were not designed to explicitly test for the effect of the TMVB formation episodes, and geographic sampling among these studies is not comparable. A comparative phylogeographic analysis would thus need sampling of co-distributed taxa and analyses with methods that allow to test inferences across community assemblages, such as hierarchical approximate Bayesian computation (Hickerson & Meyer 2008; Chan *et al.* 2014).

2.6.5. Genetic data

The interaction of climate and recent volcanism, as simplified in Fig. 2.5 and Fig. 2.6, would likely involve complex scenarios of gene flow among populations, hybridization and incomplete lineage sorting between closely related populations and species. As a consequence, similar geographic patterns of genetic variation at given loci could be produced by different processes. Resolving the causal explanations of such genetic patterns is unlikely using traditional molecular markers. Plastid sequences provide single locus information that is not representative of how isolation and admixture affects genomes, and in the case of plants, they may not provide enough resolution for population level analyses (Zhang & Hewitt 2003). Microsatellites and other size-based methods are useful to study recent population history, but do not allow for more detailed genealogical inferences (Zhang & Hewitt 2003). Incorporating nuclear multilocus data and coalescent analyses would help to estimate divergence times more accurately (McCormack *et al.* 2011), but still they may

not be sufficient to disentangle how species genomes have evolved under the complex history of the TMVB.

An exciting alternative is to use high-resolution genomic data, which can now be applied to phylogeography and phylogenetics in a cost-effective way with a variety of methods (McCormack *et al.* 2013). This type of data can provide enough number of variable loci to examine the outcome of the climate-volcanism interaction in the TMVB, but also can be used to explore the processes of speciation in novel ways. For instance, genomic approaches have just started to be used on taxa from the Mexican highlands with interesting results. Leaché *et al.* (2013) found evidence of gene flow among a set of lizard species with allopatric and parapatric distributions, opening the door to examine whether divergence occurred with gene flow or after secondary contact. And in a study of a subalpine shrub, Mastretta-Yanes *et al.* (2014) found that loci originated by recent gene duplication events account for differentiation among populations and species, thus highlighting that divergence of isolated montane populations can be examined with alternative sources of genomic differentiation.

2.7. Acknowledgements

We are thankful to Luca Ferrari for helpful advice on the geological sections of this work and for providing the original files to generate Fig. 2.4. This work was supported by Consejo Nacional de Ciencia y Tecnología doctorate scholarship to AMY (CONACYT 213538).

Table 2.1. Phylogeographic studies undertaken for the Mexican highlands

Taxa [a]	*Marker* [b]	*Type* [c]	*Highlands* [d]					*Discussion* [e]						*Main findings*	*References*
			SMO	*SME*	*TMVB*	*SMS*	*ACh*	*1*	*2*	*3*	*4*	*5*	*6*		
Fungi															
Ascomycota > Rhytismataceae															
Lophodermium nitens	nucl.	Fa	x	x	x	x		*			*			Patterns congruent with host genetic structure	Salas-Lizana *et al.* 2011
Plants															
	Gymnosperms														
	Coniferales > Pinaceae														
Abies spp.	Alloen. mtDNA, cpDNA, cpSSRs	B	x	x	x	x	x	*	*	*		*		High population differentiation. Divergence follows a model of environmental stasis and decreased extinction rate.	Aguirre-Planter *et al.* 2000, 2012; Jaramillo-Correa *et al.* 2008
Picea chihuahuana	Alloen. cpDNA, mtDNA, AFLP	Fr	x					*		*				Fragmentation, isolation and bottle necks	Ledig *et al.* 1997; Jaramillo-Correa *et al.* 2006
Picea martinezii	Alloen.	Fr		x				*						Population collapse during Holocene warming	Ledig *et al.* 2001

Pinus strobiformis, P. ayacahuite and *P. flexilis*	cpSSRs	Fa	x	x	x	x	x	*			*	Larger connectivity within SMO than within SME y FVTM populations. Ancestral contact zones. Ecological differentiation.	Ortíz-Medrano *et al.* 2008; Moreno-Letelier & Piñero 2009; Moreno-Letelier *et al.* 2013
Pinus leiophylla	cpSSRs	Fa	x	x	x			*		*		Different expansion routes within mountain ranges.	Rodríguez-Banderas *et al.* 2009
Pinus nelsonii	cpSSRs	Fr		x				*				Demographic stasis for a long period	Cuenca *et al.* 2003
Pinus rzedowzkii, P. pinceana and *P. maximartinezii*	cpSSRs	Fr	x	x	x							High genetic diversity despite small current population size. High population differentiation.	Ledig *et al.* 2001; Molina-Freaner *et al.* 2001; Delgado *et al.* 2008
Pinus montezuame and *P. pseudostrobus*	cpSSRs	Fa		x	x	x	x	*	*			Introgressive hybridization. Long time persistence of hybrid lineage	Delgado *et al.* 2007
Pseudotsuga menziesii	Alloen. cpSSRs, mtDNA, cpDNA	Fa	x	x	x			*		*	*	Southward migration into Mexico. Niche models predict refugia. Long term isolation of Mexican populations.	Li & Adams 1989; Wei *et al.* 2011; Gugger *et al.* 2011
Coniferales> Podocarpaceae													
Podocarpus matudae	cpDNA	Fr		x	x	x	x	*	*			* Extant populations are a pre-Quaternary relict . Miocene age for temperate for a of cloud forests.	Ornelas *et al.* 2010
Coniferales > Cupressaceae													

Juniperus blancoi	cpDNA, nucl.	Fr	x		x			*	*	*				Phenotypic and habitat differences among populations. Deep divergence times between TMVB and SMO, and within SMO.	Mastretta-Yanes *et al.* 2011; Moreno-Letelier *et al.* 2014
		Angiosperms													
	Asparragales > Asparagaceae														
Nolina parviflora	cpDNA, nucl.	Fr	x	x	x	x			*				*	Correlation between TMVB formation stages and diversification times	Ruiz-Sanchez & Specht 2013
	Berberales > Berbericeae														
Berberis alpina	ddRAD	Fr	x		x			*	*					Population differentiation from orthologs correlated with private paralogs. High population differentiation.	Mastretta-Yanes *et al.* 2014b
	Cucurbitales > Begoniaceae														
Begonia heracleifolia and *B. nelumbiifolia*	cpSSRs	Fr			x	x	x	*					*	Genetic structure explained by dispersal limitation but not by expected moist glacial refugia	Twyford *et al.* 2013
	Fagales > Fagaceae														
Quercus oleoides	SSR, cpDNA	Fa		x	x		x	*						Phylogeographic breaks matching leaf morphology. "Out of the tropics" scenario hypothesized to explain expansion in the temperate zone	Cavender-Bares *et al.* 2011
Q. affinis and *Q. laurina*	RFLP	Fa		x	x	x	x	*						Latitudinal and altitudinal migrations during climate fluctuations with low gene flow among populations	González-Rodríguez *et al.* 2004
Q. crassifolia and *Q. crassipes*	cpSSRs	B	x	x	x									Introgression in hybrid areas throughout the TMVB during long periods of sympatry	Tovar-Sánchez *et al.* 2008
	Lamiales > Gesneriaceae														
Moussonia	cpDNA, ITS	Fa		x	x	x	x	*	*				*	*	Multiple refugia with populations	Ornelas & González 2014

deppeana												persisting and diverging during interglacial cycles in multiple refugia	
	Gentianales > Rubiaceae												
Palicourea padifolia	cpDNA	Fa		x		x	x	*		*		Population isolation throughout glacial cycles by the TI, but no differentiation among populations at each side of the isthmus	Gutiérrez-Rodríguez *et al.* 2011
	Liliales>Smilacaceae												
Smilax hispida spp. complex	cpDNA, nucl.	B	x	x	x	x	x	*	*			Independent colonization and speciation of Mexican spp.. Miocene origin of *S. jalapensis* and Plio-Pleistocene speciation of *S. moranensis*	Zhao *et al.* 2013
	Pteridophytes												
	Cyatheales>Cyatheaceae												
Alsophila firma	nucl. cpSSRs	Fa		x	x	x	x	*		*		Range fluctuations during the Pleistocene. Interglacial population expansion and LGM population divergence	Ramírez-Barahona & Eguiarte 2014
Animals													
	Arthropods												
	Coleoptera > Curculionidae												
Dendroctonus mexicanus	mtDNA	Fa	x	x	x	x		*	*			Demographic expansion. Complex spatial patterns	Anducho-Reyes *et al.* 2008
D. pseudotsugae	mtDNA	Fa	x	x				*				Divergence of Mexican populations relatively to USA and Canada	Ruiz *et al.* 2010
D. approximatus	mtDNA, SSR	Fa	x	x	x	x		*		*		Independent colonization of Mexico through the SMO and through the SME	Sánchez-Sánchez *et al.* 2012
Coleoptera > Zopheridae													

Zopherus, Verodes and *Phloeodes* spp.	mtDNA nucl.	B	x	x	x	x	x	*	*			Narrow niche widths lead to higher probability of fragmentation during climate fluctuations and increased speciation. Lack of extinction. Population persistence.	Baselga *et al.* 2011
Araneae > Salticidae													
Habronattus pugillis	mtDNA	S	x					*				Postglacial expansion to Arizona from the SMO	Masta 2000
		Nematodes											
Rhabdochona lichtenfels	mtDNA	Fa	x		x			*				Divergence times fell within the Pleistocene but are discussed in terms past basins connectivity due to topographic changes caused by volcanism	Mejía-Madrid *et al.* 2007
		Fishes											
	Cyprinodontiformes>Poeciliidae												
Poeciliopsis & *Poecilia* spp.	mtDNA	B			x			*				Pliocene-Pleistocene vicariance driven by volcanism	Mateos 2005
		Amphibians											
	Caudata > Plethodontidae												
Pseudoeurycea leprosa	mtDNA	Fr			x			*	*			Climate and volcanism driving population differentiation	Parra-Olea *et al.* 2012
	Caudata > Ambystomatidae												
Ambystoma leorae	SSR	Fr			x							High genetic diversity and no inbreeding within the only remaining and small population	Sunny *et al.* 2014
		Reptiles											
	Squamata > Viperidae												
C. triselatus group	mtDNA	B	x	x	x			*	*	*		Basins and low elevation areas as geographic barriers between mayor phylogroups	Bryson *et al.* 2011c

C. intermedius group	mtDNA	B	x	x	x	x		*	*	*				Less divergence between South SMO and SME than between South-North break within SMO	Bryson *et al.* 2011b
Atropoides, Bothriechis and *Cerrophidion* spp.	mtDNA	B		x	x	x	x	*	*			*		Geological events impacted divergence. Widespread within-spp. genetic structure	Castoe *et al.* 2009
	Squamata > Colubridae														
Pituophis catenifer and *P. deppei*	mtDNA	B	x	x	x				*					Uprising of SMO as geographic barrier between Sonoran and Chihuahuan deserts.	Bryson *et al.* 2011a
Thamnophis rufipunctatus spp. complex	nucl. mtDNA	S, B						*		*				SMO as an archipelago of high elevation refugia	Wood *et al.* 2011
	Squamata > Anguidae														
Barisia spp.	mtDNA	B	x	x	x	x		*	*	*			*	Old lineages (up to 11 Ma) within SMO and TMVB	Zaldivar-Riverón *et al.* 2005; Bryson & Riddle 2011
	Squamata > Phrynosomatidae														
Sceloporus spp.	RRL	B	x	x	x	x	x	*						Examples of speciation with and without gene flow among parapatric and allopatric spp.	Leaché *et al.* 2013a
S. virgatus	mtDNA	S	x					*						Colonization of Arizona Sky-Islands from SMO	Tennessen & Zamudio 2008
S. *scalaris* group	mtDNA	B	x	x	x	x		*	*	*	*			Neogene divergence times, deeper divergence than expected form taxonomy	Bryson *et al.* 2012b
S. bicanthalis	mtDNA nucl.	Fa		x	x	x								Population structure and differentiation congruent with ancient fragmentation and prolonged isolation	Leaché *et al.* 2013b
Phrynosoma orbiculare	mtDNA	Fa	x	x	x				*	*				Old lineages, varieties could be spp.	Bryson *et al.* 2012a
	Birds														

	Passeriformes > Corvidae												
Aphelocoma wollweberi spp. group	mtDNA, nucl.SSR	B	x	x	x	x		*	*	*	*	*	No niche diverge during speciation process. Divergence times fell both in the Pleistocene and the Neogene	McCormack *et al.* 2008a; b, 2010, 2011
	Passeriformes > Thraupidae												
Chlorospingus ophthalmicus	mtDNA	Fa		x		x	x	*	*			*	High population differentiation and long term isolation among mountain ranges	García-Moreno *et al.* 2004; Weir *et al.* 2008
	Piciformes > Ramphastidae												
Aulacorhynchus prasinus spp. complex	mtDNA	B	x			x	x	*					Northward expansion into Mexico from Central American populations.	Puebla-Olivares *et al.* 2008
	Apodiformes > Trochilidae												
Amazilia cyanocephala	mtDNA	Fa		x		x	x	*				*	TI driving recent diversification but allowing gene flow. Morphological and environmental niche differences. Selection strong enough to counteract the effects of gene flow	Rodríguez-Gómez & Ornelas 2014
Amazilia sensu lato spp. from Mesoamerica	mtDNA	B	x	x	x	x	x	*				*	Ancestral distribution west of TI with subsequent dispersals east of the isthmus and to S. America. The diversification related to vegetation shifts and orogenesis of Mexican and C. America highlands	Ornelas *et al.* 2014
	Mammals												
	Rodentia > Cricetidae												
Peromyscus aztecus spp. group	mtDNA	B	x	x	x	x		*	*				Early differentiation of SMO's populations from other mountain	Sullivan *et al.* 1997

											ranges	
Peromyscus mexicanus spp. group	mtDNA	B				x	x	*			Southwards migration and speciation in the highlands. Pleistocene dispersal and vicariance events.	Ordóñez-Garza *et al.* 2010
Neotoma mexicana	mtDNA	B	x		x	x		*	*	*	Divergence times and fossil records support habitat extension to lower elevations during glacial periods	Edwards & Bradley 2002
Reithrodontomys sumichrasti	mtDNA nucl.	Fa		x	x	x	x	*		*	Differentiation across the TI supported by mtDNA but not nuclear loci. Potential contact zone.	Hardy *et al.* 2013
Habromys spp.	mtDNA	B		x	x	x		*		*	*In situ* diversification	León-Paniagua *et al.* 2007
	Rodentia > Geomydae											
Thomomys spp.	mtDNA, Alloen.	B	x		x			*			North to South migration and Pleistocene divergence times in the highlands and arid regions	Mathis *et al.* 2014
Multiple for joint comparative analysis												
15 non related plants, birds and rodent spp. from cloud forests	mtDNA, cpDNA, ITS	Fa, Fr		x	x	x	x	*		*	Phylogeographic breaks are shared among different taxa but occurred at different times	Ornelas *et al.* 2013

a) Studied species. If more than three species were examined together they are abbreviated as *Genus* spp.
b) Molecular markers: simple sequence repeats (SSRs); DNA sequences from nuclear, chloroplast or mitochondrial loci (nucl. cpDNA and mtDNA, respectively); DNA sequences from the internal transcribed spacer (ITS); alloenzymes (Alloen.); amplified fragment length polymorphism (AFLP); double digest Restriction Site-associated DNA sequencing (ddRAD); restriction fragment length polymorphism (RFLP) and Reduced Representation library (RRL).
c) Whether the study focused on several closely related taxa as part of a biogeographic analyses (B) or performed phylogeographic study of a single widely distributed (Fa) or rare species (Fr).

d) Mexican highlands included in the study coded as in Fig. 2.1.
e) Discussion and findings addressed: (1) Pleistocene climate fluctuations and Quaternary divergence times, (2) Neogene volcanic activity and Pre-Quaternary divergence times, (3) a North-South phylogeographic break within the SMO (likely geographically shared among species), (4) evidence for historical connectivity between the SMO and the SME, (5) evidence of the TI acting as a geographic barrier, and (6) a West-East phylogeographic break within the TMVB (at different longitudes for different taxa).

Table 2.2. Empirical sources of paleoclimatic data for the Mexican highlands

Source of evidence	Location	Altitude (masl)	Latitude	Longitude	Time (yrs. or Period)	Findings / identified taxa or ecosystems	Reference
Packrat middens	lowlands	1,000-1,500	not provided	not provided	40,000	cooler habitats on present deserts	Betancourt *et al.* 1990
Pollen and magnetic susceptibility from lake core	TMVB	2,240	19°15′ N	99°00′ W	20,600 -18,300	reduced forests with extensive grasslands	Lozano-García & Ortega-Guerrero 1994
					18,300 and 17,500	dry and warm climate with xerophyte vegetation.	
					17,500 and 10,000	increasing moisture and cooler temperatures, strong volcanic activity	
					12,000	expansion of forests	
					Holocene	oak forest expansion	
Pollen, magnetic susceptibility and loss-on-ignition from lake core	TMVB	2,330	19°30'00"N	99°0'00"W	34,000 to ca. 23,000	humid period with mesophytic and wetland taxa	Lozano-García & Ortega-Guerrero 1998
					21,000-14,000	dry and cold, expansion pine forests and then grasslands, volcanic activity	
					Holocene	increase in precipitation and oaks	
Plant macrofossils from packrats middens	lowlands	780	33°53'24"N	113°10'12"W	LGM	woodland of pine, oaks and junipers	McAuliffe & Van Devender 1998
					Holocene	oaks, junipers and desert shrub lands	
pollen	lowlands	400	31°08'N	115°15'W	44,000	dry, montane and chaparral	Lozano-García *et al.* 2002
					44,000-34,500	humid, pines, junipers, and Artemisia	

					22,000-13,000	increment in junipers, lowering altitudinal ranges of woodland/chaparral	
					present	Desert	
Magnetic minerals, total organic carbon and pollen	TMVB	1,973	19°50′N	101°40′W	52,000 - 39,000	humid conditions	Ortega *et al.* 2002
					35,000	drier	
					21,000	drier	
					14,000-4,800	driest	
pollen	TMVB	2,570	19° 8'60"N	99°29'53"W	23,000-11,600	woodlands and grasslands	Lozano-García *et al.* 2005
					21,000 to 16,000	grasses and non-arboreal pollen, glacial advance	
					12,600	grasses and non-arboreal pollen, glacial advance	
					>10,000	tree cover increased	
					3,100	human deforestation	
					present	oak forest	
pollen	SMO	1,700	28°23′06" N	108°33′09"W	12,849 - 10,300	*Abies-Pinus* dominance	Ortega-Rosas *et al.* 2008
					10,300–9200	*Pinus–Quercus*	
					last 2,000	*Quercus–Pinus*–Cyperaceae dominance	
		1,810	28°25′39"N	108°22′47"W	6,638-1,950	*Pinus–Quercus* dominance.	
					1,950 - 1,800	*Pinus–Quercus–Abies* dominance	
					1,800 - 0	*Pinus–Quercus*–Cyperaceae dominance	
		1,945	28°22′39"N	108°23′05"W	6,445 - 5,750	*Abies–Pinus*–Poaceae dominance	
					5,750 - 4,260	*Pinus–Quercus–Abies* dominance.	
					2,990 - present	*Pinus–Quercus* dominance	
paleoecology using ostracode fauna and shell chemistry	SMO	2,200	29°15′N	107°40′W	28,465-16,342	Water temperature ranged 5-10°C	Palacios-Fest *et al.* 2002
					11,000	lake shrank. Water temperature 8.2-21.3°C	

					8,900-4,000	humidity decreased, lake became intermittent. Water reached 21.3°C	
pollen	lowlands	50	18° 5'4.00"N	94°20'33.00" W	Pliocene-Holocene	*Abies, Picea, Alnus, Celtis, Fagus, Juglans, Liquidambar, Myrica, Populus, Ulmus*	Graham 1999
pollen	lowlands	650	17°8'26"N	92°42'39"W	Beginning Miocene	*Picea, Pinus*	Graham 1998
pollen	lowlands	150	27°20'N	99°40'W	Eocene	*Ilex*	Martínez Hernández *et al.* 1980
pollen	lowlands	80	15°45'N	88°42'W	Miocene-Pliocene transition and early Pliocene	*Picea, Pinus, Quercus, Juglans, Ulmus*	Graham 1998
pollen	lowlands	780	14°43'N	89°29'W	Miocene-Pliocene transition and early Pliocene	*Picea, Pinus, Quercus, Juglans, Ulmus*	
pollen	lowlands	1,120	16°48'12"N	92°15'22"W	Miocene-Pliocene transition	*Pinus*	Mártínez Hernández 1992
pollen	lowlands	50	31°32'24"N	87°30'56"W	Middle Eocene to Holocene	*Abies, Pinus, Alnus, Betula, Carya, Castanea, Celtis, Fagus, Juglans, Liquidambar, Liriodendron, Myrica, Nyssa, Ostrya-Carpinus, Platanus, Quercus, Tilia, Ulmus*	Gray 1960
diatoms	TMVB	1,880	19°55'18"N	101°08'25"W	120,000	low lake deepness	Israde-Alcántara *et al.* 2002
					42-32,000	low lake deepness and arid conditions	
					25-18,000	lake expansion	
					8,830	fluctuating lake	
					6-2,000	very low lake deepness	
pollen	TMVB	3,860	19°12'35"N	98°39'57"W	12-10,500	glacial advance	Lozano-García & Vázquez-Selem 2005
					10,900-7,200	alpine grasslands	
					7,200-6,500	alpine grasslands and close pine forest	
					6,500-5,000	alpine grasslands as in modern (4,000 m) timberline	

Pollen, diatom, and geochemical	TMVB	2,035	19° 36′ N	101° 39′ W	38-25,000	freshest and deepest lacustrine phase of the lake	Bradbury 2000
					25-13,000	cool, deep, freshwater	
					10,000	the lake became shallower and more eutrophic	
Packrat middens	lowlands	not provided	not provided	not provided	Mid-Pleistocene	drier than the late Pleistocene	Van Devender 1990a
					Late Pleistocene (full glacial, marine isotope stage 2)	more humid and 5º to 6ºC cooler than present. *Pinus quadrifolia* and *Juniperus occidentalis* and chaparral species (now 400 km northern)	
					Mid-Pleistocene	woodland of sandpaper bush and big sagebrush (Artemisia tridentata), now in northern latitudes	
					late Pleistocene and early Holocene	*Artemisia* becomes rare	
Packrat middens	lowlands	800	not provided	not provided	late Pleistocene and early Holocene	junipers and chaparral, Mediterranean climate with at least twice the winter precipitation it receives today	Rhode 2002
					present	desert shrubs and succulents	

2.8. References

Adams RP (2008) *Junipers of the World: The genus Juniperus*. Trafford Publishing.

Adriaensen (2003) The application of "least-cost" modelling as a functional landscape model. *Landscape and Urban Planning*, **64**, 233–247.

Aguirre-Planter E, Furnier GR, Eguiarte LE (2000) low levels of genetic variation within and high levels of genetic differentiation among populations of species of *Abies* from southern Mexico and Guatemala. *American Journal of Botany*, **87**, 362–371.

Aguirre-Planter E, Jaramillo-Correa JP, Gomez-Acevedo S *et al.* (2012) Phylogeny, diversification rates and species boundaries of Mesoamerican firs (*Abies*, Pinaceae) in a genus-wide context. *Molecular Phylogenetics and Evolution*, **62**, 263–274.

Almeida-Leñero, L., Escamilla, M., Giménez de Azcárate, J., González-Trápaga, A., Cleff, A. M. (2007) Vegetación alpina de los volcanes Popocatépetl, Iztaccíhuatl y Nevado de Toluca. In: *Biodiversidad de la faja volcánica transmexicana* (eds Luna-Vega I, Morrone JJ, Espinosa D), pp. 179–198. Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Zaragoza e Instituto de Biología, México.

Alvarado-Serrano DF, Knowles LL (2013) Ecological niche models in phylogeographic studies: applications, advances and precautions. *Molecular Ecology Resources*, advance online publication.

Anducho-Reyes MA, Cognato AI, Hayes JL, Zúñiga G (2008) Phylogeography of the bark beetle *Dendroctonus mexicanus* Hopkins (Coleoptera: Curculionidae: Scolytinae). *Molecular Phylogenetics and Evolution*, **49**, 930–940.

Barrier E, Velasquillo L, Chavez M, Gaulon R (1998) Neotectonic evolution of the Isthmus of Tehuantepec (southeastern Mexico). *Tectonophysics*, **287**, 77–96.

Baselga A, Recuero E, Parra-olea G, García-París M (2011) Phylogenetic patterns in zopherine beetles are related to ecological niche width and dispersal limitation. *Molecular Ecology*, **20**, 5060–5073.

Betancourt JL, Van Devender TR, Martin PS (Eds.) (1990) *Packrat Middens. The Last 40,000 Years of Biotic Change*. University of Arizona Press, Tucson.

Braconnot P, Otto-Bliesner B, Harrison S *et al.* (2007) Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features. *Climate of the Past*, **3**, 261–277.

Bradbury JP (1997) Sources of glacial moisture in Mesoamerica. *Quaternary International*, **43–44**, 97–110.

Bradbury JP (2000) Limnologic history of Lago de Pátzcuaro, Michoacán, Mexico for the past 48,000 years: impacts of climate and man. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **163**, 69–95.

Bryson RW, García-Vázquez UO, Riddle BR (2011a) Phylogeography of Middle American gophersnakes: mixed responses to biogeographical barriers across the Mexican Transition Zone. *Journal of Biogeography*, **38**, 1570–1584.

Bryson RW, García-Vázquez UO, Riddle BR (2012a) Diversification in the Mexican horned lizard *Phrynosoma orbiculare* across a dynamic landscape. *Molecular Phylogenetics and Evolution*, **62**, 87–96.

Bryson RW, García-Vázquez UO, Riddle BR (2012b) Relative roles of Neogene vicariance and Quaternary climate change on the historical diversification of bunchgrass lizards (*Sceloporus scalaris* group) in Mexico. *Molecular Phylogenetics and Evolution*, **62**, 447–457.

Bryson RW, Murphy RW, Graham MR, Lathrop A, Lazcano D (2011b) Ephemeral Pleistocene woodlands connect the dots for highland rattlesnakes of the *Crotalus intermedius* group. *Journal of Biogeography*, **38**, 2299–2310.

Bryson RW, Murphy RW, Lathrop A, Lazcano-Villareal D (2011c) Evolutionary drivers of phylogeographical diversity in the highlands of Mexico: a case study of the *Crotalus triseriatus* species group of montane rattlesnakes. *Journal of Biogeography*, **38**, 697–710.

Bryson RW, Riddle BR (2011) Tracing the origins of widespread highland species: a case of Neogene diversification across the Mexican sierras in an endemic lizard. *Biological Journal of the Linnean Society*.

Caballero M, Guerrero BO (1998) Lake Levels since about 40,000 Years Ago at Lake Chalco, near Mexico City*1. *Quaternary Research*, **50**, 69–79.

Calderón de Rzedowski G, Rzedowski J (2005) *Flora Fanerogámica del Valle de México*. Instituto de Ecología A.C. y Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Pátzcuaro, Michoacán, México.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009) stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785 –789.

Castoe TA, Daza JM, Smith EN *et al.* (2009) Comparative phylogeography of pitvipers suggests a consensus of ancient Middle American highland biogeography. *Journal of Biogeography*, **36**, 88–103.

Cavender-Bares J, Gonzalez-Rodriguez A, Pahlich A, Koehler K, Deacon N (2011) Phylogeography and climatic niche evolution in live oaks (Quercus series Virentes) from the tropics to the temperate zone. *Journal of Biogeography*, **38**, 962–981.

Challenger A, Soberón J (2008) Los ecosistemas terrestres. In: *Capital natural de México* , pp. 87–108. CONABIO, México.

Chan LM, Brown JL, Yoder AD (2011) Integrating statistical genetic and geospatial methods brings new power to phylogeography. *Molecular Phylogenetics and Evolution*, **59**, 523–537.

Chan YL, Schanzenbach D, Hickerson MJ (2014) Detecting concerted demographic response across community assemblages using hierarchical approximate Bayesian computation. *Molecular Biology and Evolution*, msu187.

Coates AG, Jackson JBC, Collins LS *et al.* (1992) Closure of the Isthmus of Panama: The near-shore marine record of Costa Rica and western Panama. *Geological Society of America Bulletin*, **104**, 814–828.

Cohen KM, Gibbard PL (2011) Global chronostratigraphical correlation table for the last 2.7 million years, v 2007. Subcommission on Quaternary Stratigraphy (International Commission on Stratigraphy), Cambridge, England.

Cuenca A, Escalante AE, Piñero D (2003) Long-distance colonization, isolation by distance, and historical demography in a relictual Mexican pinyon pine (Pinus nelsonii Shaw) as revealed by paternally inherited genetic markers (cpSSRs). *Molecular Ecology*, **12**, 2087–2097.

Delgado P, Eguiarte L, Molina-Freaner F, Alvarez-Buylla E, Piñero D (2008) Using phylogenetic, genetic and demographic evidence for setting conservation priorities for Mexican rare pines. *Biodiversity and Conservation*, **17**, 121–137.

Delgado P, Salas-Lizana R, Vázquez-Lobo A *et al.* (2007) Introgressive Hybridization in Pinus montezumae Lamb and Pinus pseudostrobus Lindl. (Pinaceae): Morphological and Molecular (cpSSR) Evidence. *International Journal of Plant Sciences*, **168**, 861–875.

Van Devender TR (1990a) Late Quaternary vegetation and climate of the Chihuahuan Desert, United States and Mexico. In: *Packrat Middens. The Last 40,000 Years of Biotic Change* (eds Betancourt JL, Van Devender TR, Martin PS), pp. 104–133. University of Arizona Press, Tucson.

Van Devender TR (1990b) Late Quaternary vegetation and climate of the Sonoran Desert, United States and Mexico. In: *Packrat Middens. The Last 40,000 Years of Biotic Change* (eds Betancourt JL, Van Devender TR, Martin PS), pp. 134–165. University of Arizona Press, Tucson.

Devitt TJ (2006) Phylogeography of the Western Lyresnake (Trimorphodon biscutatus): testing aridland biogeographical hypotheses across the Nearctic-Neotropical transition. *Molecular Ecology*, **15**, 4387–4407.

Edwards CW, Bradley RD (2002) Molecular systematics and historical phylobiogeography of the Neotoma mexicana species group. *Journal of Mammalogy*, **83**, 20–30.

Eguiluz de Antuñano, S, Marrett R, Aranda García, M (2000) Tectónica de la Sierra Madre Oriental, México. *Boletín de la Sociedad Geológica Mexicana*, **53**, 1–26.

Escalante T, Ocegueda S (2007) Introducción. In: *Biodiversidad de la faja volcánica transmexicana* (eds Luna-Vega I, Morrone JJ, Espinosa D), pp. 5–6. Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Zaragoza e Instituto de Biología.

Farjon A, Styles BT (1997) *Pinus (Pinaceae). Flora Neotropica Monograph 75*. The New York Botanical Garden, New York.

Ferrari L (2004) Slab detachment control on mafic volcanic pulse and mantle heterogeneity in central Mexico. *Geology*, **32**, 77–80.

Ferrari L, Luna-González L (in press) Evolución geológica de la Sierra Madre Occidental. In: *Biodiversidad y Paisaje de la Sierra Madre Occidental* (eds González-Elizondo MS, González-Elizondo M, Montaño C, Cortéz). Instituto Politécnico Nacional-CONABIO.

Ferrari L, Orozco-Esquivel T, Manea V, Manea M (2012) The dynamic history of the Trans-Mexican Volcanic Belt and the Mexico subduction zone. *Tectonophysics*, **522–523**, 122–149.

Ferrari L, Rosas-Elguera J, Orozco-Esquivel M *et al.* (2005) Digital geologic cartography of the Trans-Mexican Volcanic Belt.

Ferrari L, Valencia-Moreno M, Bryan S (2007) Magmatism and tectonics of the Sierra Madre Occidental and its relation with the evolution of the western margin of North America. In: *Geology of Mexico: celebrating the centenary of the Geological Society of Mexico* (eds Alaniz-Álvarez SA, Nieto-Samaniego AF), pp. 1–39. Geologicla Society of America (Geological Society of America special papers, no. 442).

Ferrusquía-Villafranca I (1990) Provincias Bióticas con énfasis en criterios morfotectónicos.

Ferrusquía-Villafranca I (1993) Geology of Mexico: a synopsis. In: *Biological Diversity of Mexico: Origins and Distribution,* (eds Ramamoorthy TP, Bye R, Lot A). Instituto de Biología, UNAM, México, D.F.

Fjeldså J, Bowie RCK, Rahbek C (2012) the role of mountain ranges in the diversification of birds. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 249–265.

Gámez N, Escalante T, Rodríguez G, Linaje M, Morrone JJ (2012) Caracterización biogeográfica de la Faja Volcánica Transmexicana y análisis de los patrones de distribución de su mastofauna. *Revista Mexicana de Biodiversidad*, **83**, 258–272.

García E (1998) Climas (clasificación de Koppen, modificado por García).

García-Moreno J, Navarro-Sigüenza AG, Peterson AT, Sánchez-González LA (2004) Genetic variation coincides with geographic structure in the common bush-tanager (*Chlorospingus ophthalmicus*) complex from Mexico. *Molecular Phylogenetics and Evolution*, **33**, 186–196.

Gómez-Tuena A, Orozco-Esquivel MT, Ferrari L (2007) Igneous petrogenesis of the Trans-Mexican Volcanic Belt. *Geological Society of America Special Papers*, **422**, 129 –181.

González-Rodríguez A, Bain JF, Golden JL, Oyama K (2004) Chloroplast DNA variation in the *Quercus affinis–Q. laurina* complex in Mexico: geographical structure and associations with nuclear and morphological variation. *Molecular Ecology*, **13**, 3467–3476.

Graham A (1992) Utilization of the isthmian land bridge during the Cenozoic—paleobotanical evidence for timing, and the selective influence of altitudes and climate. *Review of Palaeobotany and Palynology*, **72**, 119–128.

Graham A (1998) Studies in Neotropical Paleobotany. XI. Late Tertiary vegetation and environments of southeastern Guatemala: palynofloras from the Mio-Pliocene Padre Miguel group and the Pliocene Herrería formation. *American Journal of Botany*, **85**, 1409–1425.

Graham A (1999) the tertiary history of the Northern temperate element in the Northern Latin American biota. *American Journal of Botany*, **86**, 32–38.

Graham CH, Carnaval AC, Cadena CD *et al.* (2014) The origin and maintenance of montane diversity: integrating evolutionary and ecological processes. *Ecography*, doi:10.1111/ecog.00578

Gray J (1960) Temperate pollen genera in the Eocene (Claiborne) Flora, Alabama. *Science*, **132**, 808–810.

Gugger PF, González-Rodríguez A, Rodríguez-Correa H, Sugita S, Cavender-Bares J (2011) Southward Pleistocene migration of Douglas-fir into Mexico: phylogeography, ecological niche modeling, and conservation of "rear edge" populations. *New Phytologist*, **189**, 1185–1199.

Gutiérrez-Rodríguez C, Ornelas JF, Rodríguez-Gómez F (2011) Chloroplast DNA phylogeography of a distylous shrub (*Palicourea padifolia*, Rubiaceae) reveals past fragmentation and demographic expansion in Mexican cloud forests. *Molecular Phylogenetics and Evolution*, **61**, 603–615.

Halffter G (1987) Biogeography of the Montane Entomofauna of Mexico and Central America. *Annual Review of Entomology*, **32**, 95–114.

Halffter G, Llorente-Bousquets J, Morrone JJ (2008) La perspectiva biogeográfica histórica. In: *Capital natural de México* , pp. 67–88. CONABIO, México.

Hardy DK, González-Cózatl FX, Arellano E, Rogers DS (2013) Molecular phylogenetics and phylogeographic structure of Sumichrast's harvest mouse (*Reithrodontomys sumichrasti*: Cricetidae) based on mitochondrial and nuclear DNA sequences. *Molecular Phylogenetics and Evolution*, **68**, 282–292.

Hickerson MJ, Meyer C (2008) Testing comparative phylogeographic models of marine vicariance and dispersal using a hierarchical Bayesian approach. *BMC Evolutionary Biology*, **8**, 322.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**.

Holt BG, Lessard J-P, Borregaard MK *et al.* (2013) An Update of Wallace's Zoogeographic Regions of the World. *Science*, **339**, 74–78.

Israde-Alcántara I, Garduño-Monroy VH, Ortega-Murillo R (2002) Paleoambiente lacustre del Cuaternario tardío en el centro del lago de Cuitzeo. *Hidrobiológica*, **12**, 61–78.

Jaramillo-Correa JP, Aguirre-Planter E, Khasa DP *et al.* (2008) Ancestry and divergence of subtropical montane forest isolates: molecular biogeography of the genus *Abies* (Pinaceae) in southern México and Guatemala. *Molecular Ecology*, **17**, 2476–2490.

Jaramillo-Correa JP, Beaulieu J, Ledig FT, Bousquet J (2006) Decoupled mitochondrial and chloroplast DNA population structure reveals Holocene collapse and population isolation in a threatened Mexican-endemic conifer. *Molecular Ecology*, **15**, 2787–2800.

Knowles LL (2000) Tests of pleistocene speciation in montane grasshoppers (genus *Melanoplus*) from the sky islands of western North America. *Evolution,* **54**, 1337–1348.

Knowles LL, Alvarado-Serrano DF (2010) Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: insights from coupled ecological, demographic and genetic models in montane grasshoppers. *Molecular Ecology*, **19**, 3727–3745.

Knowles LL, Carstens BC (2007) Estimating a geographically explicit model of population divergence. *Evolution*, **61**, 477–493.

Lachniet MS, Vazquez-Selem L (2005) Last Glacial Maximum equilibrium line altitudes in the circum-Caribbean (Mexico, Guatemala, Costa Rica, Colombia, and Venezuela). *Quaternary International*, **138-139**, 129–144.

Leaché AD, Harris RB, Maliska ME, Linkem CW (2013a) comparative species divergence across eight triplets of spiny lizards (*Sceloporus*) using genomic sequence data. *Genome Biology and Evolution*, **5**, 2410–2419.

Leaché AD, Palacios JA, Minin VN, Bryson RW (2013b) Phylogeography of the Trans-Volcanic bunchgrass lizard (*Sceloporus bicanthalis*) across the highlands of south-eastern Mexico. *Biological Journal of the Linnean Society*, **110**, 852–865.

Ledig FT, Capó-Arteaga MA, Hodgskiss PD *et al.* (2001) Genetic diversity and the mating system of a rare Mexican piñon, *Pinus pinceana*, and a comparision with *Pinus maximartinezii* (Pinaceae). *American Journal of Botany*, **88**, 1977–1987.

Ledig FT, Jacob-Cervantes V, Hodgskiss PD, Eguiluz-Piedra T (1997) Recent evolution and divergence among populations of a rare Mexican endemic, Chihuahua spruce, following Holocene climatic warming. *Evolution*, **51**, 1815–1827.

Li P, Adams WT (1989) Range-wide patterns of allozyme variation in Douglas-fir (*Pseudotsuga menziesii*). *Canadian Journal of Forest Research*, **19**, 149–161.

Lozano-García MS, Ortega-Guerrero B (1994) Palynological and magnetic susceptibility records of Lake Chalco, central Mexico. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **109**, 177–191.

Lozano-García MS, Ortega-Guerrero B (1998) Late Quaternary environmental changes of the central part of the Basin of Mexico; correlation between Texcoco and Chalco basins. *Review of Palaeobotany and Palynology*, **99**, 77–93.

Lozano-García MS, Ortega-Guerrero B, Sosa-Nájera S (2002) Mid- to Late-Wisconsin pollen record of San Felipe Basin, Baja California. *Quaternary Research*, **58**, 84–92.

Lozano-García S, Sosa-Nájera S, Sugiura Y, Caballero M (2005) 23,000 yr of vegetation history of the Upper Lerma, a tropical high-altitude basin in Central Mexico. *Quaternary Research*, **64**, 70–82.

Lozano-García S, Vázquez-Selem L (2005) A high-elevation Holocene pollen record from Iztaccihuatl volcano, central Mexico. *The Holocene*, **15**, 329–338.

Manea VC, Manea, M. (2006) The origin of modern Chiapanecan volcanic arc in southern Mexico infered from thermal models. *Society*, **2412**, 27–38.

Marshall CJ, Liebherr JK (2000) Cladistic biogeography of the Mexican transition zone. *Journal of Biogeography*, **27**, 203–216.

Mártínez Hernández E (1992) Caracterización ambiental del Terciarlo de la región de Ixtapa, Estado de Chiapas-un enfoque palinoestratigráfico. *Revista mexicana de ciencias geológicas*, **10**, 54–64.

Martínez Hernández E, Hernández Campos H, Sánchez López M (1980) Palinología del Eoceno en el Noreste de México. *Revista mexicana de ciencias geológicas*, **4**, 155–166.

Masta SE (2000) Phylogeography of the jumping spider *Habronattus pugillis* (Araneae: Salticidae): recent vicariance of Sky Island populations? *Evolution*, **54**, 1699.

Mastretta-Yanes A (in press) Estudios filogeográficos en la Sierra Madre Occidental: la biodiversidad y los procesos evolutivos desde el nivel genético. In: *Biodiversidad y Paisaje de la Sierra Madre Occidental* (eds González-Elizondo MS, González-Elizondo M, Montaño C, Cortéz). Instituto Politécnico Nacional-CONABIO.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014a) Data from: RAD sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Dryad Digital Repository.*, doi:10.5061/dryad.g52m3.

Mastretta-Yanes A, Wegier A, Vázquez-Lobo A, Piñero D (2011) Distinctiveness, rarity and conservation in a subtropical highland conifer. *Conservation Genetics*, **13**, 211–222.

Mastretta-Yanes A, Zamudio S, Jorgensen TH *et al.* (2014b) Gene duplication, population genomics and species-level differentiation within a tropical mountain shrub. *Genome Biology and Evolution*, evu205.

Mateos M (2005) Comparative phylogeography of livebearing fishes in the genera *Poeciliopsis* and *Poecilia* (Poeciliidae: Cyprinodontiformes) in central Mexico. *Journal of Biogeography*, **32**, 775–780.

Mathis VL, Hafner MS, Hafner DJ (2014) Evolution and phylogeography of the *Thomomys umbrinus* species complex (Rodentia: Geomyidae). *Journal of Mammalogy*, **95**, 754–771.

McAuliffe JR, Van Devender TR (1998) A 22,000-year record of vegetation change in the north-central Sonoran Desert. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **141**, 253–275.

McCormack J, Bowen B, Smith T (2008a) Integrating paleoecology and genetics of bird populations in two sky island archipelagos. *BMC Biology*, **6**, 28.

McCormack JE, Heled J, Delaney KS, Peterson AT, Knowles LL (2011) Calibrating divergence times on species trees versus gene trees: implications for speciation history of *Aphelocoma* jays. *Evolution*, **65**, 184–202.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

McCormack JE, Peterson AT, Bonaccorso E, Smith TB (2008b) Speciation in the highlands of Mexico: genetic and phenotypic divergence in the Mexican jay (*Aphelocoma ultramarina*). *Molecular Ecology*, **17**, 2505–2521.

McCormack JE, Zellmer AJ, Knowles LL (2010) does niche divergence accompany allopatric divergence in *Aphelocoma* jays as predicted under ecological speciation?: insights from tests with niche models. *Evolution*, **64**, 1231–1244.

McDonald JA (1993) Phytogeography and history of the alpine-subalpine flora of northeastern Mexico. In: *Biological Diversity of Mexico: Origins and Distribution* (eds Ramamoorthy TP, Bye R, Lot A, Fa J), pp. 681–703. Oxford University Press, New York Oxford.

McRae BH (2006) Isolation by resistance. *Evolution*, **60**, 1551.

McRae BH, Dickson BG, Keitt TH, Shah VB (2008) Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, **89**, 2712–2724.

Mejía-Madrid HH, Vázquez-Domínguez E, Pérez-Ponce de León G (2007) Phylogeography and freshwater basins in central Mexico: recent history as revealed by the fish parasite *Rhabdochona lichtenfelsi* (Nematoda). *Journal of Biogeography*, **34**, 787–801.

Metcalfe SE (2006) Late Quaternary environments of the Northern deserts and central Transvolcanic Belt of Mexico. *Annals of the Missouri Botanical Garden*, **93**, 258–273.

Metcalfe SE, O'Hara SL, Caballero M, Davies SJ (2000) Records of late Pleistocene–Holocene climatic change in Mexico — a review. *Quaternary Science Reviews*, **19**, 699–721.

Mittermeier RA (2004) *Hotspots revisited*. CEMEX.

Molina-Freaner F, Delgado P, Piñero D, Pérez-Nasser N, Álvarez-Buylla E (2001) Do rare pines need different conservation strategies? Evidence from three Mexican species. *Canadian Journal of Botany*, **79**, 131–138.

Montes C, Cardona A, McFadden R *et al.* (2012) Evidence for middle Eocene and younger land emergence in central Panama: Implications for Isthmus closure. *Geological Society of America Bulletin*, B30528.1.

Mora JC, Jaimes-Viera MC, Garduño-Monroy VH *et al.* (2007) Geology and geochemistry characteristics of the Chiapanecan Volcanic Arc (Central Area), Chiapas Mexico. *Journal of Volcanology and Geothermal Research*, **162**, 43–72.

Morán-Zenteno D., Cerca M, Duncan Kepple J (2007) The Cenozoic tectonic and magmatic evolution of southwestern México: Advances and problems of interpretation. In: *Geology of México: Celebrating the Centenary of the Geological Society of México* (eds Alaniz-Álvarez SA, Nieto-Samaniego ÁF), pp. 71–90. Geological Society of America Special Paper 422.

Moreno-Letelier A, Mastretta-Yanes A, Barraclough TG (2014) Late Miocene lineage divergence and ecological differentiation of rare endemic *Juniperus blancoi*: clues for the diversification of North American conifers. *New Phytologist*, **203**, 335–347.

Moreno-Letelier A, Ortíz-Medrano A, Piñero D (2013) Niche divergence versus neutral processes: combined environmental and genetic analyses identify contrasting patterns of differentiation in recently diverged pine species. *PLoS ONE*, **8**, e78228.

Moreno-Letelier A, Piñero D (2009) Phylogeographic structure of *Pinus strobiformis* Engelm. across the Chihuahuan Desert filter-barrier. *Journal of Biogeography*, **36**, 121–131.

Morrone JJ (2006) Biogeographic areas and transition zones of Latin America and the Caribbean islands based on panbiogeographic and cladistic analyses of the entomofauna. *Annual Review of Entomology*, **51**, 467–494.

Morrone JJ (2010) Fundamental biogeographic patterns across the Mexican Transition Zone: an evolutionary approach. *Ecography*, **33**, 355–361.

Morrone JJ, Márquez J (2001) Halffter's Mexican Transition Zone, beetle generalized tracks, and geographical homology. *Journal of Biogeography*, **28**, 635–650.

Nieto-Samaniego ÁF, Alaniz-Alvarez SA, Camprubí A (2007) Mesa Central of México: stratigraphy, structure, and Cenozoic tectonic evolution. In: *Geology of México: Celebrating the Centenary of the Geological Society of México* (eds Alaniz-Álvarez SA, Nieto-Samaniego ÁF), pp. 41–70. Geological Society of America Special Paper 422.

Ordóñez-Garza N, Matson JO, Strauss RE, Bradley RD, Salazar-Bravo J (2010) Patterns of phenotypic and genetic variation in three species of endemic Mesoamerican Peromyscus (Rodentia: Cricetidae). *Journal of Mammalogy*, **91**, 848–859.

Ornelas JF, González C (2014) Interglacial genetic diversification of *Moussonia deppeana* (Gesneriaceae), a hummingbird-pollinated, cloud forest shrub in northern Mesoamerica. *Molecular Ecology*, **23**, 4119–4136.

Ornelas JF, González C, de los Monteros AE, Rodríguez-Gómez F, García-Feria LM (2014) In and out of Mesoamerica: temporal divergence of Amazilia hummingbirds pre-dates the orthodox account of the completion of the Isthmus of Panama. *Journal of Biogeography*, **41**, 168–181.

Ornelas JF, Ruiz-Sánchez E, Sosa V (2010) Phylogeography of *Podocarpus matudae* (Podocarpaceae): pre-Quaternary relicts in northern Mesoamerican cloud forests. *Journal of Biogeography*, **37**, 2384–2396.

Ornelas JF, Sosa V, Soltis DE *et al.* (2013) Comparative Phylogeographic Analyses Illustrate the Complex Evolutionary History of Threatened Cloud Forests of Northern Mesoamerica. *PLoS ONE*, **8**, e56283.

Ortega B, Caballero C, Lozano S, Israde I, Vilaclara G (2002) 52,000 years of environmental history in Zacapu basin, Michoacan, Mexico: the magnetic record. *Earth and Planetary Science Letters*, **202**, 663–675.

Ortega-Rosas CI, Peñalba MC, Guiot J (2008) Holocene altitudinal shifts in vegetation belts and environmental changes in the Sierra Madre Occidental, Northwestern Mexico, based on modern and fossil pollen data. *Review of Palaeobotany and Palynology*, **151**, 1–20.

Ortíz-Medrano A, Moreno-Letelier A, Piñero D (2008) Fragmetnación y expansión demográfica en las poblaciones mexicanas de *Pinus ayacahuite* var. *ayacahuite*. *Boletín de la Sociedad Botánica de México*, **83**, 25–36.

Palacios-Fest MR, Carreño AL, Ortega-Ramírez JR, Alvarado-Valdéz G (2002) A paleoenvironmental reconstruction of Laguna Babícora, Chihuahua, Mexico based on ostracode paleoecology and trace element shell chemistry. *Journal of Paleolimnology*, **27**, 185–206.

Parra-Olea G, Windfield JC, Velo-Antón G, Zamudio KR (2012) Isolation in habitat refugia promotes rapid diversification in a montane tropical salamander. *Journal of Biogeography*, **39**, 353–370.

Porter SC (2000) Snowline depression in the tropics during the Last Glaciation. *Quaternary Science Reviews*, **20**, 1067–1091.

Puebla-Olivares F, Bonaccorso E, De Los Monteros AE *et al.* (2008) Speciation in the emerand toucanet (*Aulacorhynchus prasinus*) complex. *The Auk*, **125**, 39–50.

Ramírez-Barahona S, Eguiarte LE (2013) The role of glacial cycles in promoting genetic diversity in the Neotropics: the case of cloud forests during the Last Glacial Maximum. *Ecology and Evolution*, **3**, 725–738.

Ramírez-Barahona S, Eguiarte LE (2014) Changes in the distribution of cloud forests during the last glacial predict the patterns of genetic diversity and demographic history of the tree fern *Alsophila firma* (Cyatheaceae). *Journal of Biogeography*, advance online publication.

Rhode D (2002) Early Holocene Juniper Woodland and Chaparral Taxa in the Central Baja California Peninsula, Mexico. *Quaternary Research*, **57**, 102–108.

Richards CL, Carstens BC, Lacey Knowles L (2007) Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography*, **34**, 1833–1845.

Rodríguez-Banderas A, Vargas-Mendoza CF, Buonamici A, Vendramin GG (2009) Genetic diversity and phylogeographic analysis of *Pinus leiophylla*: a post-glacial range expansion. *Journal of Biogeography*, **36**, 1807–1820.

Rodríguez-Gómez F, Gutiérrez-Rodríguez C, Ornelas JF (2013) Genetic, phenotypic and ecological divergence with gene flow at the Isthmus of Tehuantepec: the case of the azure-crowned hummingbird (*Amazilia cyanocephala*). *Journal of Biogeography*, **40**, 1360–1373.

Rodríguez-Gómez F, Ornelas JF (2014) Genetic divergence of the Mesoamerican azure-crowned hummingbird (*Amazilia cyanocephala*, Trochilidae) across the Motagua-Polochic-Jocotán fault system. *Journal of Zoological Systematics and Evolutionary Research*, **52**, 142–153.

Ruiz EA, Rinehart JE, Hayes JL, Zuñiga G (2010) Historical Demography and Phylogeography of a Specialist Bark Beetle, *Dendroctonus pseudotsugae* Hopkins (Curculionidae: Scolytinae). *Environmental Entomology*, **39**, 1685–1697.

Ruiz-Sanchez E, Specht CD (2013) Influence of the geological history of the Trans-Mexican Volcanic Belt on the diversification of *Nolina parviflora* (Asparagaceae: Nolinoideae). *Journal of Biogeography*, **40**, 1336–1347.

Rzedowski J (1978) *Vegetación de México*. Limusa, México.

Salas-Lizana R, Santini NS, Miranda-Pérez A, Piñero DI (2011) The Pleistocene glacial cycles shaped the historical demography and phylogeography of a pine fungal endophyte. *Mycological Progress*. doi:10.1007/s11557-011-0774-x

Salzmann U, Williams M, Haywood AM *et al.* (2011) Climate and environment of a Pliocene warm world. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **309**, 1–8.

Sánchez-Sánchez H, López-Barrera G, Peñaloza-Ramírez JM, Rocha-Ramírez V, Oyama K (2012) Phylogeography reveals routes of colonization of the bark beetle *Dendroctonus approximatus* Dietz in Mexico. *Journal of Heredity*. doi:10.1093/jhered/ess043

Sandel B, Arge L, Dalsgaard B *et al.* (2011) The Influence of Late Quaternary Climate-Change Velocity on Species Endemism. *Science*, **334**, 660 –664.

Socorro G-E M, Martha G-E, A TF, Jorge, Lizeth R-G, Lorena LE I (2012) Vegetación de la Sierra Madre Occidental, México: una síntesis. *Acta Botánica Mexicana*, 351–404.

Sullivan J, Markert JA, Kilpatrick CW (1997) Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. *Systematic Biology*, **46**, 426 –440.

Sunny A, Monroy-Vilchis O, Fajardo V, Aguilera-Reyes U (2014) Genetic diversity and structure of an endemic and critically endangered stream river salamander (Caudata: *Ambystoma leorae*) in Mexico. *Conservation Genetics*, **15**, 49–59.

Tennessen JA, Zamudio KR (2008) Genetic differentiation among mountain island populations of the striped plateau lizard, *Sceloporus virgatus* (Squamata: Phrynosomatidae). *Copeia*, **2008**, 558–564.

Toledo V (1982) Pleistocene changes of vegetation in tropical Mexico. In: *Biological diversification in the tropics* (ed Prance GT), pp. 93–111. Columbia University Press, New York.

Tovar-Sánchez E, Mussali-Galante P, Esteban-Jiménez P *et al.* (2008) Chloroplast DNA polymorphism reveals geographic structure and introgression in the *Quercus crassifolia Quercus crassipes* hybrid complex in Mexico. *Botany*, **86**, 228–239.

Twyford AD, Kidner CA, Harrison N, Ennos RA (2013) Population history and seed dispersal in widespread Central American *Begonia species* (Begoniaceae) inferred from plastome-derived microsatellite markers. *Botanical Journal of the Linnean Society*, **171**, 260–276.

Valencia S (2004) Diversidad del género *Quercus* (Fagaceae) en México. *Boletín de la Sociedad Botánica de México*, **diciembre**, 33–53.

Vázquez-Selem L, Heine K (2011) Late Quaternary Glaciation in Mexico. In: *Quaternary Glaciations - Extent and Chronology - A Closer Look* (eds Ehlers J, Gibbard PL, Hughes P), pp. 849–861. Elsevier.

Webb SD (2006) The Great American Biotic Interchange: Patterns and Processes. *Annals of the Missouri Botanical Garden*, **93**, 245–257.

Wei X-X, Beaulieu J, Khasa DP *et al.* (2011) Range-wide chloroplast and mitochondrial DNA imprints reveal multiple lineages and complex biogeographic history for Douglas-fir. *Tree Genetics Genomes*, 1–16.

Weir JT, Bermingham E, Miller MJ, Klicka J, González MA (2008) Phylogeography of a morphologically diverse Neotropical montane species, the Common Bush-Tanager (Chlorospingus ophthalmicus). *Molecular Phylogenetics and Evolution*, **47**, 650–664.

Wendt T (1993) composition, floristic affinities, and origins of the canopy tree flora of the Mexican Atlantic slope rain forests, In: *Biological diversity of Mexico, origins and distribution*, (eds Ramamoorthy TP, Bye R, Lot A, Fa J), pp 595–680, Oxford University Press, New York Oxford.

West R (1964) Surface configuration and associated geology of Middle America. In: *Handbook of Middle American Indians* (eds Wauchope, West R), pp. 33–83. University of Texas Press, Austin, USA.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution*, **24**, 90 –101.

Witt C, Rangin C, Andreani L, Olaez N, Martinez J (2012) The transpressive left-lateral Sierra Madre de Chiapas and its buried front in the Tabasco plain (southern Mexico). *Journal of the Geological Society*, **169**, 143–155.

Wood DA, Vandergast a. G, Lemos Espinal JA, Fisher RN, Holycross a. T (2011) Refugial isolation and divergence in the narrowheaded gartersnake species complex (*Thamnophis rufipunctatus*) as revealed by multilocus DNA sequence data. *Molecular Ecology*, **20**, 3856–3878.

Zachos J, Pagani M, Sloan L, Thomas E, Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to Present. *Science*, **292**, 686–693.

Zaldivar-Riverón A, Nieto-Montes de Oca A, Laclette JP (2005) Phylogeny and evolution of dorsal pattern in the Mexican endemic lizard genus *Barisia* (Anguidae: Gerrhonotinae). *Journal of Zoological Systematics and Evolutionary Research*, **43**, 243–257.

Zhang D, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563–584.

Zhao Y, Qi Z, Ma W *et al.* (2013) Comparative phylogeography of the *Smilax hispida* group (Smilacaceae) in eastern Asia and North America – Implications for allopatric speciation, causes of diversity disparity, and origins of temperate elements in Mexico. *Molecular Phylogenetics and Evolution*, **68**, 300–311.

CHAPTER 3

Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference

A version of this manuscript has been published in *Molecular Ecology Resources*

DOI: 10.1111/1755-0998.12291

Bioinformatics, ticatla 2013

Science is an exploration of all the order than appears in the world, on the scale of our physical and mental organs, on that scale only, because neither our telescopes and microscopes nor the dizziest mathematical symbols, nor any procedure whatsoever will ever enable us to go beyond it
- Simone Weil

3. 1. Abstract

Restriction site-associated DNA sequencing (RADseq) provides researchers with the ability to record genetic polymorphism across thousands of loci for non-model organisms, potentially revolutionising the field of molecular ecology. However, as with other genotyping methods, RADseq is prone to a number of sources of error that may have consequential effects for population genetic inferences, and these have received only limited attention in terms of the estimation and reporting of genotyping error rates. Here we use individual sample replicates, under the expectation of identical genotypes, to quantify genotyping error in the absence of a reference genome. We then use sample replicates to (1) optimize *de novo* assembly parameters within the program *Stacks,* by minimizing error and maximizing the retrieval of informative loci, and; (2) quantify error rates for loci, alleles and SNPs. As an empirical example we use a double digest RAD dataset of a non-model plant species, *Berberis alpina*, collected from high altitude mountains in Mexico.

3.2. Introduction

Restriction site-associated DNA sequencing (RADseq) is a genotyping method that allows subsampling of a genome at putatively homologous locations across many individuals to identify and type single nucleotide polymorphisms (SNPs) in short DNA sequences. The method was created by Baird et al. (2008) and has been subsequently developed into a family of related approaches (also called genotyping-by-sequencing and reviewed by Davey *et al.* 2011). These

approaches can be applied to non-model organisms to potentially sequence thousands of loci for hundreds of individuals, rapidly and at low cost, regardless of genome size and previous genomic knowledge. As a result, RADseq is increasingly being used across the spectrum of evolutionary analysis, ranging from phylogenetic relationships within a genus (e.g. Jones *et al.* 2013), to genome wide association studies to identify regions under selection (e.g. Parchman *et al.* 2012; Richards *et al.* 2013), through to ecological and conservation studies (Narum *et al.* 2013).

Although the validity of RADseq data has been demonstrated, genotyping errors are to be expected. RADseq is prone to both technical and human sources of error (Table 3.1.), similar to those identified for traditional molecular markers (e.g. Bonin *et al.* 2004) and for whole genome sequencing (Pool *et al.* 2010; Gompert & Buerkle 2011). Wet lab procedures, parallel sequencing and species-specific genome properties also contribute to error in several ways (Table 3.1.), leading to variance in: (a) the total number of reads per individual; (b) the number of loci represented in each individual; (c) read count per locus; and (d) the read counts of alternative alleles at polymorphic loci (Hohenlohe *et al.* 2012). For example, differences in amplification success during the PCR step may lead to variation in the depth of coverage among loci and individuals, potentially causing locus or allelic dropout (Supporting Information 1).

The consequences of error, and statistical methods to account for it, have been widely discussed for other molecular makers, from AFLPs and microsatellites (Bonin *et al.* 2004; Pompanon *et al.* 2005; Price & Casler 2012) to whole-genome sequence data (Pool *et al.* 2010; Gompert & Buerkle 2011; Nielsen *et al.* 2011). Error can lead to incorrect biological conclusions, such as an

artificial excess of homozygotes (Taberlet *et al.* 1996), false departure from Hardy–Weinberg equilibrium (Xu *et al.* 2002), overestimation of inbreeding (Gomes *et al.* 1999), unreliable inferences about population structure (Miller *et al.* 2002) and incorrectly inferring demographic expansion from the confounding influence of low frequency error-derived SNPs (Pool *et al.* 2010). These potentially inaccurate inferences can be mitigated and accounted for if error rates are reported (Bonin *et al.* 2004; Pompanon *et al.* 2005; Pool *et al.* 2010; Davey *et al.* 2011) or incorporated into data analysis (Gompert & Buerkle 2011; Nielsen *et al.* 2011; Gautier *et al.* 2013a). However, the quantification and reporting of such errors has been largely overlooked by most recent RAD studies.

In addition to errors introduced during wet lab and sequencing procedures, errors can arise during the bioinformatic processing of RADseq data (Table 3.1.). For instance, when RAD sequences are assembled into loci and alleles, often using distance-based criteria, genotyping results will vary according to the algorithm used (Davey *et al.* 2013) (note that we refer to a *locus* as a short DNA sequence produced by clustering together unique RAD *alleles*; in turn, alleles differ from each other by small number of *SNPs*). Several assembly and genotyping tools for RADseq data have recently been released, such as *RaPiD* (Willing *et al.* 2011), *RADtools* (Baxter *et al.* 2011), graph-based distance clustering approaches (Peterson *et al.* 2012), *Stacks* (Catchen *et al.* 2011, 2013), *Rainbow* (Chong *et al.* 2012) and *pyRAD* (Eaton 2014). Within a given tool it is to be expected that different parameters and settings will result in different levels of assembly-related error. For instance, *Stacks* relies on a set of core parameters (summarized in Table 3.2) to first create sets of short-read sequences that match

(i.e. stacks) within a given threshold of nucleotide differences, and to then curate and assemble these into genotyped loci within individuals. Catchen *et al.* (2013) have explored how variation in: (1) the minimum number of raw reads required to form a stack (-*m*); (2) the number of mismatches allowed between stacks (-*M*); (3) the maximum number of stacks allowed per single locus (--*max_locus_stacks*); and (4) modulating the assumed rate of sequencing error (using a bounded SNP calling model) affect the recovery of RAD loci. To do so, they ran *Stacks de novo* pipeline using different parameter values and compared results to expectations from a reference genome. They concluded that the optimal values for these parameters will depend upon the polymorphism of the genome being analysed, the amount of sequencing error and the depth of sequencing performed. The authors recommended testing a range of parameter values in order to optimize the analysis of each RADseq dataset. However, their strategy to assess if true or erroneous loci were assembled involved a reference genome, therefore alternative criteria are needed for taxa where a reference genome is not available.

Here we show how replicates can be used to not only estimate error rates, but also to optimize the *de novo* assembly of RADseq data. The central premise is that DNA replicates derived from the same DNA should have the same genotype. Thus, after running any *de novo* assembly pipeline with different combinations of parameters, one can evaluate which settings produce both a high number of loci and low differences between replicate pairs (Supporting Information 1). Optimizing *de novo* assembly is particularly important for low coverage datasets, because it facilitates the recovery of more loci than could otherwise be reliably achieved.

To demonstrate how replicates can be used to estimate error rates and optimise *de novo* assembly, we use double-digest RADseq (Parchman *et al.* 2012; Peterson *et al.* 2012) data generated from populations of *Berberis alpina,* a non-model plant species limited to high altitude mountains in Mexico. We use the program *Stacks,* an efficient and well documented software that is increasingly being used by molecular ecologists, but the principle of comparing replicates could be applied to other assembly and genotyping tools for RADseq data. Optimizing RAD data assembly is important to achieve good results (Davey *et al.* 2013) and accounting for error is essential for the robustness of any individual study or meta-analysis. However, the approach presented here could be particularly useful when focal species lack previous genomic knowledge, and when datasets are characterised by low-coverage.

3.3. Methods

3.3.1. Study system and sampling

The focal species is *Berberis alpina* (Zamudio 2009), a diploid plant with a probable genome size of between 0.5 and 1.83 Gbp, based on values of related species (Rounsaville and Ranney, 2010). *Berberis alpina* inhabits the Transmexican Volcanic Belt (TMVB), a biodiversity hotspot for temperate forest plant species (Myers *et al.* 2000) where the species is restricted to a few mountain tops (Fig. 3.1).

Seven mountains where *B. alpina* and one where *B. moranensis* (a closely related species with which *B. alpina* potentially hybridizes) occur were sampled in the TMVB and nearby areas of the Sierra Madre Oriental (SMOr) during

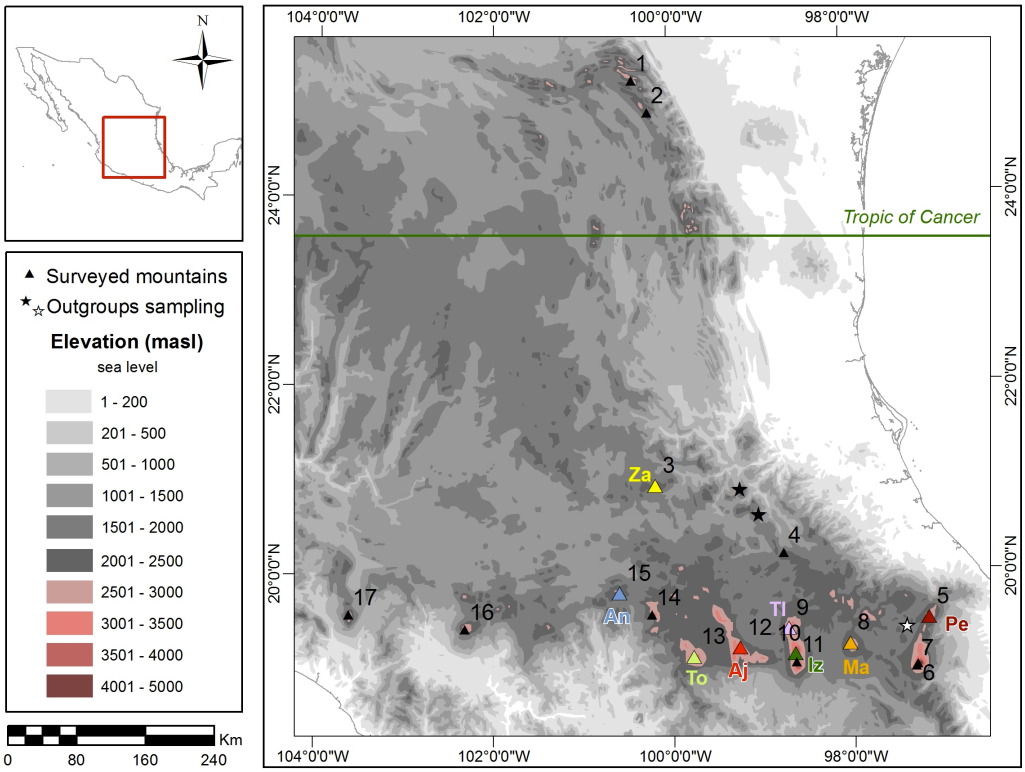


Figure 3.1. Mountains surveyed for the presence of *B. alpina* within the Sierra Madre Oriental (1-3) and the Transmexican Volcanic Belt (4-17). *B. alpina* was found on El Zamorano (Za), Nevado de Toluca (To), Ajusco (Aj), Tlaloc (Tl), Iztaccihuatl (Iz), La Malinche (Ma) and Cofre de Perote (Pe). *B. moranensis* was found on Cerro San Andrés (An). *B. pallida* (black stars) and *B. trifolia* (white star) were sampled as outgroups.

September-October 2010 and April-May 2011 (Sampling localities: doi:10.5061/dryad.g52m3). The sampling locations for *Berberis alpina* encompass the full range of the species within the TMVB (Fig. 3.1). Fresh young leaves of 6-25 specimens per mountain (depending upon population sizes) were collected and kept on ice while transported to the molecular ecology laboratory within the Instituto de Ecología, Universidad Nacional Autónoma de México (UNAM). Herbarium specimens were prepared and deposited within the Herbario Nacional in Mexico City. *Berberis pallida* and *B. trifolia* collected in the TMVB in October 2012 were used as outgroups. For each sample half the tissue was stored at -80°C at UNAM, with the remainder dried in silica gel for transport

to the University of East Anglia (UEA), England where samples were maintained at -20ºC until extraction. Samples were collected with SEMARNAT permit No. SGPA/DGGFS/712/2896/10.

3.3.2. Molecular methods

DNA extractions of *Berberis alpina* and *B. moranensis* were performed at UEA using the Qiagen DNeasy Plant Mini Kit (69106). DNA extractions of outgroup samples were performed at UNAM using a CTAB method (Vázquez-Lobo, 1996) with fresh tissue. Seventy-five specimens of *B. alpina* and *B. moranensis* (6-10 per sampling site) plus three samples of *B. trifolia* and three of *B. pallida* (outgroup species) were used to prepare double digest RAD libraries (Parchman *et al.* 2012; Peterson *et al.* 2012) using the enzymes EcoRI-HF and MseI, T4 DNA Ligase and Phusion Taq from New England Biolabs. Supporting Information 2 contains the complete lab protocol, including reaction mixes and sequencing quality details. Individual DNA extracts were randomly divided into three groups (BERL1, BERL2, BERL3), each corresponding to pools of final libraries sequenced in an independent lane. Each group was comprised of 27 *Berberis* sp. individuals and 5 replicates for a total of 32 barcoded (sequence tagged) invididuals. For each group, the 5 replicates consisted of 4 intra-library (group) replicates and 1 inter-library replicate. Replicates had the same DNA source but were processed and barcoded independently. Replicates were chosen randomly but included at least one replicate per outgroup and sampling location. Within each group of 32 barcoded individuals, positions on PCR plates were randomly selected. The digestion, ligation and PCR steps were performed in the same plate for the three groups. Samples of the same group were then pooled together and size selection

for all three groups was performed in the same gel. The three groups were each sequenced with single-end reads (100bp long) in a separate lane of an Illumina HiSeq2000, using the Lausanne Genomic Technologies Facility service provider, Switzerland.

3.3.3. Basic quality filtering and general bioinformatics pipeline

All raw reads were trimmed to 84bp because a considerable drop in quality was identified after position 85 of BERL3. Quality filtering and demultiplexing were performed with a custom Perl script equivalent to the *Stacks* program *process_radtags* (this custom script was developed prior to the release of the update of *process_radtags* that allows processing single-end double digested data). Demultiplexed data was then *de novo* assembled and genotyped using *Stacks* v. 1.02 (Catchen *et al.* 2013), first with the default settings and all samples as an exploratory run, and then with the settings and subset of samples described below for the following two experiments: (1) *exploratory analysis of Stacks key assembly parameters and SNP calling model using replicates,* and; (2) *effect of using different parameters on the output amount of data and on the detection of genetic structuring*. Trimming, demultiplexing and *Stacks de novo* assembly were performed using a computer cluster (Westmere Dual 6 core Intel X5650 2.66GHz processor systems of 12 cores with 48GB of RAM).

3.3.4. Experiment 1. Exploratory analysis of Stacks key assembly parameters and SNP calling model using replicates

We explored the effect of using different *de novo* assembly conditions and SNP calling model settings within *Stacks* on error rates and number of loci recovered.

To do so, we used the 11 replicates that sequenced successfully (yielding sufficient reads to have >50% of the mean number of loci in a first exploratory analyses of the full dataset) to run *Stacks* multiple times with a range of parameter values. For the assembly, the following key parameters were tested with the values specified in parentheses: the minimum number of raw reads required to form a stack (–*m* 2 to 15), the maximum number of mismatches allowed between stacks when processing an individual (-*M* 2 to 10), the allowed number of mismatches between loci when building the catalog (-*n* 0 to 5) and the maximum number of stacks per locus (--*max_locus_stacks* 2 to 6). Only one parameter was varied at a time while keeping the other parameters fixed to *m*=3, *M*=2, *n*=0 and *max_locus_stacks*=3. The value of –*N* was always defined as *M*+2. For the SNP calling model, we compared the default (where error rate varies freely) and the bounded model, testing different values (0.5, 0.25, 0.15, 0.1, 0.05 and 0.0056) for the upper bound (*sequencing error upper bound,* a parameter used by the bounded model: Catchen *et al.* 2013). Note that values >0.15 represent high and unrealistic levels of sequencing error. The minimum was set to 0.0056 because this was the PhiX estimate of sequencing error for BERL3 (which had the largest sequencing error of all lanes) at cycle 100 (instead of 75, to compensate for a slight quality drop at 80-84 bp). As for the remaining settings, three different minimum coverage values were explored (*m*=3, 4 and 10) and the other parameters were set to the values considered to perform better in the assembly exploratory analyses (*M*=2, *N*=4, *n*=3, *max_locus_stacks*=3, see results).

Outputs were then processed as detailed in *General processing of* Stacks *outputs* (see below) and the results were analysed in R v. 2.15.1 (R. Core Team

2012) to estimate: (1) the number of output loci and SNPs; (2) locus, allele and SNP error rates (as defined in *Error rates,* see below), and; (3) Euclidean distance matrices among individuals to build neighbour joining (NJ) dendrograms (to examine if replicate pairs cluster together, as would be expected).

3.3.5. Experiment 2. The effect of parameter values on output amount of data and the detection of genetic structuring

To examine the effect of using different *Stacks* settings on the full dataset (78 specimens) we ran *Stacks* with four *de novo* parameter profiles, namely: default, optimal, near optimal and high coverage. The default values were *m*=3, *M*=2, *N*=4, *n*=0, *max_locus_stacks*=3 and the default SNP calling model. The other parameter profiles were given values that provided the highest number of loci and SNPs at the lowest error rates in the exploratory analysis using the replicate pairs (*M*=2, *N*=4, *n*=3, *max_locus_stacks*=3 and a SNP calling model with an upper bound of 0.05, see results) but increasing the minimal coverage: *m*=3 (optimal), *m*=4 (near optimal) and *m*=10 (high coverage). Note that we define optimal as the profile that performed better in *experiment 1* for our data, and thus optimal parameter values will vary for other RADseq data. Each parameter profile was used to run *Stacks* with all individuals of *B. alpina* and *B. moranensis* (75), the three individuals of the closest outgroup (*B. trifolia*) and the replicates (14).

Outputs were then processed as detailed in *General processing of* Stacks *outputs,* and locus, allele and SNP error rates (as defined in *Error rates*) were estimated for each profile. After error rate estimation, subsequent analyses were run with only one of the replicates of each replicate pair. This dataset was used to: (1) estimate an Euclidean distance matrix based on SNPs; (2) perform a

principal coordinates analysis (PCoA) based on the distance matrix to summarise data into the four first eigenaxes that account for 90% of the total variance; (3) normalize the distance matrix and extract the distances between individuals of the same sampling location; and (4) run the population program of *Stacks* to estimate F_{ST} between population pairs using only samples from *B. alpina* and *B. moranensis*.

3.3.6. General processing of Stacks *outputs*

Stacks outputs from experiments 1 and 2 were imported to a desktop computer, where data was visualized and exported as allele and coverage matrices. These matrices were then analysed with R to: (1) estimate the number of reads and coverage per locus, per individual and per lane; (2) filter data to keep only those individuals having more than 50% of the mean number of loci per individual, and only those loci present in at least 80% of the barcoded individuals; and (3) output loci and individuals that passed the previous filter as plink format. Further analyses were performed as described above for each experiment.

3.3.7. Error rates

Replicate pairs were used to estimate three error rates using R: (1) locus error rate, corresponding to missing data at the locus level and measured as the number of loci present in only one of the samples of a replicate pair, divided by the total number of loci found; (2) allele error rate, calculated as the number of allele mismatches between replicate pairs, divided by the number of loci being compared; and (3) SNP error rate, measured as the proportion of SNP mismatches between replicate pairs.

Note that we refer to a locus as a short DNA sequence produced by clustering together unique RAD alleles; in turn, alleles differ from each other by a small number of SNPs. We define a missing locus as absent in at least one sample of a replicate pair, but present in any other individual of the dataset. In addition to the locus error rate, we further examined the distribution of missing data within replicate pairs by estimating: (1) the number of missing loci per replicate pair; (2) the proportion of missing loci (number of missing loci per replicate pair over the total); and (3) the percentage of missing loci of a given replicate that were not the same missing loci in the other replicate (proportion of missing loci different within a replicate pair). Supporting Information 1 provides a diagram detailing the differences between replicates estimated here.

The R scripts utilized here used the packages: adegenet_1.3-7 (Jombart 2008), ape_3.0-8 (Paradis *et al.* 2004), gtools_2.7.1 (Warnes *et al*. 2013), multicore_0.1-7 (Urbanek 2011) and stringr_0.6.2 (Wickham 2012).

3.4. Results and discussion

3.4.1. RAD sequencing output and coverage

An average of 1,632,914 reads per tagged-individual were obtained after demultiplexing, with no major differences between lanes or sampling localities. Full details of sequencing output are provided in Supporting Information 2. In a first exploratory analysis (using *Stacks* default settings and post-filtering the data with the >50% and 80% criteria described in the basic quality exploration section), fifteen out of the 96 samples had too few reads and therefore did not pass the filter of sharing >50% of the mean number of loci with the rest of the

individuals. Among these were the interlibrary replicate sequenced in lane BERL1 (PeB01_ir1) and one sample of a replicate pair (MaB21). Also, a strong lane effect associated with lane BERL3 was found. Samples sequenced from this lane clustered together within a NJ dendrogram, while the samples from BERL1 and BERL2 were intermixed, clustering typically by geography. The source of the lane effect was determined to be a single SNP found in position 70 of many reads, which was then identified as an artefact by the sequencing service provider. Deleting position 70 in all the demultiplexed reads removed the lane effect.

In general, mean coverage per locus was low (increasing the min. coverage -*m* from 3 to 10 produced a substantially lower number of loci, Fig. 3.2 and Table 3.3). For *Stacks,* coverage is the main filter to distinguish sequencing error from real variation. However, if coverage is generally low, a high filter threshold for coverage can lead to allele dropout, which in turn becomes genotyping error. Assembling and genotyping a low coverage RADseq dataset like that of *Berberis* is thus challenging, and may lead researchers to keep only a small fraction of the loci and alleles that have high coverage for all individuals which, as shown below, may not be the most reliable data. Many RADseq datasets may have low coverage, particularly for species for which genome size is unknown, or if a study design aims for more individuals or loci to increase the accuracy of population genetic parameters (Buerkle & Gompert 2013).

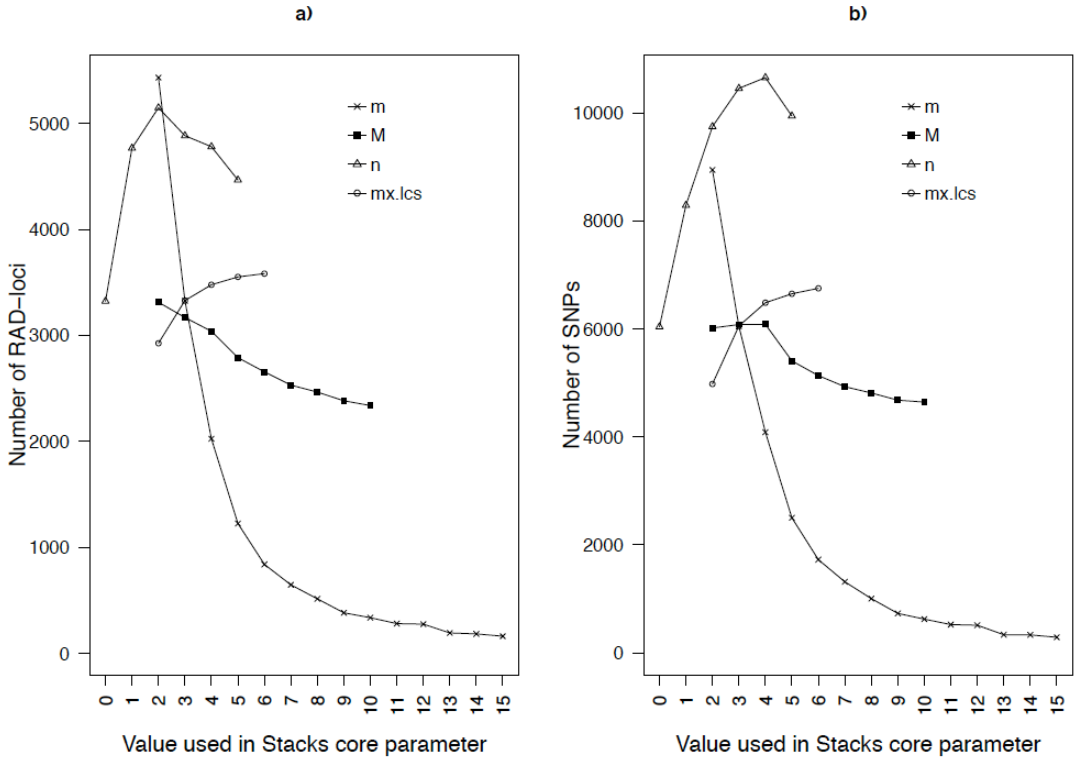


Figure 3.2. Total number of (a) RAD-loci and (b) SNPs obtained using different *Stacks* core parameters settings. For each run only one parameter varied, with the remaining set to *m*=3, *M*= 2, *n*=0 and *max_locus_stacks* (mx.lcs) = 3 and *N*= *M*+2.

3.4.2. Exploratory analysis of Stacks assembly parameters and SNP calling model using replicates

We ran *Stacks* with 11 replicate pairs (22 samples). After filtering the output so that all individuals shared >50% of the mean number of loci per individuals, most assembly parameter profiles recovered 19-20 samples and only runs with *n*≥3 recovered all 22. The samples that were not recovered for some of the parameter profiles explored for *Stacks* either had a small number of reads relative to other individuals, or belonged to the more distant outgroup (*B. pallida, OutBs*). These samples shared <50% of the mean number of loci with the remainder of the dataset and thus were excluded by the filtering step. When both samples of a replicate pair passed filtering, they clustered together in the NJ

dendogram (Supporting Information 3), with two exceptions: (1) the interlibrary replicate (PeB01) pair clustered together in only 18 of 36 parameter profiles tested, and in the remaining analyses it formed a paraphyletic group with other samples from the same sampling location, and (2) one replicate pair (AjB21) did not cluster together in 9 occasions, with each replicate clustering instead with samples from another locality. Importantly, the parameter profiles at which incorrect clustering occurred were high values for minimal coverage (-*m*) and the number of mismatches between loci when processing an individual (-*M*). This suggests that setting –*m* too high can lead to locus/allele dropout large enough to cause incorrect inferences of individual differentiation. It is less evident why setting -*M* to high values causes differences between replicates, but it is likely related to overmerging (e.g. merging paralogs as a single locus), leading to the formation of nonsensical loci (Catchen *et al.* 2013). The fact that not all replicates pairs clustered together indicates that differentiation among individuals should be interpreted with care. However, this only occurred with some parameter values, indicating that assembly settings can be tuned to minimize differences between replicates.

Across all explored parameter profiles, the number of loci recovered ranged from ~200 to >5,000 (Fig. 3.2a), the number of SNPs ranged from ~200 to >8,000 (Fig. 3.2b), and the total number of missing loci ranged from 50 to >500 (Fig. 3.3a). In general the parameters that control the minimal coverage (-*m*) and number of mismatches allowed between loci when building the catalog of loci (-*n*) contributed most to the variance of the amount of data (Fig. 3.2a) and missing loci (Fig. 3.3a and 3b). A key source of variation between replicate pairs is that the identity of most (>70%) of the missing loci in a given replicate

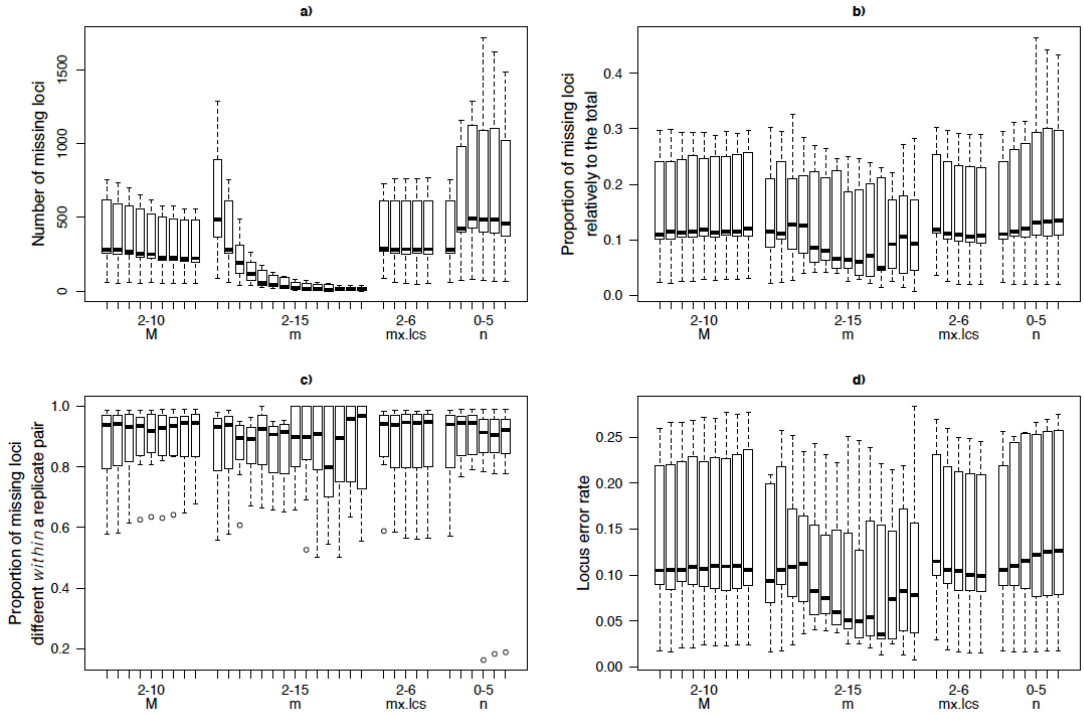


Figure 3.3. Effect of different values for *Stacks* core parameters on missing data. In each run only one parameter varied (shown on the x axis), with settings for the remainder as explained in Fig. 3. (a) total number of missing loci, (b) proportion of missing loci relative to the total, (c) proportion of missing loci different within a replicate pair and (d) locus error rate. See Supporting Information 1 for a diagram detailing the meaning of these estimates.

are not the same in the corresponding replicate (Fig. 3.3c), which leads to a locus error rate typically >10% (Fig. 3.3d) regardless of the parameter values used. As these differences are between samples from the same DNA source that were processed together, it seems that stochastic PCR/sequencing sampling events and imprecise size selection are the main sources of heterogeneous coverage among loci.

Allele error rates ranged from ~5% to >15%, depending on the parameter profile used to execute *Stacks* (Fig 4a). Allele mismatches between replicates can be caused by allelic dropout, or by the acceptance of error-based variation (likely enhanced by PCR duplicates) during assembly. Similarly, the SNP error rate ranged from ~2% to 12% (Fig. 3.4b). Again, the most important

differences were related to changes in –*m* and –*n*. Increased values of –*m* decreased the allele error rate, but not to a level below 10%, and at a cost of yielding fewer loci. Similarly, the SNP error rate was reduced from ~7% at *n*=0 to ~2.5% at *n*=3.

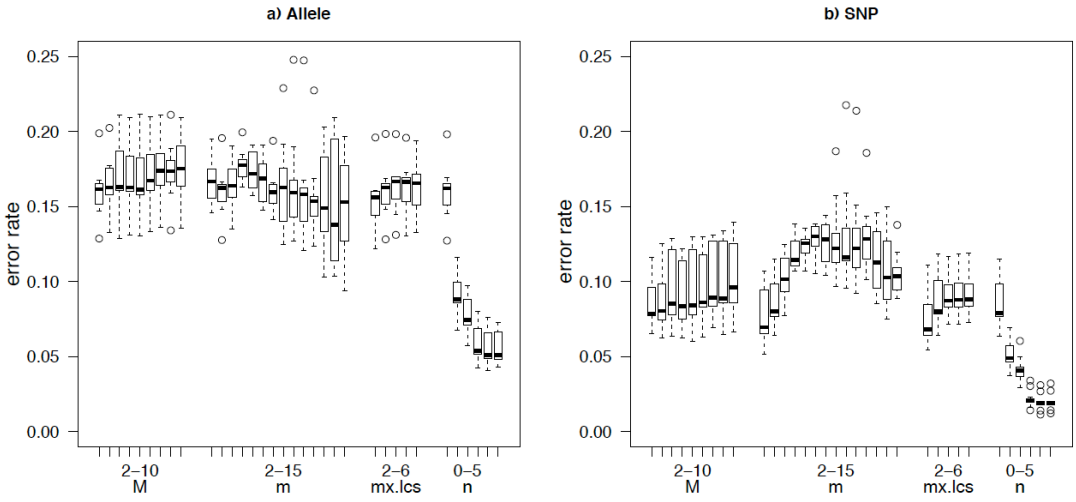


Figure 3.4. Effect of different values for *Stacks* core parameters on (a) the allele error rate and (b) the SNP error rate. In each run only one parameter varied (shown on the x axis) with settings for the remainder as explained in Fig. 3.1.

The parameter –*m* controls the total number of raw reads per individual to create a stack, so the higher it is set, the lower is the probability that there will be enough reads per locus to assemble an allele. Setting -*m* to a higher value could also result in genuine alleles being considered as secondary reads (reads that are not used to assemble reference alleles and that are set aside), and as a consequence treated as sequencing errors (see *Stacks* documentation for further details). For the *Berberis* dataset, the danger of labelling stacks with concurrent sequencing errors is reduced by the fact that the data was run in three different lanes with a randomized sample design.

The parameter –*n* modulates the maximum number of mismatches

allowed between loci when building the catalog (this is a list of all loci and alleles in the population). If *n*=0, there would be loci represented independently across individuals that are in reality homologous alleles of the same locus. When *n*>0, *Stacks* uses the consensus sequence from each locus to attempt to merge loci (*Stacks* documentation). Increasing -*n* may have resulted in significant error reduction for the *Berberis* data set because replicates involved samples from geographically isolated localities and outgroups, conditions that would be expected to result in loci that exhibit fixed differences among populations. By merging fixed alleles into a single locus the allele error rate decreased, probably because the chances of assembling the same true alleles in both replicates increased. A potential negative consequence of a high value of –*n* is the creation of erroneous loci, which can be assembled for reasons such as the acceptance of sequencing/PCR error-based stacks, and the clustering of repetitive sequence regions or paralogs (Catchen et al. 2013). However, the locus error rate did not vary significantly when –*n* varied from 0 to 5 (Fig. 3.3d), so it seems that the erroneous loci that were potentially created have less weight than the error reduction benefits gained from increasing the value of –*n*.

Regarding the SNP calling model, reducing the upper bound increases the chance of calling true heterozygous loci instead of wrongly labelling them as homozygous loci with sequencing error (Catchen et al. 2013). For the *Berberis* data, differences in genotyping errors were found only after decreasing the upper bound down to 0.0056 in the runs of *m*=3 and *m*=4 (Supporting Information 4), such that the allele error rate decreased from >5% down to approximately 2.5%. However, this increased the SNP error rate from ~2.5 % to 7%. Thus, for the *Berberis* dataset, it seems best to leave the upper bound of the

SNP calling model to a relatively high value. Finally, there were no differences in loci error rate between the SNP calling models (Supporting Information 4b).

In summary, for the *Berberis* dataset the parameter values that seemed to both increase the number of loci and reduce the SNP and allele error rates were *m*=3, *M*=2, *N*=4, *n*=3, *max_locus_stacks*=3 and a SNP calling model with an upper bound of 0.05.

3.4.3. Effect of using different parameters on the output amount of data and on detection of genetic structuring

The four combinations of *Stacks* settings (optimal, near optimal, default and high coverage) used to process the full dataset differed in the number of recovered loci, number of SNPs and error rates (Table 3.3). Among the four combinations, the optimal profile generated the highest number of RAD-loci (6,292) and SNPs (11,057) and had the lowest allele (5.9%) and SNP (2.4%) error rates, although the locus error rate (17%) was high (Table 3.3). The smallest locus error rate was found with the high coverage setting (8.8%, Table 3.3), but this parameter profile produced the highest allele and SNP error rates (8.7% and 5.7% respectively) and the smallest number of loci and SNPs (292 and 502, Table 3.3). Therefore without the replicates meaningful biological variation would have been discarded, and the data would have been assembled with settings that did not minimise error rate.

The SNP error rate is important for population genetic and phylogeographic analyses. As SNP error increases within a given dataset, so does the contribution of noise to the genetic distance between individuals. From a drift-mutation-migration equilibrium perspective, individuals collected from the

same geographic region should be expected to be genetically more similar in datasets with smaller SNP error rates. As a simple way to test this, the genetic distances between individuals from the same sampling locality were compared among the four combinations of *Stacks* settings explored here. As expected, the data with the smallest SNP error rate (optimal profile) systematically produced shorter genetic distances between individuals of the same sampling locality, when compared to the other three parameter profiles (Fig. 3.5).

To be of relevance for population genetics and phylogeographic analyses, molecular markers must not only have minimal noise, but also provide meaningful variation (Zhang & Hare 2012; Price & Casler 2012). The *Berberis* data produced by the optimal parameter profile resulted in substantial genetic variation, 80% of which was explained by the first two axes of the PCoA, which clustered samples by sampling locality (Supporting Information 5). The same axes of the PCoAs produced with the data from the high coverage and default parameter profiles explained only 47% and 57%, respectively (Table 3.3, Supporting Information 5). Also, the mean value of the pairwise F_{ST} matrix was higher (0.19) for the data with the smallest SNP error rate and larger number of loci (optimal parameter profile) compared to the default (0.07) or any of the other *Stacks* settings examined (Table 3.3). This is congruent with simulations that show that low coverage datasets with a larger sample of sites in the genome yield more accurate and precise population genetics parameter estimates (Buerkle & Gompert 2013).

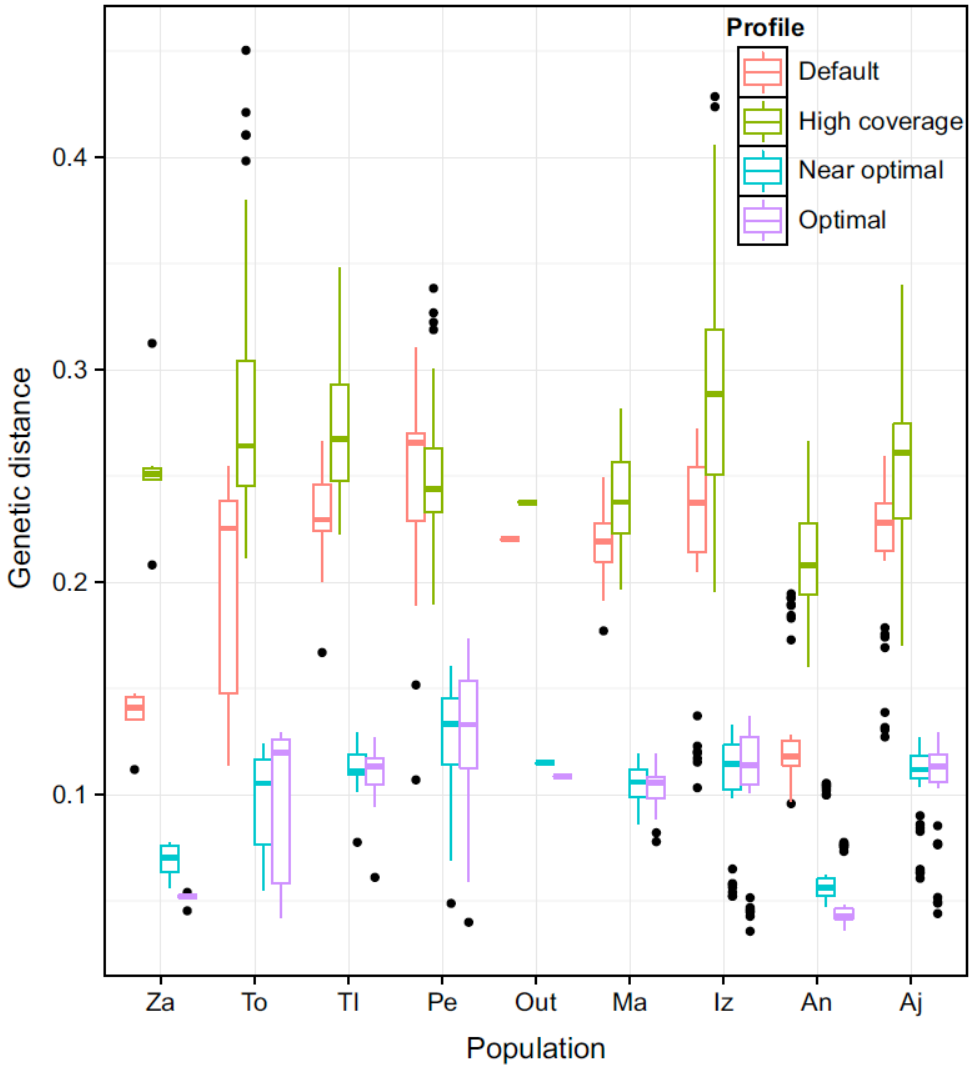


Figure 3.5. Effect of different *Stacks* profiles on the genetic distance between individuals of the same sampling location using default values and settings that were considered to perform better in the exploratory parameter analyses, but varying the minimum number of raw reads required to form a stack to: *m*=3 (optimal), *m*=4 (near optimal) and *m*=10 (high coverage).

Assembling *Berberis* data *de novo*, with the optimal parameter profile, maximised the number of informative SNPs and minimized the error that increases intra-population variation (Fig. 3.5). Regardless of the *de novo* assembly tool of choice, we advise researchers (particularly those working with previously unexplored genomes) to include replicates and follow the principles presented here (explore a range of parameter values and choose those that both

increase the number of output loci and reduce the SNP and allele error rates). In the case of RADseq datasets already produced without DNA replicates, we recommend the exploration of a range of parameter values to maximize the amount of SNPs recovered and minimize the genetic dissimilarity between individuals from the same sampling locality. This recommendation should be used as a starting point and with care, as locality may be the wrong metric to use when minimizing genetic dissimilarity in some cases (e.g. hybrid zones, breeding areas).

3.4.4. De novo *assembly tools and replicates*

RADSeq is ideal to generate genomic datasets for species for which no reference genome is available, making *de novo* assembly a crucial step of data processing. Comparative analyses of some of the available bioinformatic tools show that RAD data is reliable, but that it presents special issues that are not fully addressed by existing genotyping tools (Dou *et al.* 2012; Davey *et al.* 2013; Eaton 2014). By comparing *de novo* assembly outputs against a reference genome, Catchen *et al*. (2013) found that there may be substantial variance in the amount and quality of data recovery using different settings within *Stacks*. Using replicates in lieu of a reference genome, we also observed this variance (Figs. 3.2 & 3.4), and were able to optimise parameter values. We focused on *Stacks*, but the principle of comparing replicates can be applied to evaluate, and reduce, the amount of error produced by different assembly tools in the absence of a reference genome. However, it should be pointed out that there is no single best bioinformatic method to handle RAD data (Davey *et al.* 2013). A useful alternative to current tools would be the further development of approaches that use probabilistic

base calling (e.g. Li *et al.* 2009; McKenna *et al.* 2010), that would allow uncertainty to be incorporated into the assembly process.

3.4.5. Error rate implications and recommendations for RADseq analyses

Next-generation sequencing methods applied to population genetic inference need to account not only for sequencing error, but also for assembly error and missing data (Pool *et al.* 2010; Davey *et al.* 2011). Including DNA replicates in the preparation of RADseq libraries (see below for some recommendations) improves the characterisation of error derived from different sources (Table 3.1.) and provides the ability to partition error into locus, allele and SNP rates. High locus error rates, such as the >10% error for all combinations of parameters evaluated for *B. alpina* (Fig. 3.3d, Table 3.3) can be accommodated as missing data and mitigated by appropriate statistical corrections (Pool et al. 2010; Davey *et al.* 2011), as is possible with principal components analysis, principal coordinates analysis and STRUCTURE (Pritchard et al. 2000). However, incorrect SNP calling and allelic dropout are more problematic if data analyses are to be performed under the assumption that genotypes are known with complete certainty. Allele error can affect both allele frequency estimates and the accurate discrimination of different genotypes (Bonin et al. 2004), with the concomitant inflation of nucleotide diversity and skewing of the SNP Frequency Spectrum toward rare SNPs (Johnson & Slatkin 2008; Pool et al. 2010), thus affecting the meaningful biological interpretation of data. Excitingly, as population genomics and next-generation sequencing technology and analytical tools further develop, genotype uncertainty could be incorporated into the data analysis itself (Nielsen *et al.* 2011; Buerkle & Gompert 2013), using Bayesian

hierarchical models and genotype probabilities rather than genotypes *per se* (Gompert & Buerkle 2011; Nielsen *et al.* 2011; Buerkle & Gompert 2013; Gautier et al. 2013a). If DNA replicates are included for error rate estimation, genotype uncertainty could account not only for sequencing error, but also for the full range of sources that may affect RADseq (Table 3.1.).

The estimation of genotyping error is affected by sample size, as exemplified by the variance of error rate estimation across replicates for the *Berberis* data (Fig. 3.3, Fig. 3.4 and Table 3.3). Including multiple replicates is thus useful, but there is no minimum number for RADseq studies. For *B. alpina,* we aimed to replicate ~15% of samples, but as some failed we achieved 11%. The number of replicates for a given study will be a function of the final use of the data, the targeted coverage depth, and the precision in error rate estimation needed. Replicates should be randomly chosen while also broadly representing important data features such as geography and taxonomy. In the case of geographic sampling, we would recommend the inclusion of at least one replicate per sampling location. In addition to including replicates in the final dataset, replicates could be particularly useful during trial stages, as a way of evaluating the success of a given bench protocol.

Regarding recommendations to reduce error rate, as has been suggested for traditional molecular markers (e.g. Bonin *et al.* 2004, Pompanon *et al.* 2005), good lab practice and experimental design will help to minimize error rate. In the case of RADseq data, locus and allele recovery depend on the level of coverage of reads for each allele, locus and individual, but as shown here large numbers of markers can be recovered reliably from relatively low coverage datasets (down to ~7x, as the mean for BERL1 here). Thus, given budget

limitations, coverage depth may be traded-off for increased sampling for the number of individuals or sites in the genome, both of which can provide better estimates of population genetic parameters (Buerkle & Gompert 2013). However, studies that require very low error rates should consider increasing the coverage up to 60x (Davey *et al.* 2011). Using automated size selection methods (e.g. Pippen Prep) reduces variance among size-selected libraries, thus decreasing the amount of missing data at the wet lab stage. As we have shown here, error rates can also be reduced during *de novo* assembly by using an optimal combination of parameter values. Other recommendations have been provided elsewhere (Davey *et al.* 2013).

The acceptable error rate for RADseq studies will be case specific. In the case of *B. alpina*, the quantification of allele and SNP error rates found for the optimized *Stacks* settings (5.9% and 2.4%, respectively, Table 3.3) provides reassurance that the geographic structuring of genetic variation is biologically meaningful, but would warn against more fine-scale analyses of individual relatedness if differences between individuals fell within the error rate threshold.

Estimating error rates for a low coverage dataset allows for the recovery of more loci than could otherwise be reliably achieved, and comparing replicates can be used to aid *de novo* assembling and to validate variation. Thus, including replicates can prove particularly useful for low coverage datasets and for species lacking a reference genome. We suggest that the use of replicates for *de novo* RADseq studies should be encouraged and we consider it pertinent to extend Crawford *et al.*'s (2012) call for more transparent reporting of genotyping error to RADseq data.

Table 3.1. Potential causes of genotyping error for RADseq data

Source	Reason	Reference*
Technical and human error		
Technical	Errors related inversely to the quality of reagents and equipment, and to the organization of the laboratory in different rooms to avoid contamination	A
Human	Sample mislabelling, sample contamination, pipetting error and error during DNA concentration measurements	A
Wet lab		
Enzyme sensitivity to DNA quality and quantity	Digestion and PCR efficiency may be uneven among samples, which can result in the underrepresentation of some restriction fragments	A
Pooling concentration	Samples with higher concentration can be overrepresented in the sequencing output if they are not pooled in equimolar amounts	B, C
PCR error	PCR error may get further amplified and can appear in multiple reads resembling an alternative allele at a locus. PCR error may differ among samples depending on reaction conditions and experimental design	E
PCR bias	PCR amplification success may be variable across different alleles or barcodes, biasing their representation. Differences in amplification success lead to variation of coverage among loci and individuals, potentially resulting in allelic dropout, non-representation of some loci, or PCR duplicates	A, C, E
Size selection (double digest)	Different fragments may be selected if more than one excision is performed. Imprecise size selection can include fragments of lengths relatively distant from the size-selection target mean	C
Exposure to UV light	Can produce fragmentation (that could lead to locus/allele dropouts) and mutation of DNA strands (that introduces non-biological variation)	F

Next Generation Sequencing (NGS)		
Sequencing error	NGS introduces sequencing error (0.1–1.0% per nucleotide), that can vary across samples, RAD sites and positions in the reads for each site.	E, G
Sequencing sampling	The sampling process of a heterogeneous library inherent in NGS introduces sampling variation in the number of reads observed across RAD sites as well as between alleles at a single site	E
Barcode error	PCR or sequencing errors at the barcode of a fragment can reduce the number of reads obtained for it	E
Genome intrinsic		
GC content	At large numbers of PCR cycles RAD loci with high GC content are sequenced at higher depths compared to RAD loci with low GC content. But at the same time, high GC content loci could be under-sequenced if too few PCR cycles are performed. GC bias contributes to PCR duplicates	D
Restriction site variation	Variation in the restriction site within a locus will result in allelic dropout	D, H
DNA methylation	For some restriction enzymes digestion is impaired or blocked by methylated DNA. The same gene may or may not be methylated in different individuals or tissues	I
Bioinformatic		
Variation in coverage	Coverage is an important filter to distinguish real variation from sequencing errors, repetitive regions and duplicates. But if there is coverage heterogeneity among samples and alleles, or if the general coverage is low, setting the filters with minimal coverage values too high can lead to allele dropout. Setting it too low, however, can lead to incorrect SNP calls	E, D, J
PCR duplicates	PCR duplicates occur when more than one copy of the same original DNA molecule attaches to different beads/cells during sequencing. This can result in high coverage of PCR-error variation, or it can produce heterogeneous coverage distribution due to GC and	D

	PCR bias.	
Fragment length	Alleles will drop out as restriction fragment length decreases because RAD loci from short restriction fragments have low read depths. The efficacy of different bioinformatics tools at dealing with this varies.	D
Paralogs and repetitive regions	Paralogous and repetitive regions with similar sequences can be erroneously merged together as a single locus	E, K
Presence of indels	*Stacks* and *RADtools* are unable to handle indels, therefore indel-containing loci are not clustered together, while they can be recovered by *RaPiD* and *pyRAD*	C, D
Mapping using a reference genome	Mapping of alleles that are different from the reference genome is less probable than for a reference-matching allele, causing a bias in allele frequency toward the allele found in the reference sequence. It may additionally reduce the number of SNPs discovered and bias estimates of nucleotide diversity toward smaller values	L

* References: A) Bonin *et al*. 2004 B) Baird *et al*. 2008 C) Peterson *et al*. 2012, D) Davey *et al*. 2013 E) Hohenlohe *et al*. 2012 F) Grundemann & Schomig 1996, G) Meacham *et al*. 2011; Nielsen *et al*. 2011; Loman *et al*. 2012, H) Gautier *et al*. 2013b, I) Roberts *et al*. 2010, J) Catchen *et al*. 2013, K) Dou *et al*. 2012 and L) Pool *et al*. 2010.

Table 3.2. Role of *Stacks* core parameters in the assembly of loci and potential sources of genotyping error

Parameter	How it affects assembly and genotyping error *
minimum number of identical, raw reads required to create a stack (–*m*) default 3	Reads with convergent sequencing errors are likely to be erroneously labelled as stacks if –*m* is too low. True alleles will not be recorded and will drop out if –*m* is too high. –*m* can decrease genotyping error by distinguishing real loci from PCR and sequencing error, but it can increase error by calling a heterozygous locus as homozygous when minimum coverage is set too high and one of the alleles is therefore excluded
number of mismatches allowed between loci when processing a single individual (-*M*) default 2	If -*M* is too low, some real loci will not be formed, and their alleles will be treated as different loci (undermerging). If -*M* is too large, repetitive sequences and paralogs will form large nonsensical loci (overmerging)
number of mismatches allowed between loci when building the catalog (-*n*) default 0	For *n* = 0, there would be loci represented independently across individuals that are actually alleles of the same locus. If *n* > 0, the consensus sequence from each locus is used to attempt to merge loci. This is important for population studies where monomorphic or fixed loci may exist in different individuals. Merging fixed alleles as a single locus can increase the probability of assembling real loci, and therefore decrease the allele error rate. However, erroneous loci will be created if –*n* is too high
maximum number of stacks at a single de novo locus (--*max_locus_stacks*) default 3	The expectation for non-repetitive genomic regions is that a monomorphic locus will produce a single stack because the two sequences on the two homologous chromosomes are identical and thus indistinguishable. In contrast, a polymorphic locus will produce two stacks representing alternative alleles. Confounding cases that may arise from short, sequencing error-based stacks or from repetitive sequences, where hundreds of loci in the genome may collapse to a single putative locus. --*max_locus_stacks* allows for the identification and blacklisting of confounding cases
SNP calling model	In the default SNP calling model the error parameter is allowed to vary freely, whilst in a bounded-error model the boundary value is substituted if the maximum-likelihood value of ε exceeds a lower or

	upper bound. One consequence is that reducing the upper bound increases the chance a homozygous loci being called heterozygous. The SNP calling model allows the tolerance for false positive vs. false negative rates in calling genotypes to be tuned, which in turn influences the genotyping error

* Parameters explanation as in Catchen *et al.* 2013 and *Stacks* documentation, effect on genotyping error as discussed here.

Table 3.4. Information content, error rates and efficacy to detect structuring of genetic variation for the full dataset processed with different *Stacks* parameter settings.

	optimal	***near optimal***	***high coverage***	***default***
Number of RAD-loci	6292	2449	292	4554
Total number of SNPs	11057	4353	502	7736
Mean read coverage per sample	10.32 (SD 4.16)	15.30 (SD 5.9)	58.92 (SD 21.9)	11.50 (SD 4.65)
Mean locus error rate	0.1738 (SD 0.103)	0.1657 (SD 0.100)	0.0882 (SD 0.088)	0.1590 (SD 0.094)
Mean allele error rate	0.0592 (SD 0.013)	0.0599 (SD 0.010)	0.0879 (SD 0.023)	0.0841 (SD 0.017)
Mean SNP error rate	0.0243 (SD 0.006)	0.0321 (SD 0.006)	0.0578 (SD 0.019)	0.0423 (SD 0.010)
Variation explained by first two axes of PCoA*	80(39)%	82(34)%	47(22)%	57(32)%
Mean of F_{ST} pairwise matrix*	0.19(0.07)	0.15(0.04)	0.03(0.01)	0.07(0.04)

* Results outside parenthesis were obtained using all the samples of the dataset, and the value inside parenthesis corresponds to the results if excluding the samples from El Zamorano and the outgroup. El Zamorano (*B. alpina* population from SMOr) was excluded because it explained as much variation as the *B. trifolia* outgroup (Supporting Information 5).

3.5. Acknowledgements

We thank Subject Editor Alex Buerkle, Brant Faircloth and three anonymous referees for their constructive comments on an earlier version of the manuscript; L. Figueroa, C. Berney, T. Wyss and A. Brelsford for lab work assistance; O. Trejo for assistance with sampling permits and and SMG, JRPPK, JJRL, AOM, ROF, SSF, RAF, TSA, JAA, FDRG, FQB y MJLF for fieldwork assistance. Part of the analyses were carried out on the High Performance Computing (HPC) Cluster supported by the Research and Specialist Computing Support service at UEA. This work has been supported by a CONACYT doctorate scholarship to AMY (213538), by a CONACYT grant to DP (178245), by a SSE Rosemary Grant Student Research Award to AMY and by an SNSF grant (PP00P3_144870) to N. Alvarez.

3.6. References

Baird NA, Etter PD, Atwood TS *et al.* (2008a) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, **3**, e3376.

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE*, **6**, e19315.

Bonin A, Bellemain E, Bronken Eidesen P *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Chong Z, Ruan J, Wu C-I (2012) Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. *Bioinformatics*, **28**, 2732–2737.

Crawford LA, Koscinski D, Keyghobadi N (2012) A call for more transparent reporting of error rates: the quality of AFLP data in ecological and evolutionary research. *Molecular Ecology*, **21**, 5911-5917.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12, 499–510.

Dou, J., X. Zhao, X. Fu, W. Jiao, N. Wang, L. Zhang, X. Hu, S. Wang, and Z. Bao. 2012. Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. *Biology Direct* 7:17.

Eaton, D. A. R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics,* btu121.

Gautier M, Foucaud J, Gharbi K *et al*. (2013a) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, **22**, 3766–3779.

Gautier M, Gharbi K, Cezard T *et al.* (2013b) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Gomes I, Collins A, Lonjou C *et al.* (1999) Hardy-Weinberg quality control. *Annals of Human Genetics*, **63**, 535–538.

Gompert Z, Buerkle CA (2011) A Hierarchical Bayesian Model for Next-Generation Population Genomics. *Genetics*, **187**, 903–917.

Grundemann D, Schomig E (1996) Protection of DNA during preparative agarose gel electrophoresis against damage induced by ultraviolet light. *BioTechniques*, **21**, 898–903.

Hohenlohe PA, Catchen J, Cresko WA (2012) Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. *Methods in Molecular Biology (Clifton, N.J.)*, **888**, 235–260.

Johnson PLF, Slatkin M (2008) Accounting for bias from sequencing error in population genetic estimates. *Molecular Biology and Evolution*, **25**, 199–206.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, **22**, 2986–3001.

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

Loman NJ, Misra RV, Dallman TJ et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

McKenna A, Hanna M, Banks E et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Meacham F, Boffelli D, Dhahbi J et al. (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**, 451.

Miller CR, Joyce P, Waits LP (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics*, **160**, 357–366.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Parchman TL, Gompert Z, Mudge J *et al.* (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847–859.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

Price DL, Casler MD (2012) Simple regression models as a threshold for selecting AFLP loci with reduced error rates. *BMC Bioinformatics*, **13**, 268.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**, 945–959.

R. Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richards PM, Liu MM, Lowe N *et al.* (2013) RAD-Seq derived markers flank the shell colour and banding loci of the *Cepaea nemoralis* supergene. *Molecular Ecology*, **22**, 3077–3089.

Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, 38, D234–236.

Rounsaville TJ, Ranney TG (2010) Ploidy Levels and Genome Sizes of Berberis L. and Mahonia Nutt. Species, Hybrids, and Cultivars. *HortScience*, **45**, 1029–1033.

Taberlet P, Griffin S, Goossens B *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

Urbanek S (2011) *multicore: Parallel processing of R code on machines with multiple cores or CPUs*.

Vázquez-Lobo A (1996) Filogenia de hongos endófitos del género *Pinus*: Implementación de técnicas moleculares y resultados preliminares. Sc. Bach. Dissertation. Facultad de Ciencias, Universidad Nacional Autónoma de México, México

Warnes GR, Bolker B, Lumley et al (2013) *gtools: Various R programming tools*. R package.

Wickham H (2012) *stringr: Make it easier to work with strings*. R package.

Willing E-M, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics*, **27**, 2187–2193.

Xu J, Turner A, Little J, Bleecker ER, Meyers DA (2002) Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? *Human genetics*, **111**, 573–574.

Zamudio S (2009) Notas sobre el Género *Berberis* (Berberidaceae) en México. *Acta Botánica Mexicana*, **87**, 31–70.

Zhang H, Hare MP (2012) Identifying and reducing AFLP genotyping error: an example of tradeoffs when comparing population structure in broadcast spawning versus brooding oysters. *Heredity*, **108**, 616–625.

3.7. Data and code availability

Raw RADseq data Sequence Read Archive (SRA) accession SRP035472. Sampling information, custom R & Perl scripts and jobs with settings used to run *Stacks*, output data to compare error rates and population differentiation: doi:10.5061/dryad.g52m3. Error-rate R functions updated and versioned: https://github.com/AliciaMstt/RAD-error-rates.

3.8. Supporting information

Due to the nature of some of the Supporting information materials, they are provided as an annex at the end of this thesis.

S1. Schematic diagram of RAD data genotyping and differences between replicates

S2. Summary of ddRAD lab work reaction mixes used and the characteristics of the resulting libraries

S3. Dendrograms obtained from the analyses of replicates analyses with different *Stacks* parameters

S4. Effect on a) the allele error rate and b) the SNP error rate of using a bounded SNP calling model with different values for the upper bounder (0.0056, 0.05,

0.10, 0.15, 0.25, 0.50) or using the default SNP calling model (free) for three values of -*m*: m=3 (left), m=4 (middle) and m=10 (right).

S5. PCoA for each of the four *Stacks* parameter profiles tested (optimal, near optimal, high coverage and default). Upper panels correspond to the PCoA performed with all samples, and the bottom to the analyses removing the El Zamorano population and *B. trifolia* from the distance matrix.

CHAPTER 4

Gene duplication, population genomics and species-level differentiation within a tropical mountain shrub

A version of this chapter has been published in *Genome Biology and Evolution*

DOI: 10.1093/gbe/evu205

Arboreal fish, ticatla 2014

The genome is new, and it is old. It is a big bag of genes travelling through space-time that is constantly retooling itself. It is a rather amazing messy collection of base pairs.

- Yingguang Frank Chan

4.1. Abstract

Gene duplication leads to paralogy, which complicates the *de novo* assembly of genotyping-by-sequencing (GBS) data. The issue of paralogous genes is exacerbated in plants, because they are particularly prone to gene duplication events. Paralogs are normally filtered from GBS data before undertaking population genomics or phylogenetic analyses. However, gene duplication plays an important role in the functional diversification of genes and it can also lead to the formation of postzygotic barriers. Using populations and closely related species of a tropical mountain shrub, we examine: (1) the genomic differentiation produced by putative orthologs, and (2) the distribution of recent gene duplication among lineages and geography. We find high differentiation among populations from isolated mountain peaks and species-level differentiation within what is morphologically described as a single species. The inferred distribution of paralogs among populations is congruent with taxonomy and shows that GBS could be used to examine recent gene duplication as a source of genomic differentiation of non-model species.

Keywords: RAD-seq, de novo assembly, GBS, paralogy, Transmexican Volcanic Belt, *Berberis*

4.2. Introduction

The development of genotyping-by-sequencing (GBS) methods (reviwed by Davey et al. 2011; Poland and Rife 2012) has accelerated the use of genomic data in population genetics studies of non-model organisms. This is particularly useful for plants, where population genetic studies have often struggled to obtain sufficient resolution from DNA sequence data with traditional Sanger sequencing approaches. For example, several plant phylogeographic studies (e.g. Tovar-Sánchez et al. 2008; Gugger et al. 2011; Mastretta-Yanes et al. 2011) have been substantially less informative than studies that have used comparable sequencing effort in animal taxa within the same geographic region (e.g. Bryson et al., 2011, 2012; McCormack et al., 2008; Ornelas et al., 2013). By applying GBS techniques sufficient nucleotide variation can be harnessed within plant species to address evolutionary questions, such as genetic association of adaptive traits (Parchman *et al.* 2012) and genomic divergence of hybridizing tree species (Stölting *et al.* 2013). However, applying GBS to plants poses a unique set of challenges, or exacerbates those common to other taxa (Morrell *et al.* 2012; Schatz *et al.* 2012; Deschamps *et al.* 2012). Plant genomes typically contain a large number of transposable elements (Feschotte *et al.* 2002), which causes GBS reads to map with equal probability to multiple positions within a reference genome. Polyploidy events have also occurred frequently throughout the evolutionary history of plant species, as well as other types of gene duplication that can result in large multi-gene families (Lockton & Gaut 2005; Flagel & Wendel 2009), and thus a considerable number of paralogous loci. Paralogous loci are typically treated as a nuisance variable and filtered from GBS data, however the emergence of paralogous loci is a consequential process that

contributes to genome evolution, and can thus be examined for the quantification of genomic differentiation among populations and species.

Paralogous loci arise by gene duplication, such that both copies evolve in parallel during the history of an organism (Fitch 1970, Fig. 4.1a). Gene duplication can occur at the whole genome level (polyploidy event), but can also be limited to chromosome segments or single genes (Hurles 2004). Gene duplication can confound the assembly of genomic data because paralogs can be erroneously merged together as a single locus (Fig. 4.1c), leading to difficulty in distinguishing allelic variation from differences among closely related gene family members (Hohenlohe *et al.* 2012; Dou *et al.* 2012). This issue is caused by relatively recent gene duplications (i.e. those origination within a genus or among closely related species), because more ancient duplication events occurring over much deeper time scales are expected to have accumulated enough differences to be assembled as different loci (Fig. 4.1d). The confounding effect of gene duplication on the assembly of genomic data is particularly problematic for *de novo* assembly, but even if a reference genome is available, the short sequence reads that are typical of high-throughput sequencing may not map uniquely within a reference genome (Hohenlohe *et al.* 2012; Morrell *et al.* 2012).

Treating paralogs as a single locus generates spurious heterozygous genotype calls and can confound the estimation of genetic differentiation among individuals and populations. The magnitude of this effect will depend upon the characteristics of the focal genome, and the relatedness of the samples being analysed. With regard to focal genome characteristics, plant and fish genomes contain more duplicated genes than mammals (Volff 2004; Lockton & Gaut 2005)

and will thus, on average, provide a greater challenge for genome assembly because of paralogous loci. The evolutionary relatedness among samples is also important because paralogs are continuously arising within each evolutionary lineage (Lynch & Conery 2000; Langham *et al.* 2004; Hurles 2004). Thus, the more a focal group departs from a model of panmixia, the more paralogous loci one would expect to retrieve across all samples. In the extreme, one may expect different species, or sufficiently differentiated populations, to exhibit species-specific or population-specific paralogs.

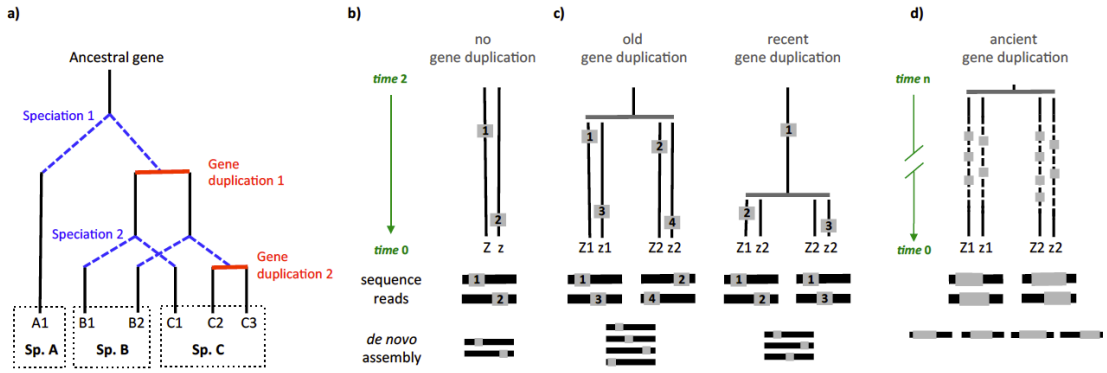


Figure 4.1. a) Paralogy and orthology relationships among six contemporary genes (A1-C3) in three species (A-C), adapted from Jensen (2001). Paralogous genes are produced by duplication events (red horizontal line) and orthologous by speciation (blue dashed inverted "Y"). A given gene in one species may have more than one ortholog in another species (e.g. B1 and B2 in species B are orthologs of A1 in species A) and paralogs are not necessarily restricted to the same species (e.g. B1 and C2 are paralogs). b) On a locus Z (with the alleles Z and z), mutation events (grey boxes) lead to the formation of two possible sequence reads (coverage not shown) that are correctly assembled as two alleles of the same locus. c) Loci that are the product of gene duplication (Z1 and Z2) produce reads that can not be distinguished from allelic variation and are assembled as a single locus with several alleles, generating erroneous SNP calls. Loci produced by relatively old duplication events would accumulate more nucleotide differences than recently duplicated loci. Therefore, if paralogs are merged as single locus, the products of old duplication events will generate more (spurious) alleles than paralogs from more recent duplication events. d) Loci produced by more ancient duplication events would accumulate enough differences to be assembled as different loci.

Paralogous loci are typically entirely filtered from GBS data. This can be done at the stage of assembly and genotyping, for instance by incorporating

differences in coverage (Dou *et al.* 2012), or by testing the independence of bi-allelic SNPs for each pair of tags (Poland et al. 2012, but se also Gayral et al. 2013; Eaton 2014 for other approaches). Filtering can also be performed on the assembled data, for example by retaining only those loci with the number of expected alleles and Hardy–Weinberg proportions (Hohenlohe *et al.* 2011; Catchen *et al.* 2013).

Despite gene duplication representing an analytical challenge for GBS, it is also a major source of evolutionary novelty (Lewis 1951; Ohno 1970). Therefore, by treating paralogs as a nuisance parameter and discarding them, potential signatures of evolution and adaptation are also being discarded. A duplicated gene copy may acquire a new function (Ohno 1970), specialize for a subset of the functions originally performed by the ancestral single-copy gene (Lynch and Force 2000) or contribute to protein dosage effects in response to environmental variables (Kondrashov *et al.* 2002). These processes are particularly relevant for plant evolution, as most plant diversity seems to have arisen following the duplication and adaptive specialization of pre-existing genes (Lockton & Gaut 2005; Moore & Purugganan 2005; Flagel & Wendel 2009). For example, many plant genes involved in pathogen recognition and herbivory defence arose through gene duplication (Moore & Purugganan 2005). However, there are also several examples of adaptive gene duplications in bacteria, yeast, fish, insect and mammal species (Kondrashov 2012). In addition to functional diversification, gene duplication can also promote speciation through the passive accumulation of genomic divergence (Lynch & Conery 2000). For example, following the duplication of an essential gene in *Arabidopsis thaliana*, populations varied with

respect to the copy that retained functionality, which acts as a postzygotic barrier among populations (Bikard *et al.* 2009).

Here, rather than seeking to remove paralogous loci, we use GBS data for the explicit purpose of investigating the distribution of putative recent gene duplication events among plant populations. We use double-digest restriction-site associated DNA sequencing (ddRAD) data sampled from the non-model plant species *Berberis alpina* (Berberidaceae) and its close relatives to characterize both (i) genomic relationships among individuals based on putative orthologs and (ii) the distribution of paralogous loci of recent origin among sampling localities and species. The inferred distribution of paralogous loci among sampling locations and species is congruent with genomic differentiation estimated from presumed orthologous loci, and reveals species-level differentiation within what is morphologically described as a single species. More broadly, our study shows that GBS can be used to study, without a reference genome, gene duplication as a source of population divergence and evolutionary novelty in non-model species.

4.3. Methods

4.3.1. Study system and sampling

Berberis alpina is a shrub that grows from 3,200-4,200 metres above sea level (masl) on alpine grasslands of the Transmexican Volcanic Belt (TMVB), a system of isolated high-altitude mountains in tropical Mexico (Fig. 4.2). The TMVB is a biodiversity hotspot (Myers *et al.* 2000) where temperate-to-cold adapted plant species are thought to have either survived through, or diversified *in situ* during,

the Pleistocene climate fluctuations (Toledo 1982; Graham 1999). *Berberis moranensis* grows at lower altitudes in the TMVB (1,800-3,150 masl, Zamudio 2009a) and is expected to be closely related to *B. alpina.*

Mountain peaks from 3,300 to 4,200 masl within the TMVB and nearby areas of the Altiplano Sur (AS) and of the Sierra Madre Oriental (SMOr) were surveyed for *B. alpina* (*sensu* Zamudio 2009b) during September-October 2010 and April-May 2011 (Fig. 4.2). The species was found in a total of seven locations, which represents its known distribution within the TMVB and the AS (Fig. 4.2). It was not found in the surveyed mountains of the SMOr. Samples of *B. moranensis,* a closely related species that grows up to 3,150 masl, were collected in Cerro San Andrés (Fig. 4.2), where *B. alpina* is absent. Samples of the outgroups *B. trifolia* and *B. pallida* were collected at lower elevations (~2,000-2,300 masl) of the TMVB (Fig. 4.2) in October 2012. Sampling was performed with SEMARNAT permission No. SGPA/DGGFS/712/2896/10. Herbarium specimens of *B. alpina* and *B. moranensis* were prepared and deposited within the Herbario Nacional in Mexico City (MEXU).

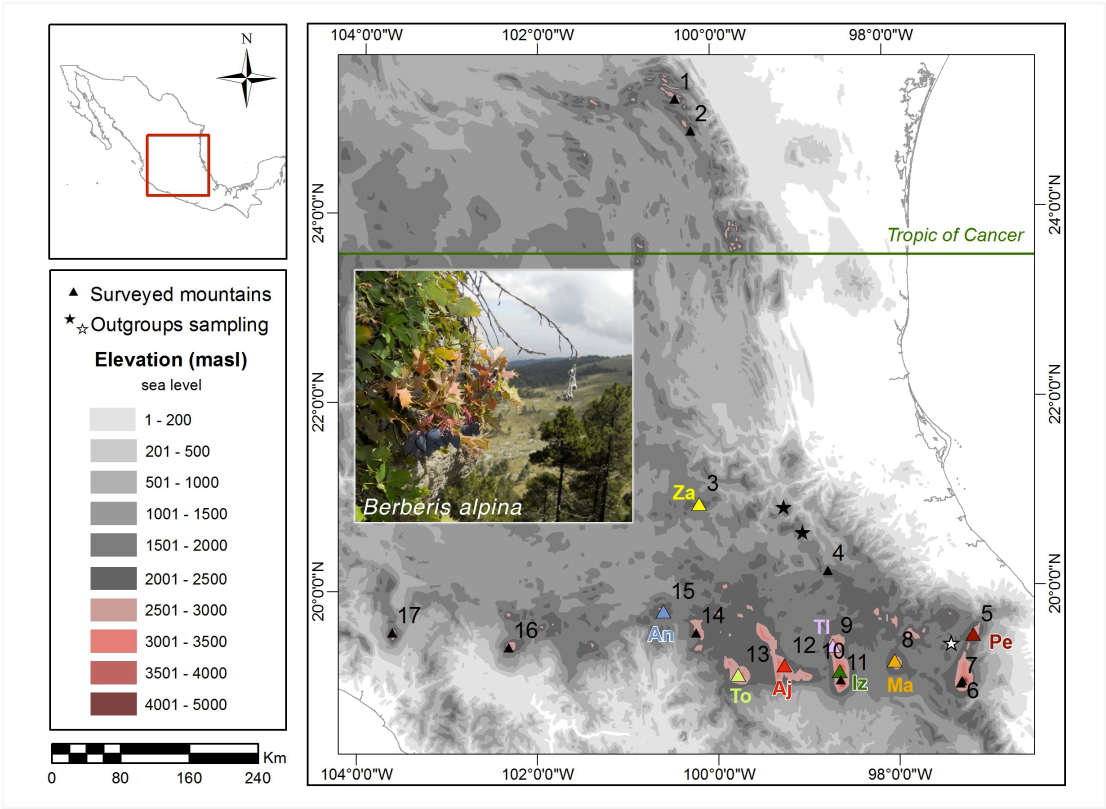


Figure 4.2. Surveyed mountains for *B. alpina* within the Sierra Madre Oriental (1-2), Altiplano Sur (3) and the Transmexican Volcanic Belt (4-17). Populations where *B. alpina* was found are Nevado de Toluca (To), Ajusco (Aj), Tlaloc (Tl), Iztaccihuatl (Iz), La Malinche (Ma) and Cofre de Perote (Pe) (To-Pe are referred as *B. alpina* ingroup) and Zamorano (Za). *B. moranensis* was collected from Cerro San Andrés (An, blue triangle). *Berberis pallida* (black stars) and *B. trifolia* (white star) were outgroups.

4.3.2. Molecular methods

Based on data from related species, the sampled *Berberis* species are likely diploid with a genome size of between 0.50 to 1.83 Gbp (Rounsaville and Ranney, 2010). We used ddRAD data from Mastretta-Yanes et al. (2014a), which consists of seventy-five individually tagged specimens of *B. alpina* and *B. moranensis* (6-10 per population), three samples of each outgroup (*B. trifolia* and *B. pallida*) and fifteen replicated samples, with at least one replicate per population or species. Briefly, the ddRAD libraries were prepared using the enzymes EcoRI-HF and MseI using a modified version of Parchman et al. (2012) and Peterson et al. (2012) protocols. Samples were divided into three groups,

each sequenced using single-end reads (100bp long) in a separate lane of an Illumina HiSeq2000.

4.3.3. De novo *assembly of RAD data*

After demultiplexing and quality trimming of raw reads, final sequences were 84 bp long. Data was *de novo* assembled using the software *Stacks* v. 1.02 (Catchen *et al.* 2011, 2013) with the parameter values *m*=3, *M*=2, *N*=4, *n*=3, *max_locus_stacks*=3, and a SNP calling model with an upper bound of 0.05. These settings (a) optimize the recovery of a large number of loci while reducing the SNP and RAD allele error rates, and (b) filter a fraction of putative paralogous loci merged as a single locus (Mastretta-Yanes et al. 2014b). After *de novo* assembly, the data were filtered to keep only those samples having more than 50% of the mean number of loci per sample, and only those loci present in at least 80% of the barcoded samples. Replicates were used to estimate error rates as in Mastretta-Yanes et al. (2014b) for each of the subsets of samples described in the sections below. For the population genomic analyses, only one sample per replicate pair was used.

Considerably fewer loci were recovered in *Berberis pallida,* which is likely explained by mutations affecting restriction enzyme cutting sites and hence a distant evolutionary relationship with the other *Berberis* species in the study. This species was therefore excluded from further analyses.

4.3.4. Identifying paralogs from recent gene duplications

Here we refer to a RAD-locus as a short DNA sequence produced by clustering together RAD-alleles; in turn, RAD-alleles differ from each other by a small

number of SNPs in certain nucleotide positions (SNP-loci). During *de novo* assembly two nucleotide mismatches (*M*=2) were allowed among reads to form a putative RAD-locus. Among individuals, loci were merged as a single locus if they presented up to three mismatches (*n*=3), which allows loci that are fixed differentially among different populations or species (thus represented independently across individuals) to be merged as a single locus (*Stacks* manual). During the formation of putative RAD-loci within individuals (determined by the -*M* parameter), and during the merging of monomorphic loci among individuals (determined by the -*n* parameter), it is expected that paralogous loci would be assembled as a single locus, leading to the formation of loci with three or more (spurious) alleles (Fig. 4.1c). Thus, if >2 alleles per locus are allowed during *de novo* assembly, data will likely contain merged paralogs. Here, a maximum of three alleles per locus was allowed (*max_locus_stacks* = 3) to filter out paralogs of relatively old origin. This filter retains paralogs derived from more recent gene duplications events, because loci produced by recent gene duplications are expected to have accumulated fewer mutations than older duplicated loci (Fig. 4.1c), and should thus produce fewer (spurious) alleles if merged as a single locus. Notice that 'old origin' is a relative term, implying that loci are still similar enough to resemble allelic variation. Paralogs from more ancient duplications, such as those ones shared across many genera and plant families (Lockton & Gaut 2005), are expected to have accumulated enough differences to be assembled as different loci (Fig. 4.1d).

PCR and sequencing error may also result in more than two alleles per locus within an individual (Hohenlohe *et al.* 2012; Catchen *et al.* 2013). However, the distribution of error-based alleles is stochastic, whilst merged paralogous

loci should produce population-wide shared polymorphism. Thus, merged paralogs can be identified by their signature on the site frequency spectrum (SFS, Hohenlohe et al. 2012): paralogous loci accumulate mutations independently, so assembling them as different alleles of the same locus produces spurious polymorphic positions at which all individuals would be heterozygous, with the exception of those that may have suffered allele dropout. This should bias the SFS towards heterozygosity with an excess of loci where the frequency of the major allele (p) is p=0.5. Here we consider any RAD-locus where p=0.5 in at least one SNP-locus within a given population to be a potentially paralogous locus. Such loci were further examined among other populations and species, because some orthologous loci may by chance be at p=0.5 in a given population, but it would be unlikely to observe this in two or more populations or within a related species. If a RAD-locus was identified as a potential paralog in two or more populations or species, it was considered to be *shared* among those taxa. However, if p=0.5 in only one population or species, the RAD-locus was considered to be a *private* potential paralog (i.e. the locus was present in other populations, but with $p \neq 0.5$).

The dataset was divided into the following three subsets of RAD-loci: (1) *All loci,* (2) *Putative orthologs* - excluding all potential paralogs, (3) *Putative orthologs within B. alpina* - excluding potential paralogs shared between two or more sampling locations of *B. alpina*, or between two or more species, which generates a subset of loci that should be orthologous within *B. alpina.* The frequency of the major allele within each locus was estimated for each of the three datasets. Allele frequencies were estimated at each SNP locus for each population and species by running the *populations* program of *Stacks* version

1.17 with the *de novo* assembled RAD-loci. The distribution of potential paralogous loci was examined and plotted with R version 2.15 (R. Core Team 2012).

4.3.5. Structuring of genetic variation and population genomic analyses

Preliminary analyses revealed the Cerro Zamorano population to be highly differentiated from other *B. alpina* populations (see discussion), so it was treated as a different lineage from *B. alpina*. Hereafter we use '*B. alpina* ingroup' to refer to the subset of *B. alpina* samples that excludes the Cerro Zamorano population.

Principal Coordinate Analysis (PCoA) was performed for all loci and for the putative orthologs. For each of these two datasets the PCoA was first performed with all samples, and then excluding the outgroup and the Cerro Zamorano population. Pairwise F_{ST} between populations were estimated using both subsets of loci. The percentage of polymorphic loci, heterozygosity, π, and F_{IS} at each nucleotide position were estimated for *B. alpina* ingroup using all loci and the subset of putative orthologs within *B. alpina*. All population genetic estimates were calculated using the *populations* program of *Stacks.*

4.3.6. Distribution of potential paralogous loci among populations and species

The distribution of shared and private potential paralogs among populations and species was further examined by controlling for unequal sample sizes, by randomly sampling four individuals (the smallest sample size) per locality. Total, shared and private potential paralogous loci were identified as described above, with the exception that shared loci were defined as those shared with *B. alpina* ingroup populations.

A linear regression was used to test if the proportion of private potential paralogous loci increases with population differentiation, the latter calculated as the mean F_{ST} per population using the putative orthologous loci subset (pairwise matrix from Table 4.1, without the outgroup). The analysis was performed using R with and without the Cerro Zamorano population.

Table 4.1. Pairwise F_{ST} for the putative orthologs subset (filtering out all putative paralogous loci)

	Iz	***Ma***	***Pe***	***Tl***	***To***	**Za**	**An**	**Out**
Aj	0.0383	0.0663	0.0972	0.0248	0.0534	0.5387	0.0757	0.4649
Iz		0.0648	0.1042	0.0299	0.0643	0.5623	0.0973	0.4909
Ma			0.0954	0.0582	0.0903	0.5634	0.1377	0.4932
Pe				0.0848	0.1216	0.4991	0.1609	0.4050
Tl					0.0534	0.5861	0.0984	0.5074
To						0.6116	0.1276	0.5339
Za							0.7225	0.6976
An								0.6393

Berberis alpina ingroup populations are shown in italics in the first five columns. *B. moranensis* (An) and *B. trifolia* (Out) are in the last columns and are shown as a reference for the values found among different species. El Zamorano (Za) population shows F_{ST} values higher than those found for *B. moranensis* (An) and *B. trifolia* (Out).

4.3.7. Morphological evaluation

We examined characters that have been informative for Mexican *Berberis* taxonomy (Zamudio 2009a; b) to assess morphological and ecological differentiation among populations of *B. alpina.* Variables included: leaf morphology (rachis length, number of leaflets, leaflet texture, shape of blades, number and length of teeth), growth habit and habitat preferences (vegetation type, substratum and altitudinal distribution). Morphological characters were examined in specimens from extant herbarium material (MEXU, ENCB, IEB, XAL)

Table 4.2. Morphological differences of *B. alpina* populations and *B. moranensis*

	B. alpina			***B. moranensis***
Character	**TMVB***	**Pe***	**Za and SMOr****	**TMVB*****
Growth habit	Low shrub 50-100 cm, or more	Low shrub 25-100 cm or more	Low shrub 10-60 cm	Shrub to tree 1-7(10) m
No. of leaflets	3-5(7)	3-5	3	(5)7-11(15)
Rachis length (terminal segment)	0.5-2(3) cm	1-2 cm	absent	(0.3)0.5-1.5(2) cm
Leaflets texture	Coriaceous	Coriaceous	Coriaceous and very rigid	Slightly coriaceous
Leaflets blades	Ovate to ample ovate	Ovate, oblong to elliptic	Oblong to elliptic	Lanceolate to ovate-lanceolate
No. of teeth by side	(3)4-7(9)	(3)4-7(12)	2-4(6)	(4)5-11(15)
Teeth length	(1)2-5 mm	2-5 mm	5-10 mm	1-2(5) mm
Substratum	Igneous rocks	Igneous rocks	Igneous or calcareous rocks	Igneous rocks
Vegetation type	Alpine grassland and upper limit of *Pinus hartwegii* and *Abies religiosa* forests	Alpine grassland and upper limit of *Pinus hartwegii* and *Abies religiosa* forests	*Abies religiosa*, *Pinus* spp. and *Quercus* spp. forests, never above timber line	*Abies religiosa*, *Pinus* spp. and *Quercus* spp. forests and secondary vegetation after perturbation, never above timberline
Altitudinal distribution (masl)	3,200-4,200	3,300-4,180	2,800-3,250	(1,800)2,000-2,800(3,150)

* TMVB refers to sampled populations for *B. alpina* in the TMVB ('*B. alpina* ingroup') as in Fig. 2, with the exception of Cofre de Perote (Pe) population. ** Cerro Zamorano (Fig. 2) and SMOr populations: Sierra del Doctor (20°47'25" N, 99°33'53" W at 3,250 masl) and Cerro Pingüical (21°09'35" N, 99°42'02.4" W at 3,060 masl). *** Several localities within the TMVB at 1,800-3,150 masl.

including two *B. alpina* populations from SMOr (Sierra del Doctor and Cerro Pingüical, Table 4.2) that was not possible to sample for the molecular analysis. Habitat characteristics and altitudinal distribution were recorded from field observations. Specimens of *B. moranensis* from throughout its distribution were also examined for comparison.

4.4. Results

4.4.1. RAD-seq data yield and error rates

The number of samples recovered (excluding one sample per replicate pair) after *de novo* assembly and quality filtering were two for *B. trifolia,* nine for *B. moranensis,* four for the Cerro Zamorano population and 6-10 for each *B. alpina* population (Table 4.1S). A total of 6,292 RAD-loci (84 bp long) and 6,105 SNP-loci were recovered after the *de novo* assembly and quality control steps. For the subset of putative orthologs (filtering all potential paralogs), a total of 4,030 RAD-loci and 3,843 SNP-loci were recovered. A total of 5,461 RAD-loci and 5,274 SNP-loci were recovered for the subset of putative orthologs within *B. alpina.* RAD-allele and SNP error rates, percentage of missing data and mean coverage per locus per sample are reported in Table 4.3. Broadly, for each dataset the allele error rate ranges from 3.5 to 5.9% and the SNP error rate from 1.3 to 2.2%, with ~20% of missing data and a mean coverage of ~10.5 (Table 4.3). Decreases from 5.9 to 4.1% for the RAD-allele error rate and from 2.2 to 1.5% for the SNP error rate represents significant differences ($p<0.001$ and $p<0.01$, respectively) between the dataset of all loci and the dataset excluding the 831 putative paralogs within *B. alpina* ingroup (see below for how these loci were defined). To

confirm that this was not a chance effect, we randomly filtered 831 loci from the entire dataset of samples and re-estimated error rates. We repeated this 100 times, and across all repetitions the RAD-allele and SNP-locus error rates were 5.9-6.0% and 2.1-2.2%, respectively, which are not significantly different from the error rates found when no loci are removed ($p>0.7$ for all repetitions for both types of error rate).

Table 4.3. RAD-seq data yield and error rate for each subset of loci

Data subset	RAD-loci	SNP-loci	RAD-locus erro rate*	RAD-allele error rate*	SNP error rate*	Missing data	Mean coverage*
All loci	6,292	6,105	17.4% (10.3)	5.9% (1.3)	2.2% (0.06)	20%	10.3 (4.2)
Putative orthologs	4,030	3,843	17.5% (10.4)	3.5% (1.1)	1.3% (0.04)	17%	11 (4.3)
Putative orthologs within *B. alpina*	5,461	5,274	17.28% (10.3)	4.1% (1.2)	1.5% (0.04)	17%	10.5 (4.3)

* SD shown between parenthesis

4.4.2. Identification and distribution of paralogous loci among populations and species

A total of 2,262 RAD-loci were identified as potential paralogous loci. When examining the subset of all loci, the frequency of the major allele for each SNP-locus reveals that the majority of loci that are polymorphic across populations are fixed within each population (Fig. 4.3a). The percentage of loci in the other

categories decreases sharply and monotonically, but then increases abruptly within the category containing loci where p=0.5. For *B. alpina* ingroup populations, the observed heterozygosity of 91% of these loci is H_{obs}=1 and the F_{IS} value of 98% of the loci is negative, with F_{IS}≤-0.5 in 77% of the cases. Out of the 2,262 potential paralogous loci, 831 have at least one SNP with p=0.5 in two or more populations or species, and were considered putative paralogs within *B. alpina* ingroup. Around 99% of these SNP-loci show negative F_{IS} values for *B. alpina* ingroup populations, with F_{IS}≤-0.5 in 69% of them and H_{obs}=1 in 57%. Retaining only the presumable orthologs within *B. alpina* ingroup does not remove the overrepresentation of SNP-loci with both alleles at equal frequency within *B. moranensis* and the Cerro Zamorano population (Fig. 4.3b), but it effectively removes the excess of loci where p=0.5 within all *B. alpina* ingroup populations (Fig. 4.3b).

The potential paralogs are not evenly distributed among sampling locations and species. In increasing order, the Cerro Zamorano population and *B. moranensis* exhibit proportionally more RAD-loci with at least one SNP where p=0.5 (Fig. 4.4 and 2S), the majority of which are private (Fig. 4.4 and 2S). In contrast, within a given population of the *B. alpina* ingroup fewer loci were found to be at p=0.5 (Fig. 4.4 and 2S). The number of private potential paralogs per population increases with their differentiation estimated from orthologous loci (Fig. 4.6), both when the Cerro Zamorano population is included (r^2=0.955, $F_{1,6}$=128.3, p<0.001) and when it is excluded (r^2=0.771, $F_{1,5}$=16.85, p<0.01). The distribution of total, private and shared potential paralogous loci is similar under unequal sample sizes (n=2-10; Fig. 4.2S) and equal sample sizes (n=4; Fig. 4.4).

4.4.3. Structuring of genetic variation

The PCoA from the subset of putative orthologous loci reveals that the Cerro Zamorano population explains as much of the variance as the outgroup, *B. trifolia*, whilst *B. moranensis* clusters closer to the remaining *B. alpina* populations (Fig. 4.5a). Excluding the Cerro Zamorano population and *B. trifolia* (Fig. 4.5b), results in separate clusters for *B. moranensis,* and for both the Cofre de Perote and Malinche populations of *B. alpina*, while Western populations (Aj, Tl and Iz; Fig. 4.2) form a single cluster.

For *B. alpina* ingroup populations, the pairwise F_{ST} matrix estimated with the putative orthologs ranges from 0.025 to 0.122 (mean = 0.070), with Cofre de Perote exhibiting the highest differentiation in all pairwise estimates (0.084-0.122, Table 4.1). Pairwise F_{ST} values of the Cerro Zamorano population against *B. alpina* ingroup populations are larger (0.499-0.612) than values obtained by comparing any *B. alpina* ingroup population against the outgroup (0.405-0.534) or against *B. moranensis* (0.076-0.161).

4.4.4. Genetic diversity within B. alpina *ingroup*

When considering all nucleotide positions (i.e. including those not polymorphic) of the presumably orthologous loci within *B. alpina* ingroup, the percentage of polymorphic loci (notice that locus here refers to a nucleotide position within the RAD-loci) ranged from 0.304 to 0.482%; the average frequency of the major allele from 0.9990 to 0.9994; H_{obs} from 0.0011 to 0.0014; and π from 0.0010 to 0.0016 (Table 4.1S). Cofre de Perote presented the highest genetic diversity (0.482% polymorphic loci, H_{obs}=0.0014 and π = 0.0016); Nevado de Toluca presented the lowest levels of genetic diversity (0.304% polymorphic loci,

H_{obs}=0.0011 and π = 0.0010), with the remainder of the populations exhibiting intermediate levels. Cofre de Perote has substantially more private alleles (1,064) than both the remaining populations (293-485, Table 4.1S) and *B. moranensis* (194, Table 4.1S). When the same statistics are estimated including all potential paralogs (Table 4.2S), the estimates of genetic diversity increase (e.g. H_{obs} increased from ≤0.0015 to ≥0.0026) and all F_{IS} values are negative.

4.4.5. Morphological variation

Specimens from Cerro Zamorano, Sierra del Doctor and Cerro Pingüical populations of *B. alpina* (Za-SMOr populations) are low rhizomatous shrubs (20-60 cm) that tend to have only three leaflets per leaf and a sessile terminal leaflet (not inserted on a conspicuous rachis' segment) (Table 4.2). In contrast, populations from the TMVB are dense, appressed shrubs (20 cm to 1 m or more), flattened against rocks or cliffs and tend to have 3-5 (max. 7) leaflets with the terminal leaflet always inserted on a conspicuous segment of the rachis (Table 4.2). A ubiquitous rachis and >5 leaflets are characteristic traits of *B. moranensis* (Table 4.2). *Berberis alpina* populations of the TMVB inhabit the highest elevations, mostly on igneous rocks of alpine grasslands, while *B. alpina* populations of the Za-SMOr do not grow beyond the timberline and can grow on calcareous rocks. TMVB populations can co-occur with *B. moranensis* in the upper limit of conifer forests (Table 4.1S).

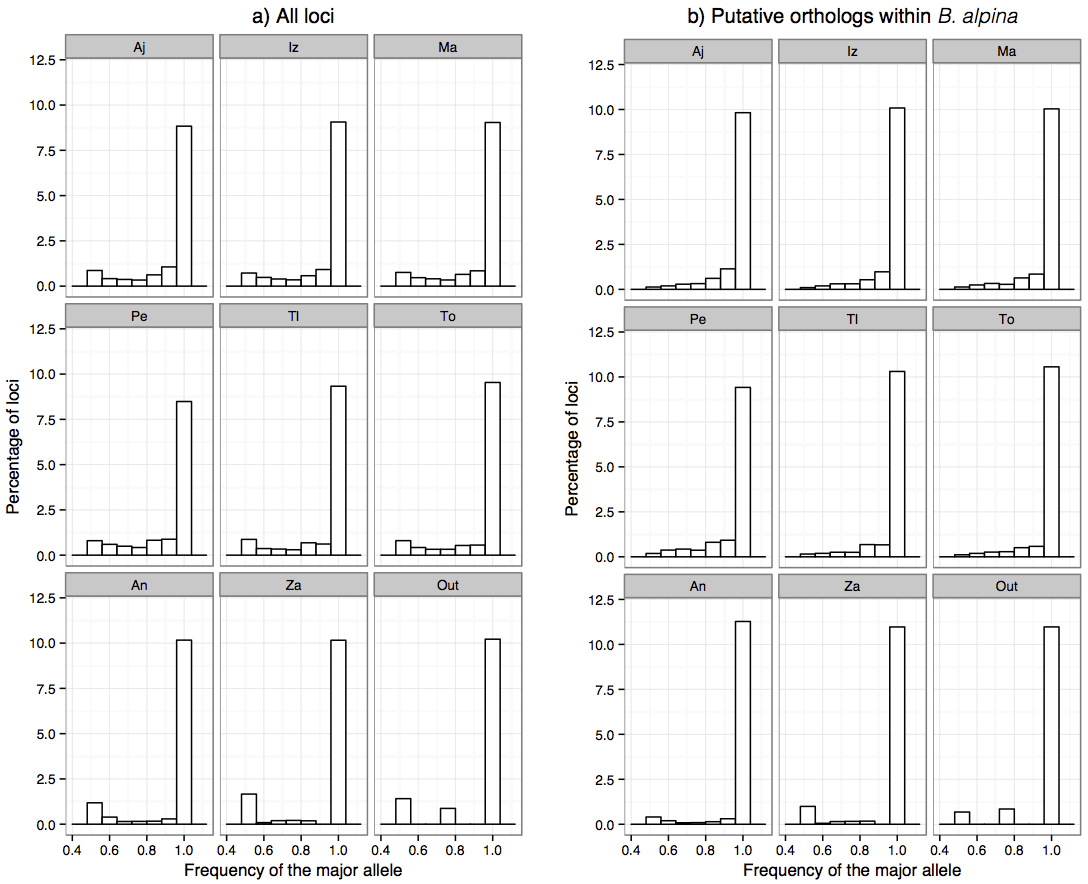


Figure 4.3. Distribution of the frequency of the major allele (*p*) for the SNP-loci for each *Berberis spp.* population. The plots on (a) correspond to all loci after *de novo* assembly and quality filtering. Notice that for every population a substantial percentage of loci is in the 0.5 category (left most bar). The plots on (b) show the distribution of the frequency of the major allele for the subset of loci presumably orthologous for *B. alpina* ingroup. This filtering removes the bias towards heterozygosity in *B. alpina* ingroup (top six panels), but not from *B. moranensis* (An) and the Zamorano population (Za). Notice that for Za and the outgroup (Out), small sampling sizes (4 and 2, respectively) affect the range of allele frequencies that can be recovered.

4.5. Discussion

4.5.1. Paralogs identification

A total of 2,262 RAD-loci were identified as potential paralogs, out of which 831 RAD-loci presented SNPs with p=0.5 in more than one population or species and were identified as putative paralogs within *B. alpina*. Removing these loci produced a set of presumably orthologous RAD-loci for the *B. alpina* ingroup.

This is similar to the approach taken by Hohenlohe et al. (2011) and Pujolar et al. (2014) to produce a dataset of putative orthologs for population genetics analyses of fish species, by removing loci with high values of observed heterozygosity. Here, we explored the excess of heterozygosity by examining if the loci where p=0.5 had high levels of H_{obs} and negative F_{IS}, as would be expected if these loci were the result of overmerging paralogous loci as a single locus. Then we examined the effect of filtering the putative paralogs on the SFS and the estimation of population genetics statistics.

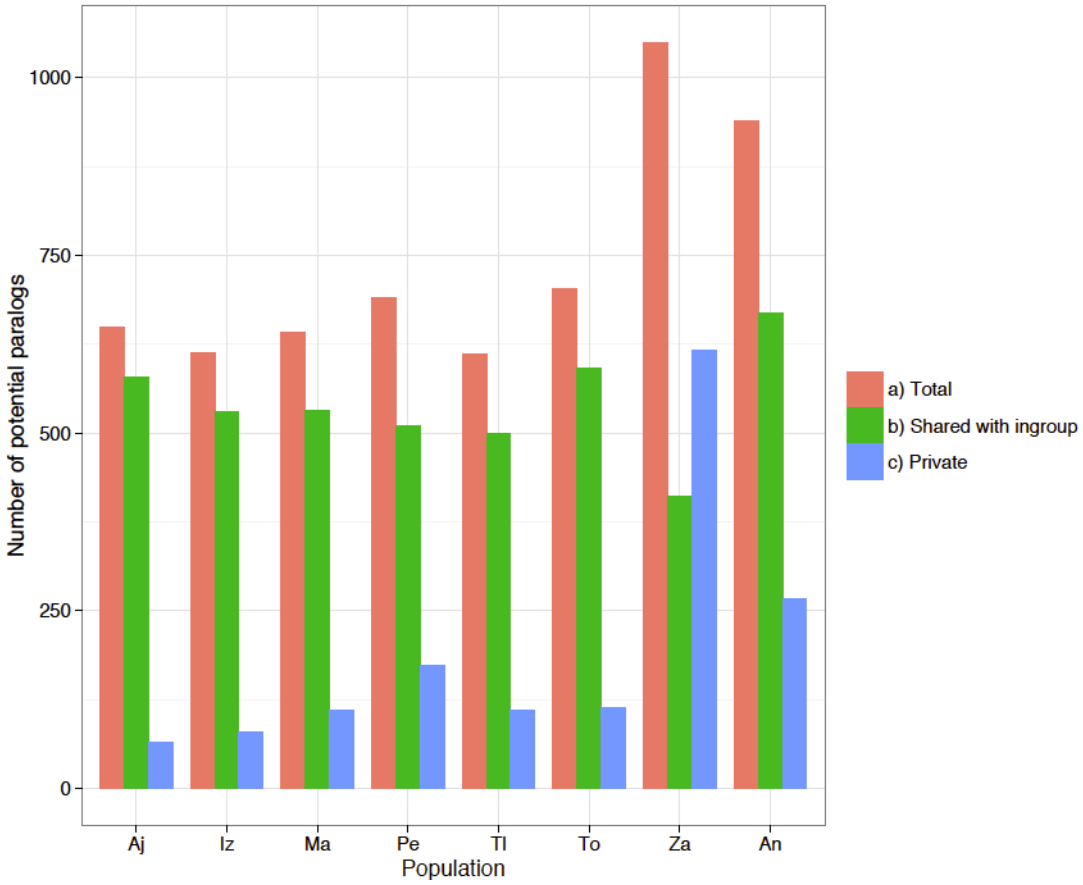


Figure 4.4. Distribution of RAD-loci with at least one SNP-locus where the frequency of the major allele equals 0.5 (potential paralogs). a) There are more loci biased towards p=0.5 in *Berberis moranensis* (An), the Zamorano population (Za) and *B. trifolia* (Out) than in *B. alpina* ingroup populations (Aj-To). b) Most of the loci where p=0.5 are the same loci in *B. alpina* ingroup and any given population or species, but c) a substantial proportion of loci show p=0.5 exclusively in *B. moranensis* or the Zamorano population. Sampling size (n=4) is the same for every population.

Filtering out the putative paralogs for *B. alpina* ingroup removed the bias towards loci with p=0.5 within these populations, but it remained noticeable for *B. moranensis* and the Cerro Zamorano population (Fig. 4.3b). This is explained by a high number of private potential paralogs within both *Berberis moranensis* and the Cerro Zamorano population (267 and 617, respectively; Fig. 4.4c). Under Hardy-Weinberg equilibrium (HWE), loci where most individuals are heterozygous are expected to be at the lowest frequency of the spectrum. While it remains possible that some of the private potential paralogous loci detected here are actually true loci where p=0.5, for *B. moranensis* and the Cerro Zamorano population they account for 18% and 37% of the non-fixed SNP-loci. Balancing selection could cause a bias towards heterozygosity but this should affect very few loci in the genome and it can not explain all (or most, as some may not be due to allele drop out) individuals being heterozygous (as shown by H_{obs}=1 in 91% and negative F_{IS} in 98% of the loci where p=0.5). Biological explanations for such extreme heterozygosity within populations are lacking, and co-occurring PCR/sequencing error cannot have produced the bias, because samples were individually tagged and randomly sequenced in different lanes. The most parsimonious explanation is therefore that the inferred heterozygosity is an artefact of the assembly of independent loci as a single locus. Therefore, the p=0.5 criterion used here for identifying potential paralogs among populations and species could be fine-tuned by formal tests of HWE deviations in datasets with sufficient sampling sizes per species. Finally, all things being equal, if the private potential paralogs were truly heterozygous loci their frequency within each population should be proportionally the same among populations. Interestingly, we found that the number of private potential paralogs increases

with the differentiation estimated using only orthologs (Fig. 4.6). This can be explained if the private potential paralogs were indeed the product of gene duplication, which is expected to occur independently within lineages and isolated populations.

Filtering deviations from HWE, such as bias towards heterozygosity caused by merged paralogs, is a necessary step for producing a set of putative orthologs, as evidenced by the following three observations. First, analyses including the putative paralogs yielded negative F_{IS} values for all populations of the *B. alpina* ingroup, and produced levels of polymorphic loci, H_{obs} and π that were found to be erroneously higher (Table 4.1S) than those obtained when these loci were excluded (Table 4.1). Second, filtering out putative paralogs increased population differentiation estimates: after putative paralogous loci within *B. alpina* are filtered out, the first axis of the PCoA of all samples increases from 81% (Fig. 4.1S) to 86% of the variance explained (Fig. 4.5), and the mean of the F_{ST} pairwise values among the *B. alpina* ingroup populations increases from 0.060 (Table 4.3S) to 0.077 (Table 4.4S). This is to be expected from the erroneous assembly of paralogous loci as a single locus, as merged paralogs generate shared polymorphism among populations. Third, the removal of paralogous loci decreased both the RAD-allele and SNP error rates (from 5.9% to 4.1%, and from 2.2% to 1.5%, respectively), likely because paralogous loci have more "alleles", and are thus more prone to allele drop out, an important source of error for low coverage GBS data (Mastretta-Yanes et al. 2014b).

4.5.2. Origin of paralogous loci in Berberis *taxa and populations.*

The older a gene duplication event is, the more nucleotide differences paralogous loci should accumulate, leading to an increased probability of recovering more than three "alleles" if they are merged as a single locus (Fig. 4.1c). Eventually the paralogs will accumulate enough differences to be assembled as different loci (Fig. 4.1d). Thus, allowing a maximum of three alleles per locus, as done here, should retain paralogs of relatively recent origin. Because gene conversion causes paralogs to maintain sequence similarity (Lynch & Conery 2000), a fraction of the putative paralogs could be older. However, gene conversion occurs mostly within multi-gene families (Semple & Wolfe 1999), which in plants tend to have an ancient origin and be largely conserved among families (Flagel & Wendel 2009).

Regarding the duplication mechanism, ancient polyploidy events within *Berberis* cannot be fully discarded. However, given that (a) the potentially paralogous loci identified here are expected to have a recent origin, (b) that they represent only a fraction of the recovered RAD-loci (from 13% of the RAD-loci, for *B. alpina* ingroup to 17% for the Cerro Zamorano population) and (c) that they are not homogeneously distributed among populations and species, it is likely that they arose by gene duplication mechanisms other than whole genome duplication. These alternative duplication mechanisms (reviewed for plants by Freeling 2009) include segmental duplication events, transposable elements and small-scale duplications (Lockton & Gaut 2005; Moore & Purugganan 2005; Flagel & Wendel 2009), and have been found to be responsible for the origin of recent paralogs within *A. thaliana* (Moore & Purugganan 2003).

4.5.3. Population differentiation and a cryptic Berberis *species*

Berberis alpina sampled from Cerro Zamorano was found to be strongly genetically differentiated from all other *B. alpina* populations, forming a distinct cluster in the PCoA that explained as much of the variation as the outgroup (Fig. 4.5). Additionally, F_{ST} values between Cerro Zamorano and the other *B. alpina* sampling locations are higher than those between the outgroup and the other *B. alpina* sampling locations (Table 4.1, Table 4.4S). The Cerro Zamorano population also exhibits a high number of RAD-loci that are likely to comprise private paralogous loci (Fig. 4.4c). Za-SMOr populations of *B. alpina* present habitat and leaf morphology differences from both TMVB populations of *B. alpina* and from *B. moranensis* (Table 4.1S). Such morphological characters are not necessarily indicative of species level differentiation, but considered together with the genomic differentiation it would appear that Za-SMOr should be recognised as a different species from the *B. alpina* TMVB populations. Species level differentiation of the *Berberis* sp. from the Cerro Zamorano from *B. alpina* from the TMVB is also congruent with (i) analyses showing that the SMOr, the AS and the TMVB are different biogeographic units (Arriaga *et al.* 1997; Morrone *et al.* 2002), with the fact that Cerro Zamorano is an old (~11 Myr old, Carrasco-Ñuñez et al. 1989) and isolated mountain (Fig. 4.2), and (iii) with data on vascular plants distributions showing that the Cerro Zamorano contains a high number endemic species or species restricted to it and to neighbour mountains in the SMOr (Rzedowski *et al.* 2012).

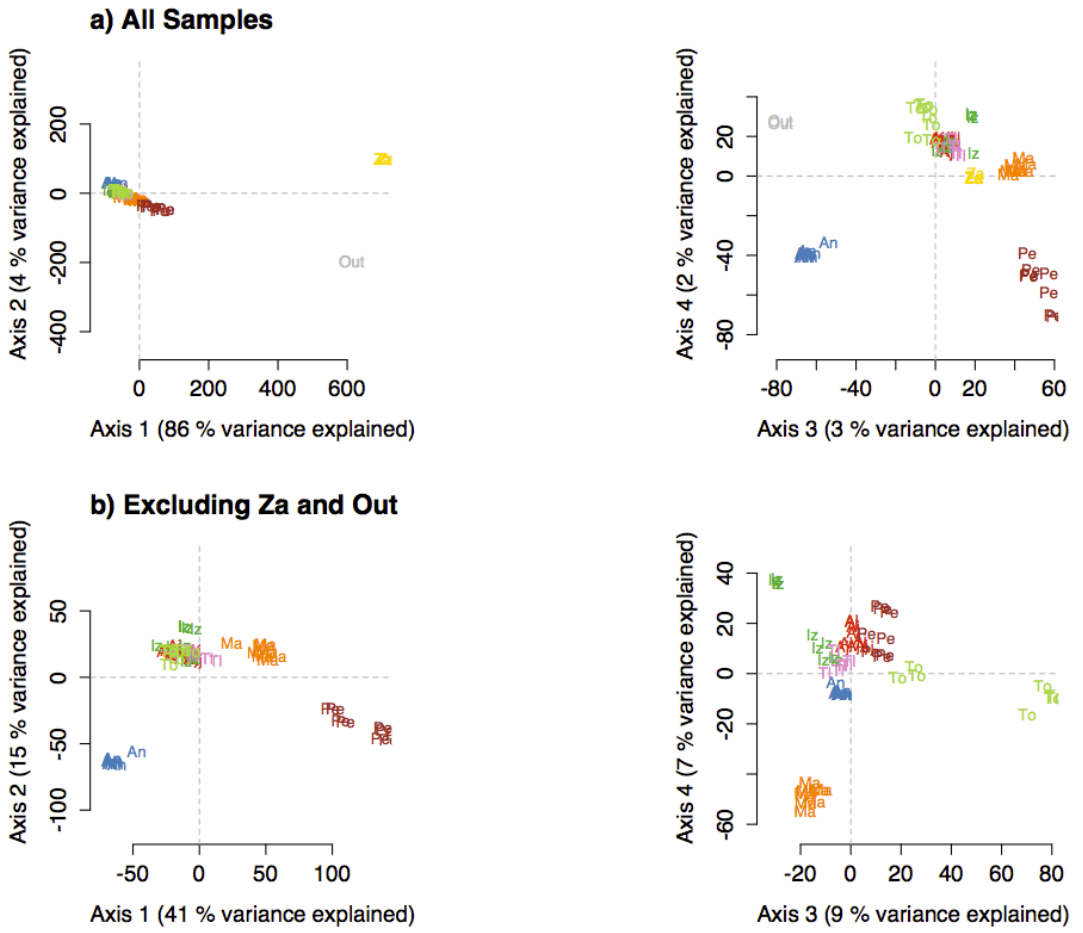


Figure 4.5. Principal coordinates analysis of the SNP-loci excluding all potential paralogs. (a) When all samples are analyzed axis 1 explains 86% of the variation and corresponds to the differences between El Zamorano-*B. trifolia* (Za and Out, respectively) to the rest of the populations. (b) If El Zamorano and *B. trifolia* are excluded, axis 1 and 2 separate *B. moranensis* (An) and the Cofre de Perote and Malinche (Pe and Ma) populations of *B. alpina*, explaining 41% and 15% of the variance, respectively. Populations ID and colors as in Fig. 4.2.

Regarding *B. alpina* ingroup populations, samples from topographically isolated mountains are expected to be genetically more differentiated than populations separated by less shallow elevations. During the Pleistocene climate fluctuations, the spatial distribution of climate variation did not undergo substantial latitudinal changes in Central Mexico, but it did undergo altitudinal shifts (Metcalfe 2006). During glacial periods cold temperatures existed at lower altitudes than today, allowing alpine grasslands to occur down to 2,500 masl, ~1,000 m below their current interglacial range (Lozano-García et al. 2005;

Metcalfe 2006; Vázquez-Selem and Heine 2011). By performing altitudinal migrations involving only short horizontal distances, species from alpine grasslands of the TMVB are expected to have persisted relatively *in situ*, with altitude being the main variable influencing possible habitat connectivity, and thus gene flow, among mountains in the past (Toledo 1982; Graham 1999). The subset of putative orthologous RAD-loci within *B. alpina* ingroup supports this expectation because the populations that are topographically more isolated (Cofre de Perote and Malinche, Fig. 4.2) present the highest F_{ST} values (0.085-0.122 and 0.068-0.100, respectively, Table 4.4S) and have the highest number of private alleles (1,067 and 485, respectively, Table 4.1S). Genomic differentiation was significant among all populations, with F_{ST} values typically greater than 0.05, (Table 4.4S), with all populations exhibiting low frequency alleles (Fig. 4.3b), as expected for old and stable populations. These genetic patterns support the hypothesis that *B. alpina* populations were able to survive *in situ* through several Pleistocene climate fluctuations. Similar conclusions have been reached for animal taxa of the TMVB using more traditional population genetic and phylogeographic approaches (e.g. McCormack et al. 2008; Bryson et al. 2011, 2012).

Berberis moranensis grows at lower elevations than *B. alpina* from the TMVB (Table 4.1S). Interestingly, the Cofre de Perote population of *B. alpina* and *B. moranensis* exhibit similar F_{ST} values against *B. alpina* ingroup populations (0.085-0.122 and 0.076-0.138, respectively; Table 4.1). However, the differentiation of Cofre de Perote is driven by a high number of private alleles (1067, Table 4.1), while *B. moranensis* has fewer private alleles (194, Table 4.1) but presents 267 RAD-loci that are presumed to be private paralogs,

approximately twice the number than in Cofre de Perote (174, Fig. 4.4c). Morphologically there are similar leaf characters (e.g. rachis and number of leaflets; Table 4.1S) between *B. alpina* ingroup populations and *B. moranensis*. This phenomenon could be explained by different scenarios of hybridization, ancestry or selection favouring duplicated loci. However, it is not possible to assess these kinds of hypotheses with our current geographical sampling of *B. moranensis*.

4.5.4. Paralogous loci as a source of genomic differentiation

A central finding of this study is that there are quantitative differences in the distribution of potential paralogous loci among populations and species: *B. moranensis* and the population likely representing a different species (Cerro Zamorano) have a high number of private paralogs (Fig. 4.4), and the populations in the *B. alpina* ingroup that are more differentiated for presumed orthologous loci also present a larger number of presumed private paralogs (Fig. 4.6).

Examining the distribution of paralogous loci among populations and species is relevant because (i) gene duplication might lead to functionally relevant, ecologically significant polymorphisms (Moore & Purugganan 2005); and (ii) the divergent evolution of recently duplicated genes can lead to postzygotic isolating barriers within existing species (Bikard *et al.* 2009). Testing whether the former phenomena were consequential for genome divergence among our *Berberis* species would require analysing the identified paralogous loci with a more detailed understanding of their genomic context and potential function. However, the paralogous loci found here are already an extra source of

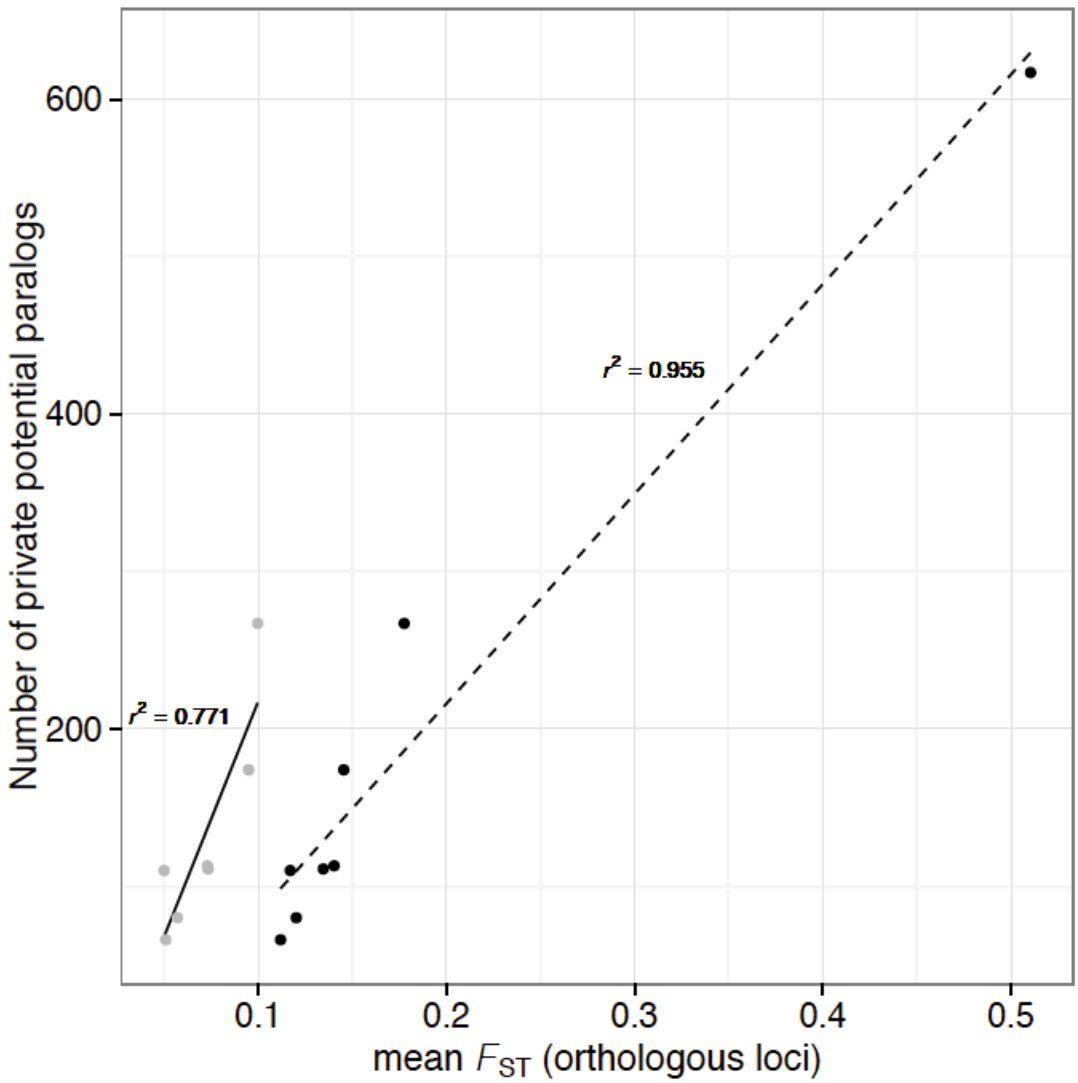


Figure 4.6. The number of private potential paralogs per population increases with their differentiation estimated from orthologous loci. The x axis corresponds to the mean F_{ST} per population from the pairwise matrix among populations and species estimated excluding all potential paralogs. The y axis corresponds to the number of private potential paralogs as in Fig. 4. Regression was performed with the Zamorano population (black dots, dashed line, $F_{1,6}$=128.3, p<0.001) and without it (grey dots, solid line, $F_{1,5}$=16.85, p<0.01).

evidence for the genomic differentiation among our *Berberis* taxa. Firstly, the fact that the population of *B. moranensis* had more paralogous loci than the most differentiated population of *B. alpina* (Cofre de Perote) shows that *B. moranensis* is more differentiated from *B. alpina* than what would be inferred from the PCoA or the F_{ST} values. This highlights that paralogous loci can be an important source

of genomic differentiation among closely related, ecologically divergent and partially sympatric plant lineages. Secondly, the distribution of potential paralogous loci among our *Berberis* species is congruent with the expectation that the independent occurrence of gene duplication within lineages should lead to different species presenting a unique set of paralogs that originated after the speciation event (Lynch & Conery 2000). This has also been shown for species of *Arabidopsis* (Moore & Purugganan 2003) and *Drosophila* (Zhou *et al.* 2008) so in the case of our *Berberis* species it highlights that Cerro Zamorano population is indeed likely to be a different species. The rate of gene duplication could not be estimated due to the uncertainty about divergence in the absence of gene flow between our populations and species, as well as lack of calibration points and reliable nuclear mutation rates for our *Berberis* data. Nevertheless, the independent accumulation of paralogs seems to be linearly correlated with the differentiation estimated from orthologous loci (Fig. 4.6) although the number of private potential paralogous of Cerro Zamorano seems an underestimate based on the trajectory of the previous points. This could be an effect of Cerro Zamorano species being too divergent, leading to the existence of paralogs of older origin that our method would have filtered.

4.5.5. Conclusion

The genomic study of paralogous loci has typically been restricted to highly annotated genomes, or requires transcriptome sequencing (e.g. Lynch & Force 2000; Zhou et al. 2008; Bikard et al. 2009; Warren et al. 2014; Kondrashov et al. 2002). Here, we have shown that GBS can be used to quantify the differential distribution of recently generated paralogs among non-model plant populations

and species. Thus, in addition to producing large amounts of genomic data for traditional population genetics analyses, GBS methods may also be used to investigate gene duplication as a source of population genomic differentiation. As shown here, this is possible despite short sequence reads and lack of previous genomic knowledge of the analysed taxa.

Incorporating gene duplication to population genetics and phylogenetic analyses of GBS data could be then taken further by: (a) including quantitative measurements of paralogous loci into diversity indexes, and (b) by developing analytical tools, such that paralogous loci are not excluded from marker-based datasets, but incorporated into models of allele and genome divergence. This may be relevant for a broad range of taxa, but should be particularly important for plants where gene duplication plays a fundamental role in their evolution.

4.7. Acknowledgements

We thank Associate Editor Yves Van De Peer and two anonymous referees for their constructive comments on an earlier version of the manuscript. Part of the analyses were carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at UEA. This work was supported by Consejo Nacional de Ciencia y Tecnología, by Rosemary Grant Award for Graduate Student Research from the Society for the Study of Evolution to AMY and by Swiss National Science Foundation grants to N. Alvarez and to N. Arrigo.

4.8. Data and code availability

Raw RADseq is available at data Sequence Read Archive SRP035472. Quality processing and assembling details are in Dryad Repository doi:10.5061/dryad.g52m3. R scripts, *Stacks* jobs and processed data as used here are available at doi:10.5061/dryad.n3jk5. Scripts are also available and versioned at https://github.com/AliciaMstt/Berberis_phylogeo.

4.9. References

Arriaga L, Aguilar C, Espinosa D, Jiménez R (1997) Regionalización ecológica y biogeográfica de México. Taller de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad.

Bikard D, Patel D, Metté CL *et al.* (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*, **323**, 623–626.

Bryson RW, García-Vázquez UO, Riddle BR (2012) Relative roles of Neogene vicariance and Quaternary climate change on the historical diversification of bunchgrass lizards (Sceloporus scalaris group) in Mexico. *Molecular Phylogenetics and Evolution*, **62**, 447–457.

Bryson RW, Murphy RW, Lathrop A, Lazcano-Villareal D (2011) Evolutionary drivers of phylogeographical diversity in the highlands of Mexico: a case study of the Crotalus triseriatus species group of montane rattlesnakes. *Journal of Biogeography*, **38**, 697–710.

Carrasco-Ñuñez G, Milán M, Verma SP (1989) Geología del Volcán Zamorano, Estado de Querétaro. *Revista Instituto de Geología*, **8**, 194–201.

Catchen J, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Deschamps S, Llaca V, May GD (2012) Genotyping-by-Sequencing in Plants. *Biology*, **1**, 460–483.

Dou J, Zhao X, Fu X *et al.* (2012) Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. *Biology Direct*, **7**, 17.

Eaton DAR (2014) PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics*, 10.1093/bioinformatics/btu121.

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, **3**, 329–341.

Fitch WM (1970) Distinguishing Homologous from Analogous Proteins. *Systematic Zoology*, **19**, 99.

Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants. *New Phytologist*, **183**, 557–564.

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, wegmental, or by transposition. *Annual Review of Plant Biology*, **60**, 433–453.

Gayral P, Melo-Ferreira J, Glémin S *et al.* (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genet*, **9**, e1003457.

Graham A (1999) The Tertiary history of the Northern temperate element in the Northern Latin American biota. *American Journal of Botany*, **86**, 32–38.

Gugger PF, González-Rodríguez A, Rodríguez-Correa H, Sugita S, Cavender-Bares J (2011) Southward Pleistocene migration of Douglas-fir into Mexico: phylogeography, ecological niche modeling, and conservation of "rear edge" populations. *New Phytologist*, **189**, 1185–1199.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Hohenlohe PA, Catchen J, Cresko WA (2012) Population genomic analysis of model and non-model organisms using wequenced RAD tags. *Methods in molecular biology (Clifton, N.J.)*, **888**, 235–260.

Hurles M (2004) Gene Duplication: The genomic trade in spare parts. *PLoS Biol*, **2**, e206.

Jensen RA (2001) Orthologs and paralogs - we need to get it right. *Genome Biology*, **2**, interactions1002.

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 5048–5057.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. *Genome Biology*, **3**(2):RESEARCH0008.

Langham RJ, Walsh J, Dunn M *et al.* (2004) Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics*, **166**, 935–945.

Lewis EB (1951) Pseudoallelism and gene evolution. *Cold Spring Harbor Symposia on Quantitative Biology*, **16**, 159–174.

Lockton S, Gaut BS (2005) Plant conserved non-coding sequences and paralogue evolution. *Trends in Genetics*, **21**, 60–65.

Lozano-García S, Sosa-Nájera S, Sugiura Y, Caballero M (2005) 23,000 yr of vegetation history of the Upper Lerma, a tropical high-altitude basin in Central Mexico. *Quaternary Research*, **64**, 70–82.

Lynch M, Conery JS (2000) The Evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Lynch M, Force A (2000) The Probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459–473.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014a) Data from: RAD sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Dryad Digital Repository.*, doi:10.5061/dryad.g52m3.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014b) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12291

Mastretta-Yanes A, Wegier A, Vázquez-Lobo A, Piñero D (2011) Distinctiveness, rarity and conservation in a subtropical highland conifer. *Conservation Genetics*, **13**, 211–222.

McCormack JE, Peterson AT, Bonaccorso E, Smith TB (2008) Speciation in the highlands of Mexico: genetic and phenotypic divergence in the Mexican jay (*Aphelocoma ultramarina*). *Molecular Ecology*, **17**, 2505–2521.

Metcalfe SE (2006) Late Quaternary environments of the Northern deserts and central Transvolcanic Belt of Mexico. *Annals of the Missouri Botanical Garden*, **93**, 258–273.

Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15682–15687.

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology*, **8**, 122–128.

Morrell PL, Buckler ES, Ross-Ibarra J (2012) Crop genomics: advances and applications. *Nature Reviews Genetics*, **13**, 85–96.

Morrone JJ, Espinosa-Organista D, Llorente-Bousquets J (2002) Mexican biogeographic provinces: Preliminary scheme, general characterizations, and synonymies. *Acta Zológica Mexicana*, **85**, 83–108.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Ohno S (1970) *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Ornelas JF, Sosa V, Soltis DE *et al.* (2013) Comparative phylogeographic analyses illustrate the complex evolutionary history of threatened cloud forests of Northern Mesoamerica. *PLoS ONE*, **8**, e56283.

Parchman TL, Gompert Z, Mudge J *et al.* (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, **7**, e32253.

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome Journal*, **5**, 92.

Pujolar JM, Jacobsen MW, Als TD *et al.* (2014) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, **23**, 2514–2528.

R. Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rounsaville TJ, Ranney TG (2010) Ploidy Levels and genome sizes of *Berberis* L. and *Mahonia* nutt. species, hybrids, and cultivars. *HortScience*, **45**, 1029–1033.

Rzedowski J, Calderón de Rzedowski G, Zamudio S (2012) La flora vascular endémica en el estado de Querétaro. I. Análisis numéricos preliminares y definición de áreas de concentración de las especies de distribución restringida. *Acta botánica mexicana*, 91–104.

Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biology*, **13**, 243.

Semple C, Wolfe KH (1999) Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *Journal of Molecular Evolution*, **48**, 555–564.

Stölting KN, Nipper R, Lindtke D *et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology*, **22**, 842–855.

Toledo V (1982) Pleistocene changes of vegetation in tropical Mexico. In: *Biological diversification in the tropics* (ed Prance GT), pp. 93–111. Columbia University Press, New York.

Tovar-Sánchez E, Mussali-Galante P, Esteban-Jiménez P *et al.* (2008) Chloroplast DNA polymorphism reveals geographic structure and introgression in the Quercus crassifolia Quercus crassipes hybrid complex in Mexico. *Botany*, **86**, 228–239.

Vázquez-Selem L, Heine K (2011) Late Quaternary Glaciation in Mexico. In: *Quaternary Glaciations - Extent and Chronology - A Closer Look* (eds Ehlers J, Gibbard PL, Hughes P), pp. 849–861. Elsevier.

Volff J-N (2004) Genome evolution and biodiversity in teleost fish. *Heredity*, **94**, 280–294.

Warren IA, Ciborowski KL, Casadei E *et al.* (2014) Extensive local gene duplication and functional divergence among paralogs in Atlantic salmon. *Genome Biology and Evolution*, evu131.

Zamudio S (2009a) Familia Berberidaceae. *Flora del Bajio y regiones adyacentes*, **163**, 1–40.

Zamudio S (2009b) Notas sobre el Género Berberis (Berberidaceae) en México. *Acta Botánica Mexicana*, **87**, 31–70.

Zhou Q, Zhang G, Zhang Y *et al.* (2008) On the origin of new genes in Drosophila. *Genome Research*, **18**, 1446–1455.

4.10. Supporting Information

Table 4.1S. Summary statistics for *B. alpina* ingroup and *B. moranensis* estimated with presumably orthologous loci within *B. alpina*.

Pop. ID	*Ns*	*N*	Priv.	Sites	%poly	P	H_{obs}	π	F_{IS}
Variant positions									
B. alpina									
Pe	8	6.13	1067	5500	39.64	0.9141	0.1186	0.1342	0.0363
Ma	8	6.43	485	5474	31.61	0.9360	0.0934	0.1025	0.0212
Iz	10	8.05	363	5484	30.94	0.9429	0.0883	0.0909	0.0137
Tl	6	4.78	315	5504	28.15	0.9410	0.0874	0.0973	0.0227
Aj	10	8.55	451	5480	34.27	0.9380	0.0972	0.0988	0.0068
To	8	6.32	293	5485	24.92	0.9470	0.0876	0.0850	-0.0001
B. moranensis									
An	9	7.71	194	5498	15.71	0.9518	0.0958	0.0644	-0.0587
All positions (variant and fixed)									
B. alpina									
Pe	8	6.71	1067	450390	0.482	0.9990	0.0014	0.0016	0.0004
Ma	8	6.91	485	450412	0.386	0.9992	0.0011	0.0013	0.0003
Iz	10	8.64	363	450395	0.377	0.9993	0.0011	0.0011	0.0002
Tl	6	5.06	315	450393	0.343	0.9993	0.0011	0.0012	0.0003
Aj	10	9.07	451	450413	0.419	0.9992	0.0012	0.0012	0.0001
To	8	6.78	293	450410	0.304	0.9994	0.0011	0.0010	0.0000
B. moranensis									
An	9	8.18	194	450377	0.191	0.9994	0.0012	0.0008	-0.0007

Results are split into those calculated for only nucleotide positions that are polymorphic in at least one population (top, "Variant positions"), as well as all nucleotide positions across all RAD sites regardless of whether they are polymorphic or fixed (bottom, "All positions"). The first column shows the number of individuals per population that were used for the analysis (Ns). Next are the average number of individuals genotyped at each locus (N), the number of variable sites unique to each population (Priv.), the number of polymorphic (top) or total (bottom) nucleotide sites across the data set (Sites), percentage of polymorphic loci (% poly), the average frequency of the major allele (P), the average observed heterozygosity per locus (H_{obs}), the average nucleotide diversity (π), and the average Wright's inbreeding coefficient (F_{IS}). Populations are ordered East to West, top to bottom. Population IDs as in Fig. 2.

Table 4.2S. Summary statistics for *B. alpina* ingroup with the dataset including putative paralogous loci.

Pop. ID.	*Ns*	*N*	Priv.	Sites	%poly	P	H_{obs}	π	F_{IS}
Variant positions									
Pe	8	5.87	1189	6900	49.55	0.8642	0.2221	0.1976	-0.0385
Ma	8	6.15	551	6919	42.59	0.884	0.1982	0.1668	-0.053

Iz	10	7.67	415	6895	42.29	0.8868	0.1983	0.1596	-0.0653
Tl	6	4.58	353	6876	39.24	0.8856	0.1955	0.1688	-0.0428
Aj	10	8.19	505	6908	45.17	0.8803	0.2106	0.1665	-0.0802
To	8	6	351	6917	36.48	0.8914	0.1972	0.155	-0.0724
All positions (variant and fixed)									
Pe	8	6.67	1189	519559	0.658	0.9982	0.0029	0.0026	-0.0005
Ma	8	6.86	551	519581	0.567	0.9985	0.0026	0.0022	-0.0007
Iz	10	8.58	415	519559	0.561	0.9985	0.0026	0.0021	-0.0009
Tl	6	5.03	353	519522	0.519	0.9985	0.0026	0.0022	-0.0006
Aj	10	9	505	519576	0.6	0.9984	0.0028	0.0022	-0.0011
To	8	6.72	351	519581	0.486	0.9986	0.0026	0.0021	-0.001

Results are split into those calculated for only nucleotide positions that are polymorphic in at least one population (top, "Variant positions"), as well as all nucleotide positions across all RAD sites regardless of whether they are polymorphic or fixed (bottom, "All positions"). The first column shows the number of individuals per population that were used for the analysis (Ns). Next are the average number of individuals genotyped at each locus (N), the number of variable sites unique to each population (Priv.), the number of polymorphic (top) or total (bottom) nucleotide sites across the data set (Sites), percentage of polymorphic loci (% poly), the average frequency of the major allele (P), the average observed heterozygosity per locus (H_{obs}), the average nucleotide diversity (π), and the average Wright's inbreeding coefficient (F_{IS}). Populations are ordered East to West, top to bottom. Population IDs as in Fig. 2.

Table 4.3S. Pairwise F_{ST} with the dataset including putative paralogous loci (all loci)

	Iz	*Ma*	*Pe*	*Tl*	*To*	Za	An	Out
Aj	0.0338	0.0618	0.0848	0.0253	0.0442	0.3658	0.0720	0.3186
Iz		0.0596	0.0857	0.0253	0.0526	0.3769	0.0904	0.3315
Ma			0.0750	0.0512	0.0761	0.3614	0.1129	0.3196
Pe				0.0725	0.1000	0.3320	0.1327	0.2717
Tl					0.0447	0.3597	0.0821	0.3073
To						0.3981	0.0990	0.3501
Za							0.4548	0.3960
An								0.3993

Berberis alpina ingroup populations are shown in italics in the first five columns. *B. moranensis* (An) and *B. trifolia* (Out) are shown as a reference for the values found among different species. Cerro Zamorano (Za) population shows F_{ST} values higher than those found for *B. moranensis* (An) and *B. trifolia* (Out).

Table 4.4S. Pairwise F_{ST} for the putative orthologs within *B. alpina* subset

	Iz	*Ma*	*Pe*	*Tl*	*To*	Za	An	Out
Aj	0.0425	0.0789	0.1060	0.0327	0.0584	0.4570	0.0970	0.3995
Iz		0.0773	0.1120	0.0349	0.0706	0.4758	0.1260	0.4191
Ma			0.0992	0.0678	0.0999	0.4586	0.1577	0.4093
Pe				0.0928	0.1288	0.4155	0.1766	0.3412
Tl					0.0589	0.4551	0.1184	0.3956
To						0.5060	0.1441	0.4499
Za							0.5862	0.4889
An								0.5216

Berberis alpina ingroup populations are shown in italics in the first five columns. *B. moranensis* (An) and *B. trifolia* (Out) are shown as a reference for the values found among different species. Cerro Zamorano (Za) population shows F_{ST} values higher than those found for *B. moranensis* (An) and *B. trifolia* (Out).

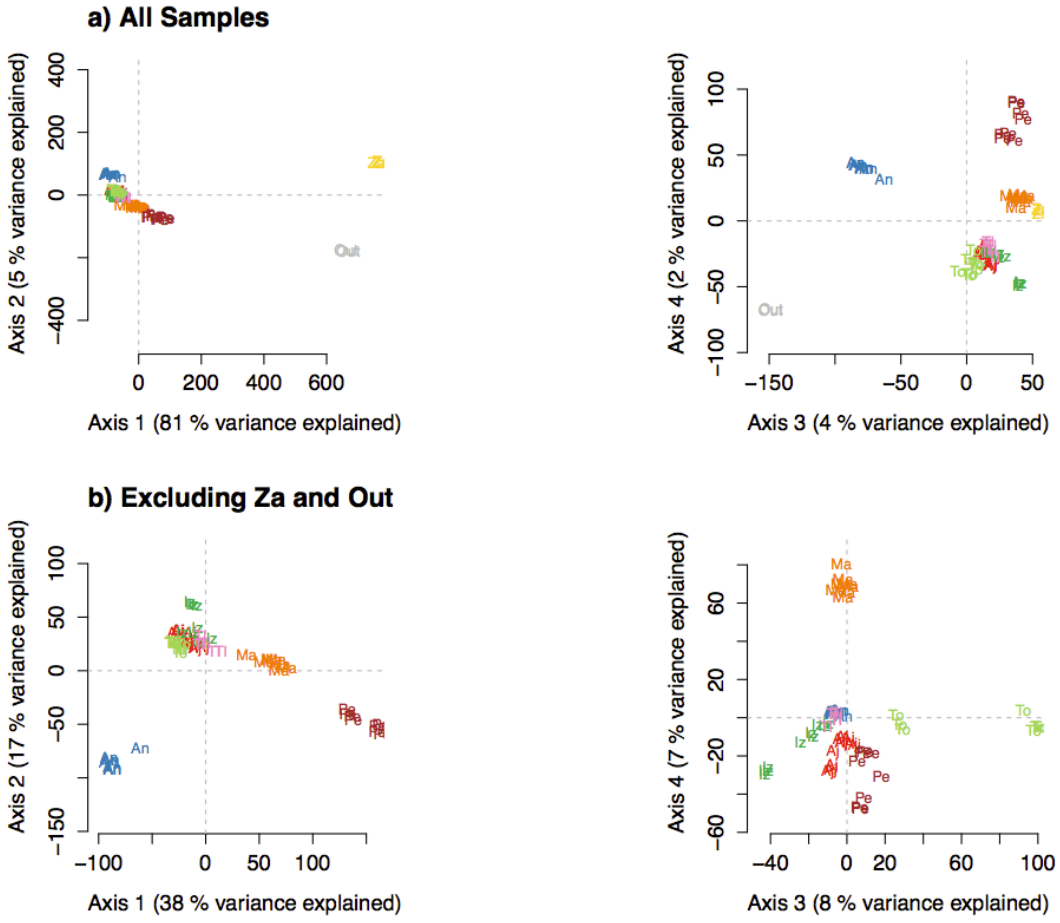


Figure 4.1S. Principal coordinates analysis of the SNP-loci including paralogs showing the four first axes that explain most of the variation. (a) When all samples are analyzed axis 1 explains 81% of the variation and corresponds to the differences between Cerro Zamorano-*B. trifolia* (Za and Out, respectively) and the rest of the populations. (b) If Cerro Zamorano and *B. trifolia* are excluded, axis 1 and 2 separate *B. moranensis* (An) and the Cofre de Perote and Malinche (Pe and Ma) populations of *B. alpina*, explaining 38% and 17% of the variance, respectively. Populations ID and colors as in Fig. 4.2.

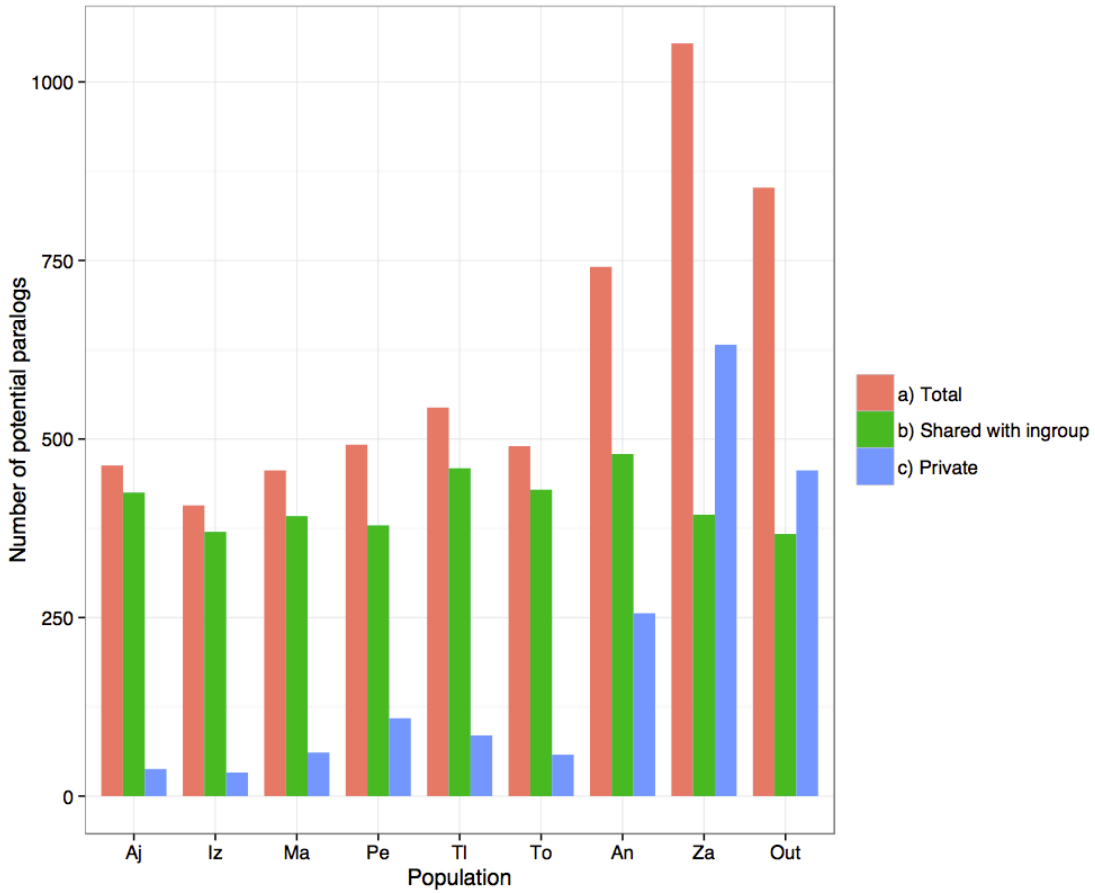


Figure 4.2S. Distribution of RAD-loci with at least one SNP-locus where the frequency of the major allele (*p*) equals 0.5 (potential paralogs) under unequal sampling size among populations (6-10 individuals for the first seven populations, four for Za and two for Out). a) There are more loci biased towards *p*=0.5 in *Berberis moranensis* (An), the Zamorano population (Za) and *B. trifolia* (Out) than in *B. alpina* ingroup populations (Aj-To). b) Most of the loci where *p*=0.5 are the same loci in *B. alpina* ingroup and any given population or species, but c) a substantial proportion of loci show *p*=0.5 exclusively in *B. moranensis*, the Zamorano population or *B. trifolia*. The distribution of total, private and shared potential paralogous loci is similar to what was found under equal sampling sizes (n=4; Fig. 4). The difference being that ~130 more potential paralogous loci per population are found when decreasing the sampling size from 6-10 to 4.

CHAPTER 5

Patterns of genetic differentiation on tropical mountains: a comparative landscape genomics approach

One must have a mind of winter
to regard the frost and the boughs
of the pine-trees crusted with snow,

And have been cold a long time
To behold the junipers shagged with ice,
The spruces rough in the distant glitter

-A fragment of *The Snow Man* by Wallace Stevens

Biodiversity thinking, ticatla 2012

5.1. Abstract

Tropical mountains are thought to be areas of high species diversity and endemism due to historical variables, namely that they: (1) allow for long-term population persistence despite global climate fluctuations, and (2) promote diversification by creating fragmented and isolated habitats that are prone to allopatric speciation. These two processes have been examined with species occurrence data and estimations of species divergence times. However, there remains a need for intraspecific analyses of the mechanisms by which endemism may emerge from its most fundamental evolutionary origin: genetic differentiation among populations. Here, we use genomic SNP data of two plant species and landscape analyses to test for habitat persistence and population genetic differentiation within the Transmexican Volcanic Belt, an archipelago of tropical sky-islands. We show that mountains have facilitated population persistence throughout glacial/interglacial cycles within a short geographic distance, and that genetic differentiation can be explained by the degree of glacial habitat connectivity among mountains. Our study supports, from an intraspecific perspective, the role of tropical mountains as cradles for biodiversity.

Keywords: ddRAD, *Juniperus monticola, Berberis alpina*, Transmexican Volcanic Belt, endemism, biodiversity distribution

5.2. Introduction

Low-latitude mountains are biodiversity hotspots (Myers *et al.* 2000). Their level of species richness is particularly high due to the presence of both taxa with wide-ranging distributions, as well as a high aggregation of locally endemic species (Kruckeberg & Rabinowitz 1985; Jetz *et al.* 2004). Contemporary environmental variables can provide good explanation for the regional variation in richness of wide-ranging species, but the excess of endemism present in tropical mountains exceeds what can be predicted using macro-ecological variables alone (Jetz & Rahbek 2002; Rahbek *et al.* 2007). This excess can, however, be explained if analyses incorporate the history of species and their habitats (Jetz *et al.* 2004; Graham *et al.* 2006; Fjeldså *et al.* 2012). The main conclusion of this integrative approach is that tropical mountains are rich in biodiversity because they promote both species diversification and long-term population persistence (Fjeldså *et al.* 2012). This new approach represents an exciting advance that calls for evolutionary data, such as that provided by phylogenetic and phylogeographic approaches, because it can increase our understanding of how biodiversity is structured geographically in a temporal context.

High levels of endemism within tropical mountains have been associated with both the increasing isolation and decreasing surface area of high mountain regions, leading to small and fragmented populations. Such populations should be prone to allopatric speciation, therefore enhancing the evolution of many new, endemic taxa (Kessler 2002). This has been found in several studies and is the most commonly cited explanation for elevational patterns of endemism (Kessler

2002). Parapatric speciation can also occur, although it seems to be a less frequent phenomena (Weir 2009; Cadena *et al.* 2011; Päckert *et al.* 2012). As for promoting population persistence trough time, tropical mountains have been found to be areas of low climate change velocity, meaning they are areas where biodiversity can survive through global climate fluctuations by undertaking altitudinal shifts instead of long latitudinal movements (Loarie *et al.* 2009; Sandel *et al.* 2011). Areas of low climate change velocity thus allow for long term population persistence relatively *in situ,* in contrast to the longer range shifts or extinctions that the Pleistocene climate fluctuations caused at higher latitudes and shallower lands (Hewitt 1996; Sandel *et al.* 2011). Population persistence is meaningful for the accumulation of endemism because it can be translated into lack of extinction, thus leading to the local aggregation of old endemic species (Fjeldså *et al.* 1999). The diversification and long-term persistence hypotheses have been examined using species occurrences (e.g. Sandel *et al.* 2011; Krömer *et al.* 2013) and more recently incorporating molecular data for estimating species divergence times (e.g. Smith *et al.* 2014). Study areas range from coarse continental data (Rahbek *et al.* 2007; Sandel *et al.* 2011; Fjeldså *et al.* 2012) to more detailed analyses of specific mountain ranges such as the Andes (e.g. Fjeldså *et al.* 1999; Kessler 2002), the Himalayas (e.g. Päckert *et al.* 2012) and the Eastern Arc Mountains of Tanzania and Kenya (Fjeldså & Bowie 2008). Although these studies have included an evolutionary perspective by analysing species ranges among with phylogenetic data, there remains a need for intraspecific analyses of the mechanisms by which endemism may emerge from its most basal evolutionary origin: genetic differentiation among populations.

Here, we aim to address this knowledge gap by testing for habitat persistence and population genetic differentiation within recently emerged (Pleistocene) high-altitude tropical mountains. We take a population level approach because it is expected that areas that facilitate population persistence over phylogeographic (intraspecific) timescales should, in the absence of further geological change, also be stable across phylogenetic (interspecific) timescales, such that regions of genetic endemism will eventually lead to regions of high species diversity (Hugall *et al.* 2002; Carnaval *et al.* 2009). Thus, testing for (i) areas that facilitate long-term population persistence, and (ii) topographic variables that promote population genetic differentiation can contribute to the evolutionary understanding of tropical mountain biodiversity.

Our study area includes the highest mountains (>3,000 masl) of the Transmexican Volcanic Belt (TMVB, Fig. 5.1). The area comprises an archipelago of sky-islands at ~19ºN, within which the highest stratovolcanoes emerged during the last 1.5 Myr (Ferrari *et al.* 2012). Species in these mountains have likely been restricted to high-elevation refugia during the interglacial periods of the Pleistocene, such as now, where divergence could be promoted by restricted gene flow. During glacial periods such species may be expected to experience genetic admixture at lower elevations, as their ranges spread to lower altitudes (Toledo 1982).

Here we suggest that mountains of the TMVB where alpine-grasslands presently exist may have provided long-term environmentally stable conditions for this ecosystem to have persisted continually throughout glacial/interglacial cycles within a short geographic distance. We then hypothesise that genetic differentiation among populations and private genetic variation within

populations of species characteristic of the timberline-alpine grasslands would be a function of the historical environmental isolation. To examine this, we focus on two timberline-alpine grassland plant species of the TMVB for which we generated thousands of genomic SNP data, and on the glacial/interglacial distribution of their habitat type. First, we examine if suitable conditions persisted within the same relative area through glacial/interglacial stages. Second, we examine if genetic differentiation can be better explained by the degree of historical or present habitat connectivity among mountains. Finally, we examine if patterns of private allelic variation (as a surrogate of genetic endemism) are related to the isolation degree of each mountain.

5.3. Methods

5.3.1. Study system and sampling

Juniperus montiocla (Cupressaceae) and *Berberis alpina* (Berberidaceae) are shrubs that grow from 3,300 to 4,200 metres above sea level (masl) on rocky formations from the timberline and alpine grasslands of the TMBV and nearby highlands. They are closely related to *J. flaccida* and *B. moranensis*, respectively, which grow at lower altitudes (800-2,600 masl and 1,800-3,150 respectively). *Berberis alpina* populations from outside the TMVB likely represent a different species (Mastretta-Yanes *et al.* 2014c) and were excluded from this analysis. *Juniperus monticola* populations from the TMVB are recognized as the varieties *J. m. compacta* and *J. m. orizabensis* (Adams 2008), and what used to be considered the most northern populations (Cerro Potosí, on the Sierra Madre Oriental) are

now accepted as a different species (*J. zanonii*) belonging to a different clade of junipers (Adams *et al.* 2010).

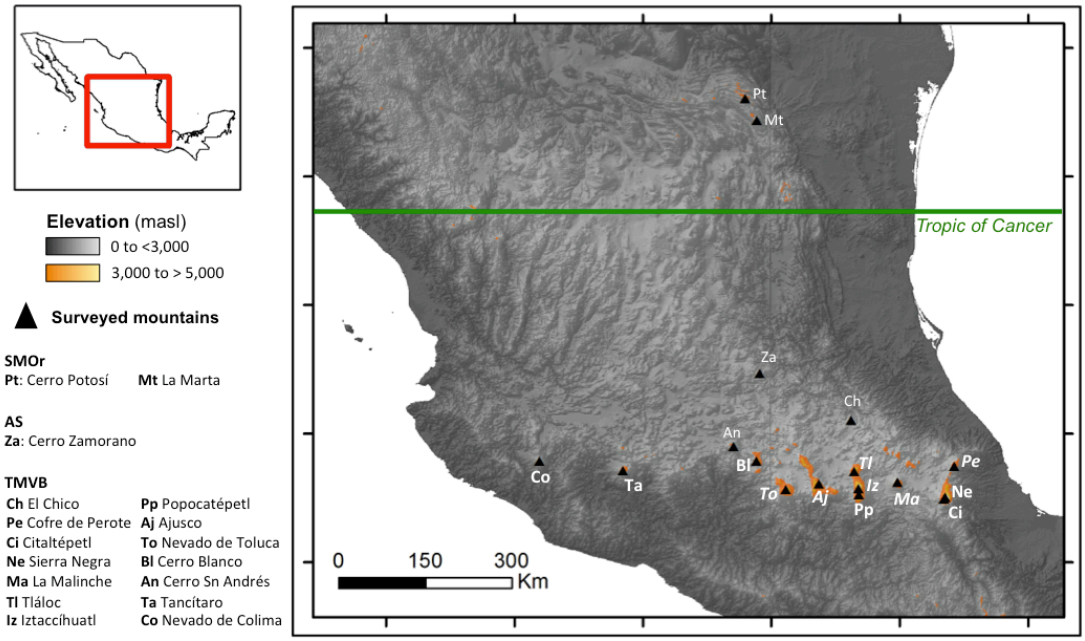


Figure 5.1. High elevation mountains with timberline - alpine grasslands surveyed (triangles) for *Juniperus monticola* and *Berberis alpina* in the Sierra Madre Oriental (SMOr), the Altiplano Sur (AS) and the Transmexican Volcanic Belt (TMVB). *Juniperus monticola* was found in populations Ch, Pe, Ci, Ne, Ma, Tl, Iz, Pp, Aj, To, Bl, Ta and Co (bold) and *B. alpina* in populations Pe, Ma, Tl, Iz, Aj and To (italics).

Mountain peaks from >3,000 masl within the TMVB and nearby areas of the Altiplano Sur (AS) and the Sierra Madre Oriental (SMOr) were surveyed for *B. alpina* and *J. monticola* during September-October 2010 and April-May 2011 (Fig. 5.1). *B. alpina* was found in a total of six locations, and *J. monticola* in 13, which represent their known distribution within the TMVB and the AS. Samples of the closely related species and outgroups *B. moranensis, B. trifolia, B. pallida, J. flaccida, J. zanonii* and *J. deppeana* were collected at lower elevations (~2,000-3,150 for *Berberis* and 800-2,500 masl for *Juniperus*) of the TMVB and at northernmost localities of the SMOr and Sierra Madre Occidental (SMOcc) in October 2010 and 2012. Sampling was performed with SEMARNAT permission

No. SGPA/DGGFS/712/2896/10. Herbarium specimens of *B. alpina, B. moranensis, J. flaccida* and *J. monticola* were prepared and deposited within the Herbario Nacional in Mexico City (MEXU) or within Herbario CIIDIR in Durango.

5.3.2. Molecular methods

Based on data from related species, the sampled *Berberis* species are likely diploid with a genome size of between 0.50 to 1.83 Gbp (Rounsaville and Ranney, 2010), while the *Juniperus* are also likely diploid but with a genome size of 9 to 10 Gbp (Zonneveld 2012). For both taxa ddRAD libraries were prepared using modified versions of protocols by Parchman et al. (2012) and Peterson et al. (2012). For *Berberis* the enzyme pair EcoRI-HF and MseI was used while for *Juniperus* the rare cutter SbfI-HF was used instead of EcoRI-HF, thus allowing for a narrower subsampling of the juniper's large genome. Samples were randomly divided into three (*Berberis*) or 10 (*Juniperus*) groups with a common sequencing index (ddRAD libraries hereafter). All *Berberis* and two *Juniperus* libraries were sequenced using single-end reads (100bp long) in a separate lane of an Illumina HiSeq2000, while two libraries were sequenced in a single lane of the same platform for the rest of the *Juniperus* libraries. Further details on *Berberis* laboratory protocol and sequencing output are detailed in Mastretta-Yanes et al. (2014a). For *Juniperus* this information is available in Supporting Information 1.

The *Berberis* dataset consists of 75 individually tagged specimens of *B. alpina* and *B. moranensis* (6-10 per mountain), three samples of each outgroup (*B. trifolia* and *B. pallida*) and 15 replicated samples, with at least one replicate per population or species. The *Juniperus* dataset consists of 137 individually

tagged specimens of *J. monticola* (10 per mountain), four of *J. flaccida* and one of *J. deppeana,* one of *J. zanonii,* 10 negative controls and 20 replicated samples, with at least one replicate per sampling locality or species (excepting *J. deppeana*).

5.3.3. Sequencing output, de novo *assembly and loci filtering of RAD data*

Complete details of *Berberis* sequencing output and quality filtering are available in Mastretta-Yanes *et al.* (2014b). Briefly, after demultiplexing and quality trimming of *Berberis* raw reads, final sequences were 84 bp long. *Juniperus* raw reads were demultiplexed and quality filtered using *Stacks* v. 1.17 by: (1) truncating final read length to 87 (because there was a quality drop after this position in library 10); (2) removing any reads with an uncalled base; (3) discarding reads with low quality scores (score limit 22 to 28, depending on the library); (4) discarding reads that have been marked by Illumina's chastity filter as failing; (5) filtering adapter sequences, and; (6) rescuing tags (maximum distance of one between barcodes). See Supporting Information 1 for full details on *Juniperus* lab protocol and bioinformatics pipeline.

Here we refer to a RAD-locus as a short DNA sequence produced by clustering together RAD-alleles; in turn, RAD-alleles differ from each other by a small number of SNPs in certain nucleotide positions (SNP-loci). Data was *de novo* assembled using the software *Stacks* (Catchen *et al.* 2011, 2013). Data from *Berberis* had been previously assembled in *Stacks* v. 1.02 with the parameter values *m*=3, *M*=2, *N*=4, *n*=3, *max_locus_stacks*=3 and a SNP calling model with an upper bound of 0.05 (Mastretta-Yanes *et al.* 2014b). Stacks v. 1.17 was used for *Juniperus* with the parameter values *m*=10, *M*=2, *N*=4, *n*=3, *max_locus_stacks*=4

and default SNP calling model. These settings were chosen after testing a wide range of parameters as in Mastretta-Yanes *et al.* (2014b), and optimising the recovery of a large number of loci while reducing the SNP and RAD allele error rates. After *de novo* assembly, the data were filtered to keep only those samples having more than 50% and 35% of the mean number of loci per sample for *Berberis* and *Juniperus,* respectively, and only those loci present in at least 80% of *Berberis* samples and 70% of *Juniperus*. Putative paralogous loci of the *Berberis* dataset were filtered by identifying loci where the frequency of the major allele equalled p=0.5 in more than one population or species, as detailed in Mastretta-Yanes *et al.* (2014c). For the *Juniperus* dataset the same procedure was followed, but with the following modifications: (1) putative paralogous loci had to meet the extra condition of showing deviations from Hardy-Weinberg Equilibrium (HWE, H_{obs} >0.9, negative F_{IS} or F_{IS}=1), and (2) putative paralogous loci private to a single population of *J. monticola* were also excluded by identifying loci where p=0.5 in any single sampling location, present in more than three individuals of that population and showing deviations from HWE. To ameliorate the effect of missing data on population genetics statistics, RAD-loci that were present in several sampling locations but represented by only one individual in any given population were also filtered. These extra conditions were not performed in the *Berberis* dataset due to the small sample sizes for some sampling locations. Replicates were used to estimate error rates for both taxa as in Mastretta-Yanes *et al.* (2014b). For the population genomic analyses, only one sample for each replicate pair was used, along with all the remaining non-replicated samples.

Considerably fewer loci were recovered in *Berberis pallida*, compared to

the other *Berberis* species, which is likely explained by mutations affecting restriction enzyme cutting sites and hence a distant evolutionary relationship with the other species in the study. This species was therefore excluded from further analyses.

5.3.4. Population genomics statistics and population differentiation

The *populations* program of *Stacks* was used to estimate the number of private alleles, the percentage of polymorphic loci, heterozygosity, π, and F_{IS} at each nucleotide position for each sampling location (mountain) of the ingroup species. Pairwise F_{ST} values were estimated, defining each sampling location as a population. Only the first SNP of each RAD-locus was used for these estimations. SNP data was exported to plink format and analysed with custom R v. 2.15.1 (R. Core Team 2012) scripts to perform Principal Coordinate Analyses (PCoA) both with and without outgroups.

5.3.5. Timberline- alpine grassland distribution of glacial and interglacial periods

Juniperus monticola and *B. alpina,* do not occur in all mountains where suitable habitat (timberline-alpine grassland) occurs within the TMVB. Thus, rather than independently modelling each species distribution with few data points the distribution of their habitat was modelled using confirmed data points of timberline-alpine grasslands of the TMVB. This "ecosystem approach" is similar to how Graham *et al.* (2006) model rainforest expansion and contraction across climate fluctuations to examine the effect of habitat persistence on rare species occurrence, and although this approach has been shown to perform below average with respect to model sensitivity, it excelled in specificity statistics and

robustness against extrapolations far beyond training data, suggesting that the ecosystem approach is well suited to reconstruct historical biogeographies and glacial distributions (Roberts & Hamann 2012).

As presence points we used alpine grassland herbarium records (n=72), *Pinus hartwegii* (a pine species characteristic of the forests reaching the timberline of the TMVB and present in all mountains with alpine grasslands) occurrence points (n=7) and the sampling points of *J. monticola* of the present study (n=13). Alpine grasslands records come from specimens in the herbaria ENCB, IEB, MEXU and XAL having "alpine grassland" or "pastizal alpino" in the vegetation description, and were corroborated in the field. Occurrence points of *Pinus hartwegii* were downloaded from GBIF using the following filters: boundary box (-108.457031 23.241346,-108.457031 14.306969,-89.736328 14.306969,-89.736328 23.241346,-108.457031 23.241346), without spatial issues, with coordinates, and recorded after 1997 (since previous years contained mostly entries whose geographic coordinates were not obtained directly with a GPS, thus making them less reliable). All occurrences were visually inspected on Google Earth to ensure they were likely on *P. hartwegii* forest. Occurrences with duplicated coordinates were filtered leaving only a unique point. Since spatial autocorrelation can lead to over-prediction, all presence points that were closer together than the minimum resolution of the climate layers (1 km) were filtered keeping only one of the points. Geographic distances were calculated using the Geographic Distance Matrix Generator v. 1.2.3 (http://biodiversityinformatics.amnh.org/open_source/gdmg/), using a WGS84 spheroid. The final number of presence points was 45.

As environmental data, the 19 bioclimatic layers of Hijmans *et al.* (2005) were reduced to the area of interest (a polygon ranging from 22-19°N and 96-106°W). A Principal Component Analysis was performed to avoid over fitting. Only the independent variables with the highest contribution to variance were considered. Maxent v. 3.3.3k (Phillips *et al.* 2006) was used for the timberline-alpine distribution modelling. This method uses a maximum entropy approach and presence-only data. The potential distribution of the timberline-alpine grassland was projected to the LGM using the bioclimatic layers obtained from CCSM and MIROC initiatives (Braconnot *et al.* 2007). The analyses were performed with 10 bootstrap replicates and a random seed.

5.3.6. Measuring effective distances

Resistance distances (McRae 2006) were used to estimate the effective distance among sampling localities. This method is based on circuit theory and considers multiple potential paths of least resistance between sampling points (McRae 2006), thus performing better than similar approaches like least-cost path analysis (McRae & Beier 2007; Moore et al. 2011).

To estimate resistance distances, the pairwise mode of the program *Circuitscape* v. 3.5.8 (McRae 2006; McRae & Beier 2007) was used using the sampling locations of *B. alpina* and *J. monticola* as focal points and using as a conductance grid (the reciprocal of the resistance) the 13 resistance surfaces described below. The cell connection scheme was set to eight neighbours and connection calculation was performed based on average resistance. The average effective distance of each sampling locality to the rest of the sampling localities was estimated from the pairwise distance matrix.

The 13 resistance surfaces used here were based on: (i) environmental modelling ("present" and "CCSM" and "MIROC" for the LGM); (ii) a "flat" landscape, and; (iii) elevation data (above 1800, 2000, 2300, 2500, 2700, 3000, 3300, 3500 and 4000 masl). All grids where reclassified so that cells suitable for the occurrence of populations (thus promoting gene flow by admixture) had a value of 1 (high conductance) and those unsuitable were set to 0.1 (high resistance, Fig. 5.2). Reclassifying of each grid was performed with the *raster* R package (Hijmans *et al.* 2014). To define suitable conditions for the "present" surface, a threshold was defined based on the cell values where the presence points fell, so that cells below the value of the point with the lowest probability were classified as unsuitable and above (inclusive) were set as suitable. For the models of the LGM a similar strategy was followed, but the threshold was defined based on the value obtained for a cell where fossil records indicate the occurrence of grasslands during the LGM (Lozano-García *et al.* 2005). In a 'flat' landscape surface all grid cells had the same value. This is equivalent to testing for isolation by distance (IBD) using Euclidean distances, but it takes into account the fact that the underlying landscape is bounded and not infinite (Lee‐Yaw *et al.* 2009; Moore *et al.* 2011). The flat surface was generated by reclassifying the raster from the elevation model, such that all cell values were equal to one. The elevational surfaces were generated by reclassifying the cell values of an elevation raster such that values above (inclusive) a given altitude were set as suitable, and below as unsuitable. All 13 resistance surfaces and sampling points are shown in Fig. 5.3.

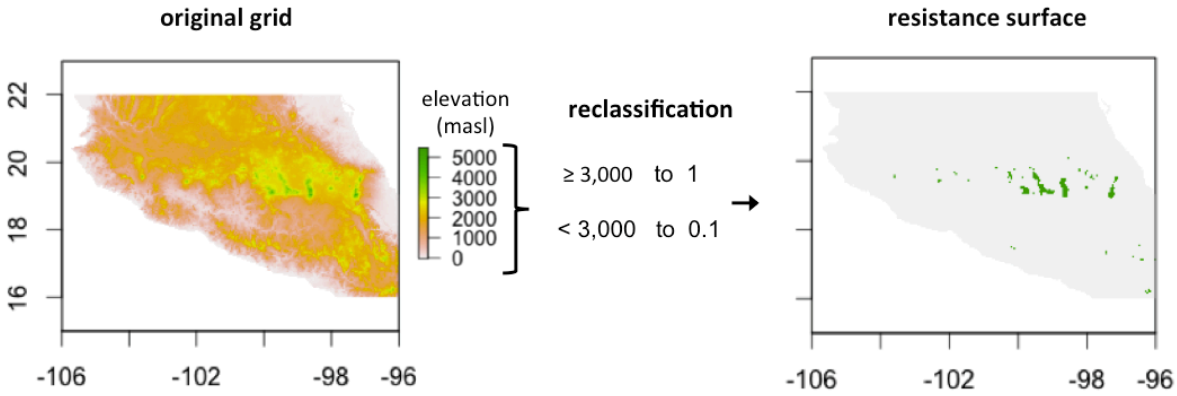


Figure 5.2. Generation of resistance surfaces. The example illustrates how cells from an elevation grid with altitudes equal to or higher than 3,000 masl are assigned a value of high conductance (1, green) and cells with lower altitudes a value of high resistance (0.1, grey). Numbers on the x and y axes represent latitude and longitude, respectively.

5.3.7. Landscape genomics analyses

To examine if genetic differentiation and endemism can be explained by the degree of historical spatial isolation among mountains we tested for (i) isolation by resistance (IBR) *vs* IBD, and (ii) a relationship between the isolation degree of each sampling site and its number of private alleles. To test for IBR a Mantel test with permutations and a linear regression were performed between the pairwise effective distances for each resistance surface and the genetic differentiation matrices. Mantel tests were performed with 10,000 permutations using the F_{ST} pairwise matrix of each species. For the linear regression the genetic differentiation matrices were linearized using the formula for isolation by distance F_{ST} /(1 − F_{ST}) as advocated by Rousset (1997). To test for a positive relation between isolation and genetic endemism, linear regressions were performed between the mean effective distance of each sampling site and the number of private alleles per population. Tests were carried out independently for both species and for the *J. monticola* subset of populations, excluding Nevado

de Colima and Tancítaro (see discussion for reasons). Analyses and plotting were performed with R using the packages *ape* (Paradis *et al.* 2004) and *ggplot2* (Wickham & Chang 2013).

5.4. Results

5.4.1. RAD-seq data yield and error rates

Berberis data used here correspond to Mastretta-Yanes *et al.* (2014c) subset of "putative orthologs within *B. alpina*". In total, the dataset contains 5,461 RAD-loci and 5,274 SNP-loci with error rates for RAD-locus, RAD-allele and RAD-SNP of 17.28% (SD 10.3), 4.1% (SD 1.2) and 1.5% (SD 0.04), respectively, 17% of missing data and mean coverage of 10.5 (SD 4.3).

For the *Juniperus* data, a total of 3,249 RAD-loci, containing 11,407 SNPs (i.e. most RAD-loci had three of more SNPs) with a mean coverage of 84.60 (SD 50.06) were recovered after filtering potential paralogous loci and loci not sufficiently represented among individuals of each sampling location. Only the first SNP of each RAD-locus was used for population genomics analyses, yielding a of total 3,181 SNPs when the outgroups were included, with a RAD-locus error rate of 21% (SD 15), an allele error rate of 1.8% (SD 2.3), a SNP error rate of 1.5% (SD 1.4) and 18% missing data. For the *J. monticola* ingroup dataset 2,925 SNPs were recovered, with a RAD-locus error rate of 21% (SD 15), an allele error rate of 1.8% (SD 2.3), a SNP error rate of 1.4% (SD 0.08) and 16% of missing data (Supporting Information 1).

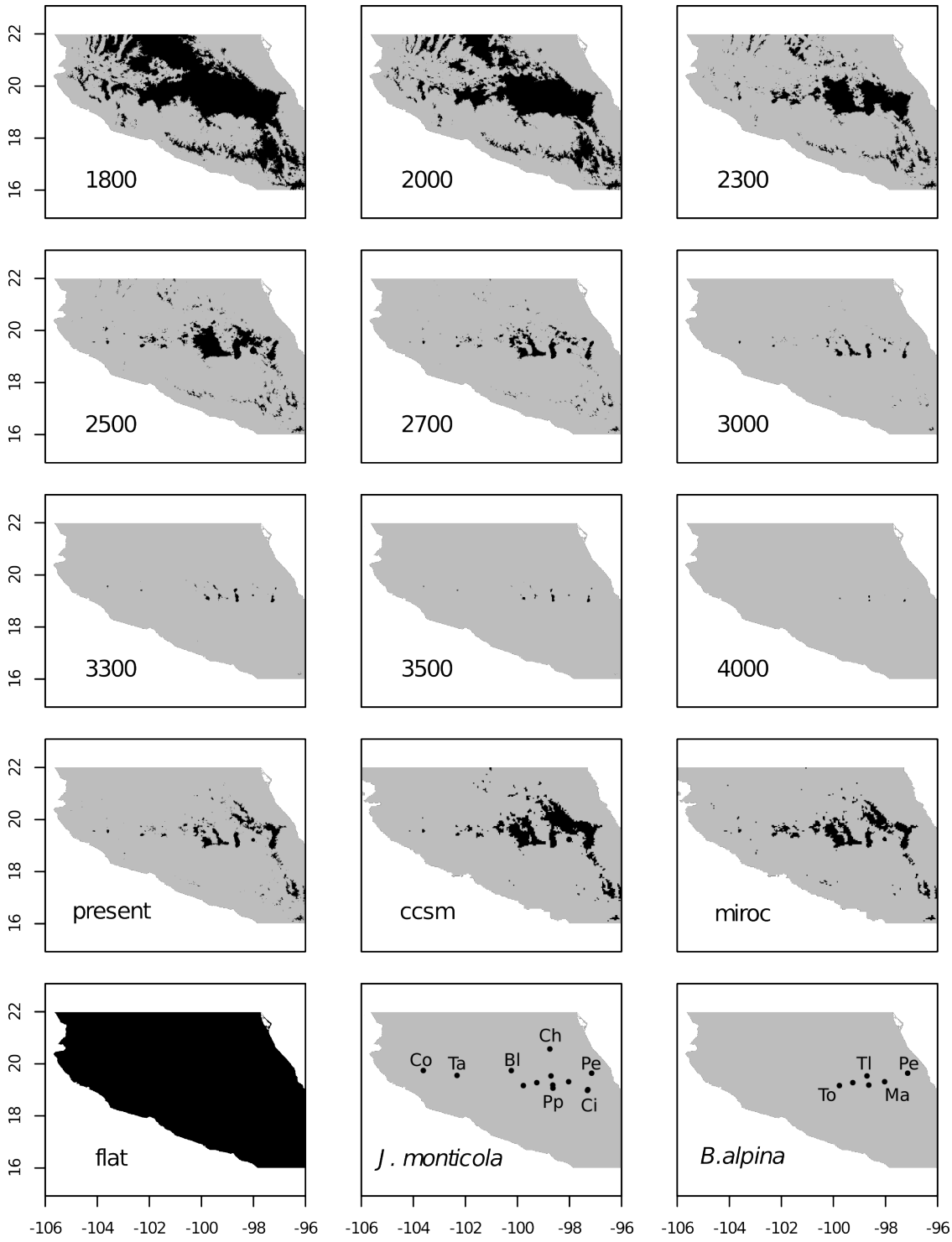


Figure 5.3. Resistance surfaces used to estimate effective distances among populations. Areas allowing the highest gene flow are shown in black. The first three rows show the surfaces using the elevation data; the fourth row uses the distribution modeling for the timberline-alpine grassland for the present and the LGM (CCSM and MIROC layers); the last row shows a landscape where all cells have high conductance ('flat' landscape) and sampling points for *J. monticola* and *B. alpina.* Some mountain names are indicated for reference (ID codes as in Fig. 5.1). For all panels, numbers on the x and y axes represent latitude and longitude, respectively.

5.4.2. Population genomics statistics and population differentiation

When considering only the variant positions (polymorphic in at least one population) for *B. alpina* the percentage of polymorphic loci (notice that locus here refers to a nucleotide position within the RAD-loci) within a given population ranged from 26% to 41%; the average frequency of the major allele from 0.9108 to 0.9449; H_{obs} from 0.091 to 0.123; π from 0.088 to 0.139 and F_{IS} from 0.0004 to 0.0374 (Table 5.1). Cofre de Perote has substantially more private alleles (1,101) than both the remaining populations (332-503, Table 5.1). For *J. monticola* the percentage of polymorphic loci ranged from 19% to 32%; the average frequency of the major allele from 0.9421 to 0.9549; H_{obs} from 0.0495 to 0.0936; π from 0.0706 to 0.0936 and F_{IS} from 0.0326 to 0.0777 (Table 5.1). The El Chico population has substantially more private alleles (608) compared to other populations (206-431).

Pairwise F_{ST} values for *B. alpina* populations ranged from 0.056 to 0.123 and were significant, with the Cofre de Perote population showing the highest levels of differentiation and Tlaloc the smallest (Table 5.2). For *J. monticola* F_{ST} ranged from 0.022 to 0.074, with La Malinche population showing the highest values of differentiation and Tlaloc the smallest (Table 5.3).

Table 5.2. Pairwise F_{ST} among *B. alpina* populations

	Pe	***Ma***	***Tl***	***Iz***	***Aj***
Ma	0.0997				
Tl	0.0928	0.0682			
Iz	0.1121	0.0776	0.0350		
Aj	0.1060	0.0796	0.0325	0.0427	
To	0.1289	0.1003	0.0590	0.0707	0.0577

Population IDs as in Fig. 5.1.

Table 5.3. Pairwise F_{ST} among *J. monticola* populations

	Ch	*Pe*	*Ci*	*Ne*	*Ma*	*Tl*	*Iz*	*Pp*	*Aj*	*To*	*Bl*	*Ta*
Pe	0.035											
Ci	0.045	0.036										
Ne	0.037	0.031	0.011									
Ma	0.049	0.052	0.052	0.050								
Tl	0.026	0.032	0.042	0.032	0.042							
Iz	0.028	0.038	0.044	0.039	0.043	0.018						
Pp	0.031	0.042	0.045	0.040	0.051	0.023	0.022					
Aj	0.029	0.038	0.049	0.041	0.051	0.023	0.027	0.031				
To	0.042	0.061	0.067	0.060	0.074	0.038	0.037	0.046	0.049			
Bl	0.034	0.044	0.052	0.047	0.058	0.032	0.035	0.038	0.035	0.052		
Ta	0.043	0.053	0.064	0.061	0.073	0.047	0.044	0.049	0.050	0.067	0.052	
Co	0.035	0.045	0.055	0.048	0.062	0.034	0.038	0.038	0.041	0.054	0.039	0.034

Table 5.1. Summary population genetic statistics for *B. alpina* and *J. monticola*

Pop. ID	***Ns***	***N***	**Priv.**	**Sites**	**%poly**	**P**	H_{obs}	**π**	F_{IS}
B. alpina									
Pe	6	6.11	1101	5312	41.10	0.9108	0.1234	0.1395	0.0374
Tl	10	4.76	332	5314	29.30	0.9383	0.0917	0.1020	0.0235
Ma	10	6.42	503	5327	32.72	0.9338	0.0967	0.1060	0.0219
Iz	8	8.03	375	5314	32.14	0.9404	0.0924	0.0951	0.0141
Aj	8	8.54	477	5323	35.54	0.9357	0.1006	0.1025	0.0073
To	8	6.31	326	5324	25.85	0.9449	0.0908	0.0883	0.0004
J. monticola									
Ch	8	7.03	608	8173	32.93	0.9421	0.0689	0.0936	0.0650
Pe	5	4.04	206	8097	19.38	0.9549	0.0577	0.0741	0.0326
Ci	10	9.06	176	8185	26.74	0.9515	0.0583	0.0757	0.0465
Ne	10	8.70	177	8183	27.32	0.9522	0.0582	0.0756	0.0460
Ma	9	7.27	175	8168	22.55	0.9543	0.0487	0.0713	0.0564
Tl	10	8.81	324	8195	32.29	0.9461	0.0661	0.0860	0.0554
Iz	8	6.64	265	8170	28.60	0.9482	0.0651	0.0844	0.0492
Pp	8	7.00	208	8182	26.77	0.9497	0.0587	0.0804	0.0553
Aj	7	4.90	194	8157	22.21	0.9515	0.0505	0.0785	0.0622
To	8	6.13	154	8160	20.50	0.9549	0.0495	0.0706	0.0491
Bl	9	7.62	327	8177	27.64	0.9486	0.0626	0.0808	0.0468
Ta	10	8.40	431	8178	28.10	0.9467	0.0594	0.0826	0.0580
Co	8	5.37	309	8141	24.49	0.9477	0.0508	0.0849	0.0777

Results include only nucleotide positions that are polymorphic in at least one population. The first column shows the number of individuals per population that were used for the analysis (Ns). Next are the average number of individuals genotyped at each locus (N), the number of variable sites unique to each population (i.e. private alleles, Priv.), the number of polymorphic nucleotide sites across the data set (Sites), percentage of polymorphic loci (% poly), the average frequency of the major allele (P), the average observed heterozygosity per locus (H_{obs}), the average nucleotide diversity (π), and the average Wright's inbreeding coefficient (F_{IS}). Populations are ordered East to West, top to bottom. Population IDs as in Fig. 5.1.

5.4.3. Alpine grassland distribution during glacial/interglacial stages

The uncorrelated environmental variables used for the timberline-alpine grassland modelling were isothermality, mean temperature annual range, temperature in the wettest quarter, precipitation seasonality and precipitation in the coldest quarter (Fig. 5.4a). The potential distribution found for the present is congruent with the known distribution of the alpine grasslands in the TMVB (mostly >3,500 masl), and the projection to the LGM shows that this ecosystem occurred in the same geographic areas, but with a larger distribution extending to relatively lower elevations (Fig. 5.4b).

5.4.4. Effective spatial isolation

The plots of the resistance surfaces (Fig. 5.3) show that although sampling points are separated by similar horizontal distances (except for Nevado de Colima and Tancítaro, far West of *J. monticola*'s distribution) there are important differences on the connectivity among points depending on the elevation or distribution model used to set the conductance values. In general, the sampled mountains start to be connected from 3,000 masl to lower altitudes. Central-West populations (Cerro Blanco, Nevado de Toluca and Ajusco), Central (Tláloc, Iztaccihuatl and Popocatepetl) and Eastern populations (Cofre de Perote, Citlatépetl and La Negra) are joined in the surfaces from the LGM models and, in the surfaces from the elevation data, from 2,700 masl to lower altitudes. Contrasting, La Malinche (second *B. alpina's* sampling point from East to West) remains isolated until a connectivity as low as 2,500 masl is allowed, and Nevado de Colima and Tancítaro (*J. monticola*'s Westernmost sampling points) are (relatively) connected only when setting as suitable altitudes as low as 1,800-

2,000 masl. At such low elevations the rest of the sampling localities are completely connected as in a flat surface.

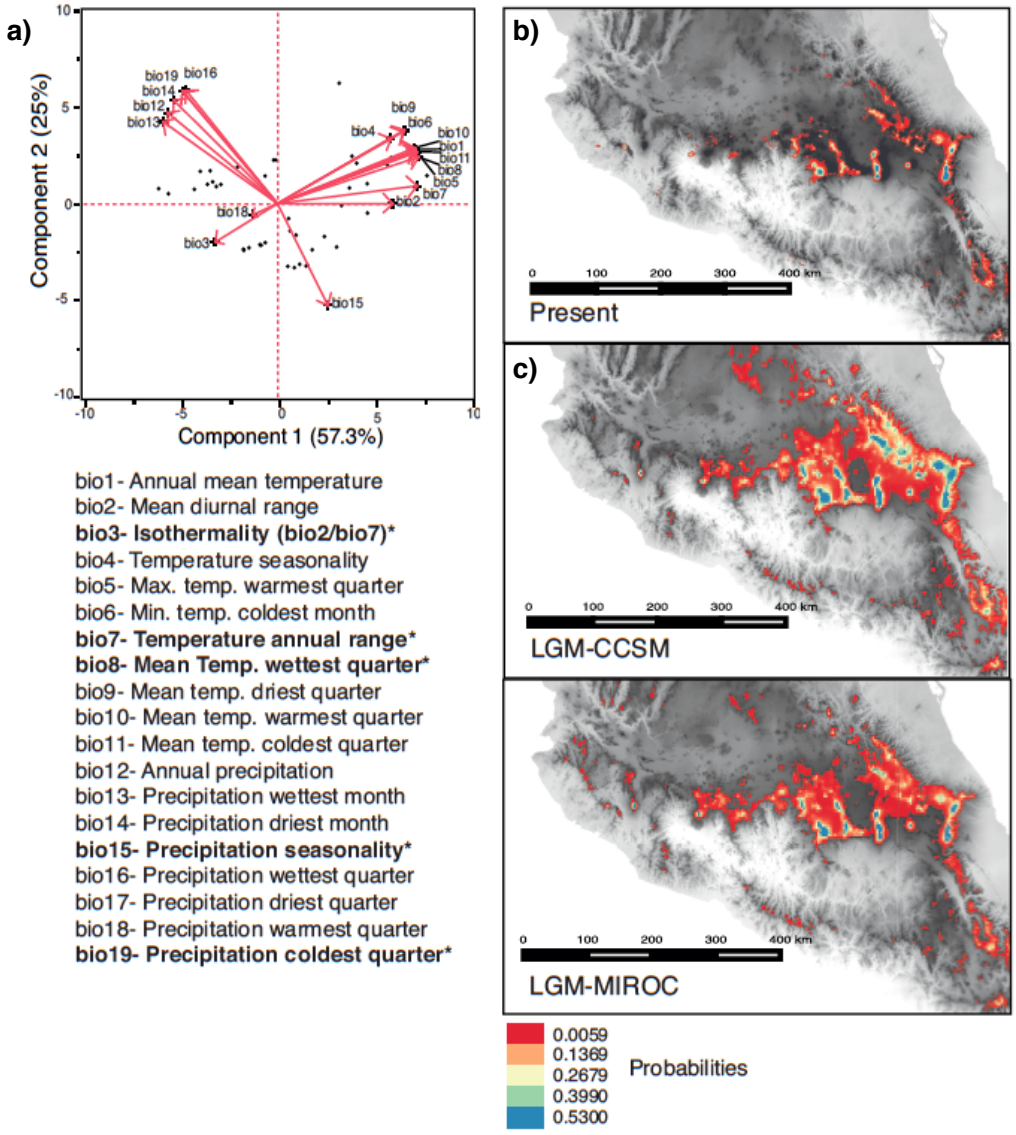


Figure 5.4. Environmental analyses and distribution models of the timberline - alpine grassland for interglacial and glacial conditions on the Trasmexican Volcanic Belt. (a) Principal component analysis of 19 bioclimatic variables. The independent variables with the highest contributions to variance were selected for the potential distribution models and are indicated with an asterisk. Potential distribution models of the alpine grassland for the present (b) and Last Glacial Maximum (c). Two sets of environmental layers were used for the projection to the LGM: CCSM and MIROC (details in the methods). The yellow to blue color gradient of *b* indicates areas where the alpine grasslands are known to occur in the present interglacial. Projections to the LGM show that this ecosystem likely occurred in the same mountains, but with a larger distribution extending to lower altitudes.

5.4.5. Isolation by resistance

Both the Mantel test and the linear regression yielded positive significant results for IBR for all resistance surfaces and species or groups of populations, but with different explanatory power depending on the surface used (Table 5.4). The 'flat' landscape (i.e. isolation by distance) was outperformed by some of the scenarios considering the environmental modelling or the elevation grids. The surface with the highest explanatory power varied between species and populations tested. For *B. alpina* the highest explanatory power was provided by the resistance surface of 3,000 masl (Mantel $r = 0.940$, $p < 0.001$; regression $r^2 = 0.883$, $p < 0.01$ Table 5.4). For *J. monticola,* considering all populations, the surface with the highest explanatory power was the flat surface (Mantel $r = 0.504$, $p < 0.01$; regression $r^2 = 0.148$, $p < 0.001$), and when excluding the populations of Nevado de Colima and Tancítaro, environmental modelling for the LGM using the CCSM layers provided the highest explanatory power (Mantel $r = 0.686$, $p < 0.001$; regression $r^2 = 0.465$, $p < 0.01$, Table 5.4).

5.4.6. Effect of historical isolation on private alleles

Testing for the effect of the mean effective isolation of each mountain on the number of private alleles yielded a significant effect for all taxa, but not for all elevation surfaces (Table 5.5). For those surfaces yielding a significant effect, the number or private alleles was found to increase as the mean effective distance of the mountain to the rest of the sampling localities increases. For *B. alpina* the 2,500 masl elevation model surface yielded the highest r^2 value ($r^2 = 0.857$, $p<0.01$), while for *J. montiocla* it was the surface of 3,000 masl, both when

considering all populations (r^2 = 0.387, p<0.05), or excluding the Tancítaro and Nevado de Colima populations (r^2 = 0.507, p<0.05, Table 5.5).

Table 5.4. Isolation by resistance

	B. alpina		*J. monticola* all pops.		*J. monticola* excluding Co & Ta	
Surface	***r***	***r²***	***r***	***r²***	***r***	***r²***
present	0.792**	0.620 ***	0.472 *	0.220 ***	0.662 ***	0.431***
ccsm	0.667*	0.439 **	0.404 *	0.161 ***	0.686 ***	0.465***
miroc	0.797**	0.627 ***	0.433 *	0.185 ***	0.675 ***	0.450***
flat	0.879**	0.776 ***	0.504 **	0.248 ***	0.579 ***	0.327***
1,800	0.883***	0.789 ***	0.330 NS	0.107 **	0.575 ***	0.325***
2,000	0.887**	0.797 ***	0.302 NS	0.090 **	0.566 ***	0.315***
2,300	0.821**	0.683 ***	0.322 NS	0.102 **	0.555 ***	0.303***
2,500	0.897**	0.811 ***	0.387 *	0.148 ***	0.530 **	0.275***
2,700	0.929**	0.862 ***	0.447 *	0.196 ***	0.550 **	0.296***
3,000	0.940**	0.883 ***	0.378 *	0.140 ***	0.331 NS	0.106*
3,300	0.904**	0.818 ***	0.394 *	0.151 ***	0.353 NS	0.120**
3,500	0.832*	0.693 ***	0.391 *	0.149 ***	0.340 NS	0.112*
4,000	0.680*	0.464 **	0.383 *	0.143 ***	0.335 NS	0.108*

Associations between genetic differentiation (F_{ST} or linearized F_{ST}, see main text) and pairwise effective distances at different surfaces. Mantel test *r* value (left column) and the r^2 of the linear regression (right column) are reported for each species. Significance codes are as follows: < 0.001 '***', <0.01 '**', < 0.05 '*', and not significant 'NS'. Underlined cells correspond to the surface with the highest prediction value for each taxon.

Table 5.5. Private alleles

Surface	***B. alpina***		***J. monticola* all pops.**		***J. monticola excluding* Co & Ta**	
present	0.342	NS	0.077	NS	0.001	NS
ccsm	0.248	NS	0.050	NS	0.217	NS
miroc	0.455	NS	0.042	NS	0.056	NS
flat	0.616	NS	0.082	NS	0.097	NS
1,800	0.651	NS	0.037	NS	0.026	NS
2,000	0.659	*	0.044	NS	0.010	NS
2,300	0.607	NS	0.062	NS	0.001	NS
2,500	0.857	**	0.081	NS	0.014	NS
2,700	0.719	*	0.130	NS	0.122	NS
3,000	0.679	*	0.387	*	0.507	*
3,300	0.708	*	0.380	*	0.489	*
3,500	0.663	*	0.365	*	0.407	*
4,000	0.668	*	0.211	NS	0.183	NS

Results show the r^2 from linear regressions of private alleles on mean effective distances at different surfaces for each species. Significance codes are as follows: <0.01 '**', < 0.05 '*', and not significant 'NS'. Underlined cells correspond to the surface with the highest prediction value.

5.5. Discussion

5.5.1. Local long-term persistence of alpine grasslands

As expected, the potential distribution of the timberline-alpine grassland matches the highest mountains of the TMVB (Fig. 5.4b). In general, the present modelling is congruent with the known distribution of the timberline-alpine grasslands in this region, but it may represent a slight overestimate because it is predicting suitable areas slightly below 3,000 masl, when strictly alpine taxa occur >3,900 masl (Lauer 1978; Calderón de Rzedowski & Rzedowski 2005), and a general decay of forest cover and grassland extension occurs not lower than 3,500 masl (Beaman 1962; Almeida-Leñero, L. *et al.* 2007). This overestimate may be due to the inclusion of a few presence points located in mountains that are too small for the resolution of the grid used, thus resembling conditions of lower elevation. However, at a regional scale the modelling matches the known distribution of this vegetation type (Rzedowski 1978; Calderón de Rzedowski & Rzedowski 2005).

The projection to the LGM shows that the timberline-alpine grasslands could have extended to lower elevations of the TMVB both under the CCSM and MIROC scenarios (Fig. 5.4c). This is congruent with fossil pollen suggesting the existence of reduced forests (similar to open forests close to timberline) and grasslands down to 2,300-2,500 masl (Lozano-García & Ortega-Guerrero 1994, 1998; Lozano-García *et al.* 2005) and with moraines showing that snow lines dropped around 1,000 m during the glacial periods (Lozano-García & Vázquez-Selem 2005; Vázquez-Selem & Heine 2011). For example, the mean altitude of the glacier terminus on Iztaccíhuatl volcano, today at above 4,700 masl, was at

3,390±160 masl during the LGM (Vázquez-Selem and Heine, 2011). The genetic data also supports a scenario of long-term population persistence in both species. Genomic differentiation was significant among all populations, with F_{ST} values typically greater than 0.05 (Table 5.2 and 3) and (2) all populations exhibited low frequency alleles (data not shown), as expected for old and stable populations, as opposed to lack of low frequency alleles expected after foundation events or bottlenecks (Hartl & Clark 2007).

Considered together, the palynological, geological and niche modelling data suggest that the LGM open forests and grasslands could have extended down to 2,300-2,500 masl at the LGM, and that suitable conditions for alpine vegetation (i.e. not covered with permanent ice but close to the glacial limit) could have existed up to 3,300 masl. If open forests and grasslands occurred at relatively low altitudes of the TMVB during the glacial maxima, they were likely replaced by other vegetation types (semi-desert scrublands to conifer forests) during the interglacial periods, similar to their present distribution. However, on mountains >3,000 masl, and particularly on the highest stratovolcanoes that reach >3,500 masl, environmental conditions suitable for alpine grasslands appear to have existed continuously over glacial and interglacial periods (Fig. 5.4b-c). This demonstrates that since their emergence during the last 1.5 Myr (Ferrari *et al.* 2012), the highest volcanoes of the TMVB have provided stable conditions throughout glacial-interglacial cycles suitable for continuous population persistence for subalpine and alpine taxa.

However, it is important to note that our modelling approach and the available palynological and geological data are not species specific. Each taxon may respond differently to subtle environmental differences or have different

tolerance thresholds (Araújo & Guisan 2006; Roberts & Hamann 2012) thus delimiting their distribution within the broader range of alpine grasslands. Even within the present distribution of alpine grasslands, some species occur only far above the timberline, while others can be found both in the grasslands and at the timberline transition (Calderón de Rzedowski & Rzedowski 2005). Nonetheless, broadly speaking, the present and past distributions of timberline-alpine taxa from the TMVB are highly dependent on temperature or temperature associated variables, which in turn are highly related to altitude (Beaman 1962; Lauer 1978; Almeida-Leñero, L. *et al.* 2007). Thus, it is expected that the altitude of the landscape separating the highest peaks of the TMVB would play a key role for population connectivity, or isolation, of species currently inhabiting the timberline-alpine grasslands of the TMVB.

5.5.2. Isolation by resistance in sky-islands of the TMVB

Testing for IBR with resistance surfaces using present and past potential habitat distributions shows that, as predicted, accounting for topography-driven connectivity better explains population differentiation than plain geographic distance. This is supported by some of the resistance surfaces having a higher explanatory power than the flat landscape (Table 5.4).

For *B. alpina,* a pattern of IBD was found to significantly explain population differentiation (Mantel $r = 0.879$ and regression $r^2 = 0.776$, Table 5.4). The resistance surfaces allowing low altitude connectivity (1,800-2,000 masl, Table 5.4), had a similar explanatory power to the flat landscape surface used for the IBD test. But interestingly, the explanatory power increased with altitude, reaching a maximum with the surface allowing for connectivity at 3,000 masl

(Mantel r = 0.940 and regression r^2 = 0.883, Table 5.4), and then decreasing (Table 5.4). This indicates that although simple geographic distance has explanatory power, more of the variance is explained if certain locality pairs are considered to be effectively less (or more) distant than others (Fig. 5.5).

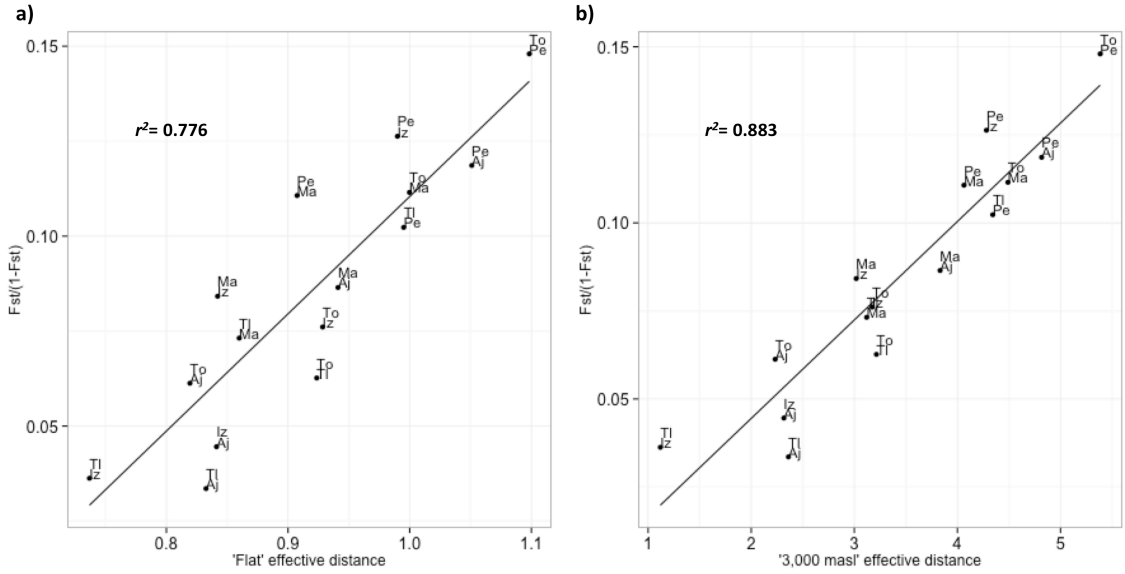


Figure 5.5. Test for (a) isolation by distance for *Berberis alpina* using the 'flat' surface ($F_{1,13}$ = 53.9, p < 0.0001) and for (b) isolation by resistance using the resistance surface that provided the highest explanatory power (elevation above 3,000 masl, $F_{1,13}$ = 97.8, p < 0.0001). Labels show populations of each pair-wise comparison. Codes as in Fig. 5.1.

For *J. monticola* a pattern of IBD was also found (Fig. 5.6a), but it did not explain a high amount of the variance (Mantel r = 0.504 and regression r^2=0.248, for the flat landscape, Table 5.4). However, when Nevado de Colima and Tancítaro populations were excluded from the analysis, the explanatory power of the 'flat' landscape increased (from r^2 = 0.248 to 0.327, Table 5.4) and instead of IBD explaining more of the variance, IBR with the LGM-CCSM surface held more explanatory power (r^2 = 0.465, Table 5.4). The Nevado de Colima and Tancítaro mountains are considerably further away from the remaining high mountains of the TMVB (Fig. 5.1), and in areas that have not been connected by

alpine grasslands to the Central TMVB in the Pleistocene glaciations (they remain isolated in both LGM models and when allowing connectivity in altitudes as low as 2,300 masl Fig. 5.3). It thus seems more likely that these populations are the product of long distance colonisation, and would not be under a climate mediated regime of gene flow with other populations.

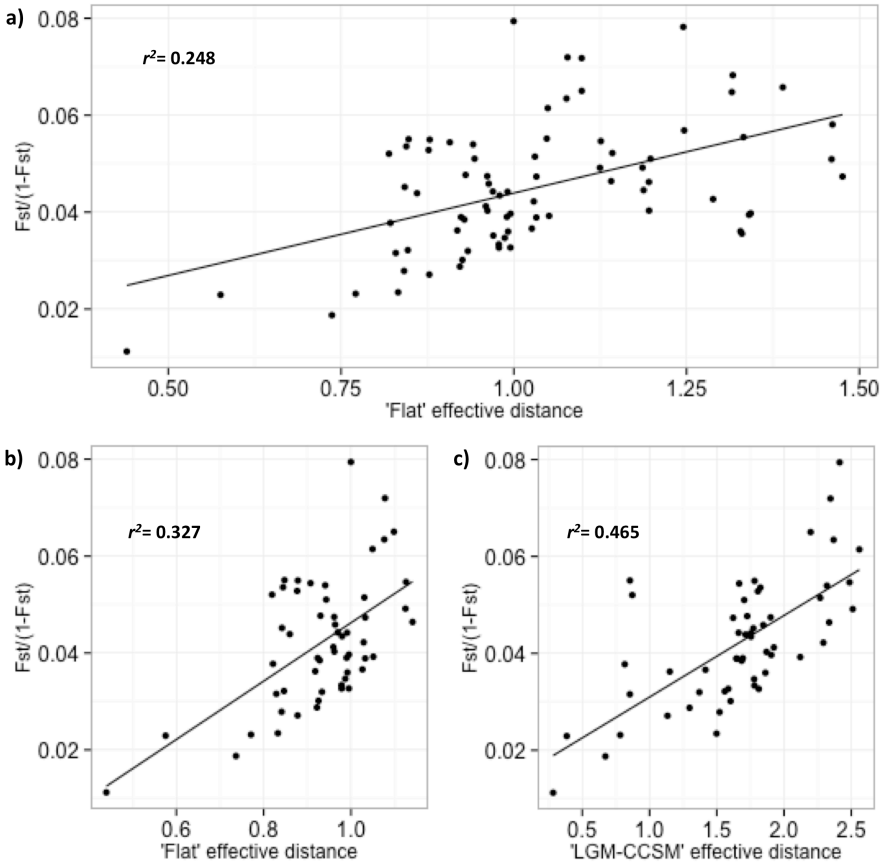


Figure 5.6. Test for isolation by distance for *Juniperus monticola* using the 'flat' surface and (a) all populations ($F_{1,76}$=16.9, p < 0.0001) or (b) excluding the Tancítaro and Nevado de Colima populations ($F_{1,53}$=17.6, p < 0.001). (c) Test for isolation by resistance using the surface that provided the highest explanatory power when excluding the populations Tancítaro and Nevado de Colima (SDM with LGM-CCSM conditions, $F_{1,53}$=46, p < 0.0001).

Compared to *B. alpina,* less of the variance could be explained by historical connectivity for *J. monticola*, even when removing the Nevado de Colima and Tancítaro populations (Mantel *r* = 0.940 for *B. alpina vs* 0.686 for the

juniper; and highest regression r^2 = 0.883 for *B. alpina*, *vs* 0.465 for the juniper Table 5.4). To examine the unexplained variance within *J. monticola*, it could be possible to test for the effect of local environmental differences, or for the role of mountain age, for instance using models of isolation by environment, isolation by colonisation and multivariate analyses (Orsini *et al.* 2013; Wang 2013). However, despite the lower predictive power of IBR for *J. monticola* relative to *B. alpina,* results are consistent with population differentiation among TMVB's subalpine taxa being influenced by the landscape surrounding the mountain peaks. Interestingly, some resistance surfaces performed better than others which, as discussed below, can be used to examine whether present or historical connectivity better explain patterns of genetic diversity.

5.5.3. Population differentiation under a sky-island dynamic

Under a sky-island dynamic, montane species inhabiting tropical mountains are expected to (i) have been restricted to high-elevation refugia during the interglacial periods of the Pleistocene, where divergence could be promoted by restricted gene flow; and (ii) to have extended ephemerally to lowlands during glacial periods, where the probability of genetic admixture would be increased (Toledo 1982). For *J. monticola* and *B. alpina*, population differentiation was tested against different interglacial and glacial landscape connectivity scenarios. This allows for an evaluation of which scenario provides a better explanation for the distribution of genetic diversity, similar to tests of which landscape features influence population structure (e.g. McRae *et al.* 2008; Moore *et al.* 2011).

Interestingly, the population genetic differentiation of both species was better explained by resistance surfaces (3,000 masl and LGM-CCSM for *B. alpina*

and *J. monticola*, respectively, Table 5.4) occupying areas ~1,000 m below the elevation where the species are more abundant in their current altitudinal ranges. This fits with the prediction of gene flow occurring during glacial periods, and seems to indicate that historical population connectivity has played a more important role than current isolation for population differentiation. This result is not surprising when considering that: (1) the timberline attained its present altitude only 3,000 yr ago (Lozano-García & Vázquez-Selem 2005); (2) the last 700,000 yr have been dominated by major glacial periods with a ~100,000 yr cycle interrupted by relatively short warm interglacials (Webb & Bartlein 1992); so that (3) resent distributions could be considered a perturbation of the "historical average", and (4) that these species are slow growing and live decades or hundreds of years (Francis 2004; Adams 2008), so that the number of generations living in the present distribution could be relatively small. Also, studies in other montane areas within biodiversity hotspots, have also shown that historical measures of population connectivity among stable areas are correlated with gene flow estimates (Devitt *et al.* 2013).

It is important to note that during the glacial scenarios with the highest explanatory power, species seem to have had a fragmented (island-like) distribution (Fig. 5.3): suitable glacial conditions in the lowlands can connect some of the currently isolated populations, but not all (e.g. Tláloc and Iztaccíhuatl-Popocatepetl are joined at 3,000 masl and below, whilst La Malinche remains isolated even at 2,500 masl and in the LGM modelling Fig. 5.3). In other words, for the two subalpine species examined here, glacial admixture could occur more readily among certain population clusters (e.g. Tláloc-Iztaccihuatl-

Popocatepetl, Fig. 5.3), while other populations would remain similarly isolated as during the interglacial stages.

5.5.4. Population differentiation and genetic endemism on tropical mountains

Population differentiation is highly explained by the pairwise effective distance among populations during the glacial periods. It was expected, therefore, that the number of private alleles of each mountain would be positively related to the relative isolation of each locality, measured as the mean effective distance of each mountain to the rest. However this expectation was not met by our data. Although a significant and positive relationship was found when using some the elevation surfaces (Table 5.5), this is largely driven by the effect of the Cofre de Perote and El Chico populations acting as outliers for *B. alpina* and *J. monticola,* respectively (Fig. 5.7b and d). If these outlier populations are removed, there remain too few points for the analysis, or the pattern is significantly lost. Explaining genetic endemism remains thus an open question.

Further analyses examining the number of private alleles per population could explore alternative measurements of isolation and the genetic history of the private alleles formation. For instance isolation could be quantified using the mean effective distance in a given radius, rather than against the entire range of the species, or grouping sampling points according to population structuring analyses. Also, the genetic history of the private alleles could be examined in a more detailed way to evaluate their ancestry to shared alleles. This could be particularly relevant for examining if the outlier populations have an excess of 'old' private alleles, which could be expected if time since isolation is playing a role on the accumulation of private alleles.

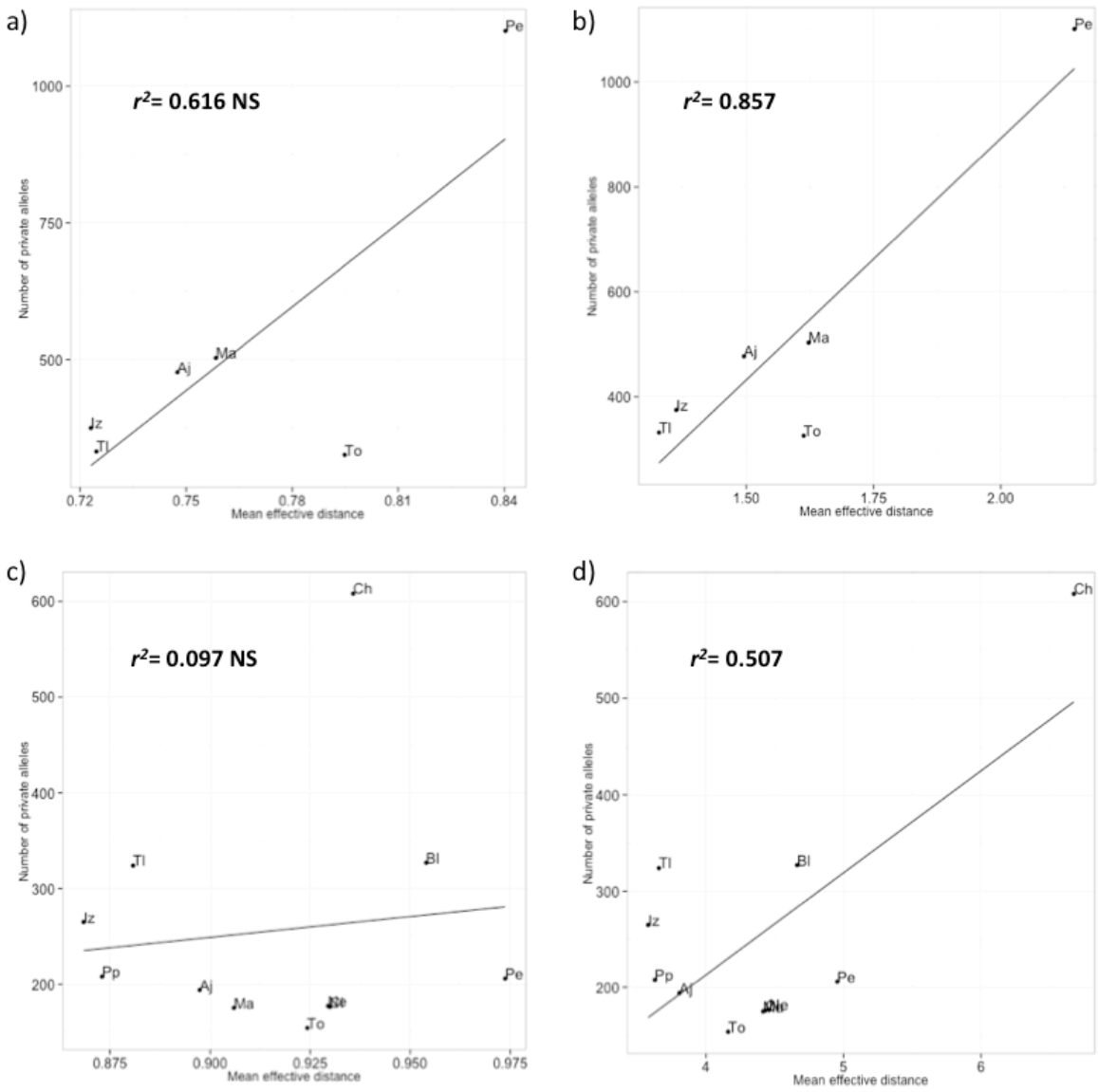


Figure 5.7. Test of the number of private alleles as a function of the mean effective distance for *B. alpina* using (a) the flat landscape surface ($F_{1,4}$=6.42, p = 0.064), (b) the surface that provided the highest explanatory power (elevation above 2,500 masl, $F_{1,4}$=24, p < 0.01). Results of the same test for *J. monticola* using (c) the flat landscape surface ($F_{1,9}$=0.115, p = 0.742) and the surface that provided the highest explanatory power (elevation above 3,000 masl, $F_{1,9}$=91, p < 0.05). In both cases plots show analyses when excluding populations Nevado de Colima and Tancítaro.

Regardless of the unexplained excess of private alleles, we have shown that: (1) the highest stratovolcanoes of the TMVB facilitated the existence of timberline-alpine grasslands throughout glacial/interglacial cycles; and (2) population genetic differentiation of species from this ecosystem can be

explained by the degree of habitat connectivity among mountains during the glacial periods. Similar conclusions have been postulated for taxa of the TMVB of slightly lower altitudes using more traditional population genetic and phylogeographic approaches (e.g. McCormack *et al.* 2008; Bryson *et al.* 2011, 2012; Gutiérrez-Rodríguez *et al.* 2011; Ornelas *et al.* 2013). Additionally, the role of topography as a barrier to present dispersal has been examined for a lizard species (Parra-Olea *et al.* 2012). However, to our knowledge this is the first time that present vs past historical connectivity are assessed in a landscape explicit and quantitative way for this region. An emergent advantage of this approach is that we can relate population differentiation to the Pleistocene glacial cycles despite not having used the molecular data for population divergence dating. This is relevant because in the TMVB climate fluctuations and volcanic changes co-occurred during the Pleistocene (see Chapter 2), and previous phylogeographic studies focusing on divergence times (e.g. Ornelas *et al.* 2010; Bryson *et al.* 2012a; b; Leaché *et al.* 2013) have not been able to distinguish between the confounding effect of climate and geological change.

In conclusion, our findings from a population-level perspective indicate that tropical mountains: (1) allow for long-term *in situ* population persistence throughout periods of climate fluctuation; and (2) promote population differentiation as a function of topographic isolation. This highlights that the importance of this region for conservation resides not only on its species richness *per se*, but on that specific areas promote long-term survival and further diversification.

5.6. Acknowledgments

We acknowledge the international modelling groups for providing their data for analysis, the Laboratoire des Sciences du Climat et de l'Environnement (LSCE) for collecting and archiving the model data. The PMIP 2 Data Archive is supported by CEA, CNRS and the Programme National d'Etude de la Dynamique du Climat (PNEDC). The analyses were performed using version 01/20/10 of the database. More information is available on http://pmip2.lsce.ipsl.fr/. Part of the analyses were carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at UEA.

5.7. References

Adams RP (2008) *Junipers of the world: The genus Juniperus*. Trafford Publishing.

Adams RP, Schwarzbach AE, Morris JA (2010) *Juniperus zanonii,* a new species form Cerro Potosi, Nuevo Leon, Mexico. *Phytologia*, **92**, 105–117.

Almeida-Leñero, L., Escamilla, M., Giménez de Azcárate, J., González-Trápaga, A., Cleff, A. M. (2007) Vegetación alpina de los volcanes Popocatépetl, Iztaccíhuatl y Nevado de Toluca. In: *Biodiversidad de la faja volcánica transmexicana* (eds Luna-Vega I, Morrone JJ, Espinosa D), pp. 179–198. Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Zaragoza e Instituto de Biología.

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Beaman JH (1962) The Timberlines of Iztaccihuatl and Popocatepetl, Mexico. *Ecology*, **43**, 377–385.

Braconnot P, Otto-Bliesner B, Harrison S *et al.* (2007) Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features. *Clim. Past*, **3**, 261–277.

Bryson RW, García-Vázquez UO, Riddle BR (2012a) Relative roles of Neogene vicariance and Quaternary climate change on the historical diversification of bunchgrass lizards

(Sceloporus scalaris group) in Mexico. *Molecular Phylogenetics and Evolution*, **62**, 447–457.

Bryson RW, García-Vázquez UO, Riddle BR (2012b) Diversification in the Mexican horned lizard Phrynosoma orbiculare across a dynamic landscape. *Molecular Phylogenetics and Evolution*, **62**, 87–96.

Bryson RW, Murphy RW, Lathrop A, Lazcano-Villareal D (2011) Evolutionary drivers of phylogeographical diversity in the highlands of Mexico: a case study of the Crotalus triseriatus species group of montane rattlesnakes. *Journal of Biogeography*, **38**, 697–710.

Cadena CD, Kozak KH, Gómez JP *et al.* (2011) Latitude, elevational climatic zonation and speciation in New World vertebrates. *Proceedings of the Royal Society B: Biological Sciences*, rspb20110720.

Calderón de Rzedowski G, Rzedowski J (2005) *Flora Fanerogámica del Valle de México*. Instituto de Ecología A.C. y Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Pátzcuaro, Michoacán, México.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009) Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785 –789.

Catchen J, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Devitt TJ, Devitt SEC, Hollingsworth BD, McGuire JA, Moritz C (2013) Montane refugia predict population genetic structure in the Large-blotched Ensatina salamander. *Molecular Ecology*, **22**, 1650–1665.

Ferrari L, Orozco-Esquivel T, Manea V, Manea M (2012) The dynamic history of the Trans-Mexican Volcanic Belt and the Mexico subduction zone. *Tectonophysics*, **522–523**, 122–149.

Fjeldså J, Bowie RCK (2008) New perspectives on the origin and diversification of Africa's forest avifauna. *African Journal of Ecology*, **46**, 235–247.

Fjeldså J, Bowie RCK, Rahbek C (2012) The Role of Mountain Ranges in the Diversification of Birds. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 249–265.

Fjeldså J, Lambin E, Mertens B (1999) Correlation between endemism and local ecoclimatic stability documented by comparing Andean bird distributions and remotely sensed land surface data. *Ecography*, **22**, 63–78.

Francis JK (2004) Mahonia aquifolium (Pursh) Nutt. *Wildland shrubs of the United States and its territories: thamnic descriptions*, **1**, 461.

Graham CH, Moritz C, Williams SE (2006) Habitat history improves prediction of biodiversity in rainforest fauna. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 632–636.

Gutiérrez-Rodríguez C, Ornelas JF, Rodríguez-Gómez F (2011) Chloroplast DNA phylogeography of a distylous shrub (Palicourea padifolia, Rubiaceae) reveals past fragmentation and demographic expansion in Mexican cloud forests. *Molecular Phylogenetics and Evolution*, **61**, 603–615.

Hartl DL, Clark AG (2007) *Principles of population genetics*. Sinauer Associates, Sunderland, MA, EEUUA.

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

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**.

Hijmans RJ, Etten J van, Mattiuzzi M *et al.* (2014) *raster: Geographic data analysis and modeling*. v. 2.3-12. CRAN repository.

Hugall A, Moritz C, Moussalli A, Stanisic J (2002) Reconciling paleodistribution models and comparative phylogeography in the Wet Tropics rainforest land snail Gnarosophia bellendenkerensis (Brazier 1875). *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6112 –6117.

Jetz W, Rahbek C (2002) Geographic range size and determinants of avian species richness. *Science*, **297**, 1548 –1551.

Jetz W, Rahbek C, Colwell RK (2004) The coincidence of rarity and richness and the potential signature of history in centres of endemism. *Ecology Letters*, **7**, 1180–1191.

Kessler M (2002) The elevational gradient of Andean plant endemism: varying influences of taxon-specific traits and topography at different taxonomic levels. *Journal of Biogeography*, **29**, 1159–1165.

Krömer T, Acebey A, Kluge J, Kessler M (2013) Effects of altitude and climate in determining elevational plant species richness patterns: A case study from Los Tuxtlas, Mexico. *Flora - Morphology, Distribution, Functional Ecology of Plants*, **208**, 197–210.

Kruckeberg AR, Rabinowitz D (1985) Biological aspects of endemism in higher plants. *Annual Review of Ecology and Systematics*, **16**, 447–479.

Lauer W (1978) Timberline studies in Central Mexico. *Arctic and Alpine Research*, **10**, 383.

Leaché AD, Palacios JA, Minin VN, Bryson RW (2013) Phylogeography of the Trans-Volcanic bunchgrass lizard (*Sceloporus bicanthalis*) across the highlands of south-eastern Mexico. *Biological Journal of the Linnean Society*, **110**, 852–865.

Lee-Yaw JA, Davidson A, Mcrae BH, Green DM (2009) Do landscape processes predict phylogeographic patterns in the wood frog? *Molecular Ecology*, **18**, 1863–1874.

Loarie SR, Duffy PB, Hamilton H *et al.* (2009) The velocity of climate change. *Nature*, **462**, 1052–1055.

Lozano-García MS, Ortega-Guerrero B (1994) Palynological and magnetic susceptibility records of Lake Chalco, central Mexico. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **109**, 177–191.

Lozano-García MS, Ortega-Guerrero B (1998) Late Quaternary environmental changes of the central part of the Basin of Mexico; correlation between Texcoco and Chalco basins. *Review of Palaeobotany and Palynology*, **99**, 77–93.

Lozano-García S, Sosa-Nájera S, Sugiura Y, Caballero M (2005) 23,000 yr of vegetation history of the Upper Lerma, a tropical high-altitude basin in Central Mexico. *Quaternary Research*, **64**, 70–82.

Lozano-García S, Vázquez-Selem L (2005) A high-elevation Holocene pollen record from Iztaccihuatl volcano, central Mexico. *The Holocene*, **15**, 329–338.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014a) Data from: RAD sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Dryad Digital Repository.*, doi:10.5061/dryad.g52m3.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014b) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, doi:10.1111/1755-0998.12291

Mastretta-Yanes A, Zamudio S, Jorgensen TH *et al.* (2014c) Gene duplication, population genomics and species-level differentiation within a tropical mountain shrub. *Genome Biology and Evolution*, evu205.

McCormack JE, Peterson AT, Bonaccorso E, Smith TB (2008) Speciation in the highlands of Mexico: genetic and phenotypic divergence in the Mexican jay (*Aphelocoma ultramarina*). *Molecular Ecology*, **17**, 2505–2521.

McRae BH (2006) Isolation by resistance. *Evolution*, **60**, 1551.

McRae BH, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences*, **104**, 19885 –19890.

McRae BH, Dickson BG, Keitt TH, Shah VB (2008) Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, **89**, 2712–2724.

Moore JA, Tallmon DA, Nielsen J, Pyare S (2011) Effects of the landscape on boreal toad gene flow: does the pattern–process relationship hold true across distinct landscapes at the northern range margin? *Molecular Ecology*, **20**, 4858–4869.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Ornelas JF, Ruiz-Sánchez E, Sosa V (2010) Phylogeography of *Podocarpus matudae* (Podocarpaceae): pre-Quaternary relicts in northern Mesoamerican cloud forests. *Journal of Biogeography*, **37**, 2384–2396.

Ornelas JF, Sosa V, Soltis DE *et al.* (2013) Comparative phylogeographic analyses illustrate the complex evolutionary history of threatened cloud forests of Northern Mesoamerica. *PLoS ONE*, **8**, e56283.

Orsini L, Vanoverbeke J, Swillen I, Mergeay J, De Meester L (2013) Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Molecular Ecology*, **22**, 5983-5999.

Päckert M, Martens J, Sun Y-H *et al.* (2012) Horizontal and elevational phylogeographic patterns of Himalayan and Southeast Asian forest passerines (Aves: Passeriformes). *Journal of Biogeography*, **39**, 556–573.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.

Parchman TL, Gompert Z, Mudge J *et al.* (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

Parra-Olea G, Windfield JC, Velo-Antón G, Zamudio KR (2012) Isolation in habitat refugia promotes rapid diversification in a montane tropical salamander. *Journal of Biogeography*, **39**, 353–370.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) double digest radseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

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

Rahbek C, Gotelli NJ, Colwell RK *et al.* (2007) Predicting continental-scale patterns of bird species richness with spatially explicit models. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 165–174.

R. Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberts DR, Hamann A (2012) Method selection for species distribution modelling: are temporally or spatially independent evaluations necessary? *Ecography*, **35**, 792–802.

Rounsaville TJ, Ranney TG (2010) Ploidy levels and genome sizes of *Berberis* L. and *Mahonia* nutt. species, hybrids, and cultivars. *HortScience*, **45**, 1029–1033.

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.

Rzedowski J (1978) *Vegetación de México*. Limusa, México.

Sandel B, Arge L, Dalsgaard B *et al.* (2011) The influence of late Quaternary climate-change velocity on species endemism. *Science*, **334**, 660 –664.

Smith BT, McCormack JE, Cuervo AM *et al.* (2014) The drivers of tropical speciation. *Nature*, doi: 10.1038/nature13687

Toledo V (1982) Pleistocene changes of vegetation in tropical Mexico. In: *Biological diversification in the tropics* (ed Prance GT), pp. 93–111. Columbia University Press, New York.

Vázquez-Selem L, Heine K (2011) Late Quaternary Glaciation in Mexico. In: *Quaternary Glaciations - Extent and Chronology - A Closer Look* (eds Ehlers J, Gibbard PL, Hughes P), pp. 849–861. Elsevier.

Wang IJ (2013) Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation. *Evolution*, **67**, 3403–3411.

Webb T, Bartlein PJ (1992) Global changes during the last 3 million years: climatic controls and biotic responses. *Annual Review of Ecology and Systematics*, **23**, 141–173.

Weir JT (2009) Implications of genetic differentiation in Neotropical montane forest birds. *Annals of the Missouri Botanical Garden*, **96**, 410–433.

Wickham H, Chang W (2013) *ggplot2: An implementation of the Grammar of Graphics*. v. 0.9.3.1. CRAN repository.

Zonneveld BJM (2012) Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nordic Journal of Botany*.

CHAPTER 6

General discussion and conclusions

Marty, it's perfect, you're just not thinking fourth-dimensionally!
-The Doc on *Back to the Future II*

Telescope, ticatla 2012

Studying natural populations with molecular tools has had a dramatic influence on our comprehension of life on Earth: biodiversity distribution changes in space and time (which we knew from the fossil record), and as a consequence species become genetically structured, with modern populations still reflecting the effect of historical events in their genomes (Avise 1994; Hewitt 1996, 2004). This became the established wisdom over the first decades of performing phylogeographic analyses using DNA sequences and other traditional molecular markers (Avise 2009; Hickerson *et al.* 2010). Over the last five years we have seen how genomic data can be feasibly obtained for non-model organisms and large sample sizes (Davey *et al.* 2011; Narum *et al.* 2013; Seehausen *et al.* 2014). This has the potential to accelerate the fields of molecular ecology and biodiversity genetics, and to more fully address key questions and open new lines of investigation as to how speciation occurs (Seehausen *et al.* 2014). The present thesis spans the transition years between phylogeographic studies being restricted to low resolution molecular markers, and new methods that facilitate the generation of orders of magnitude more data (Davey & Blaxter 2010; McCormack *et al.* 2013). As such, this thesis focuses on two main points. Firstly, on the methodological aspects of utilising a genotyping-by-sequencing method (double digest RAD-seq, ddRAD) for individual-based population genetics and phylogeography of non-model plant species. Secondly, on applying the obtained data to examine one of the classic, but as yet not fully explained, patterns of biodiversity distribution: the biodiversity excess within tropical mountains.

6.1. Genotyping-by-sequencing for individual-based genomic analyses

Obtaining large amounts of genomic data from non-model species became possible because: (1) parallel sequencing technologies have become cheaper, (2) molecular techniques were developed for subsampling genomes at homologous locations, and (3) bioinformatic tools were developed for the assembly and analysis of short sequencing reads (Davey *et al.* 2011). In this way, it is now possible to sequence thousands of loci for hundreds of individuals, rapidly and at low cost, regardless of genome size and previous genomic knowledge. As a result, studies such as the ones presented in this thesis, can move from analyses using a limited number of informative SNPs in plastid loci (Chapter 1), to examining population differentiation using thousands of SNPs (Chapter 5). However, as I discussed in Chapter 3, ddRAD and similar methods are in their early adolescence at the most, and methodological improvements are still on their way. The effect of missing data (Huang & Knowles 2014), bias on the genome regions being recovered (e.g. Arnold *et al.* 2013; DaCosta & Sorenson 2014), handling of PCR duplicates (Tin *et al.* 2014) and other special features of RAD data (Davey *et al.* 2013) are examples of how the RAD laboratory and quality-filtering resources continue to develop. To this end, Chapter 3 (published as Mastretta-Yanes *et al.* 2014), represents a contribution by (1) drawing attention to the need for genotyping error estimation, (2) proposing a method to do so with DNA replicates, and (3) further using the replicates to aid *de novo* assembly, by minimizing error and maximizing the retrieval of informative loci.

It is also possible to see how genotyping-by-sequencing methods are rapidly being improved because: (1) bioinformatic tools, such as *Stacks* (Catchen

et al. 2011, 2013), are being updated continuously, and (2) new analytical approaches continue to be published. For instance, approximate Bayesian computation (Robinson *et al.* 2014) and statistical methods powerful enough to discriminate among recent, non-equilibrium histories (Hearn *et al.* 2013) have now been developed to analyse RAD data. Thus, it is likely that in the next few years genotyping-by-sequencing methodology will reach a more mature stage and will become common practice among molecular ecology research groups. However, it has been suggested that genotyping-by-sequencing and similar methods will quickly become obsolete because undertaking whole-genome sequencing will be more feasible than what it is currently (Slavov *et al.* 2012; Seehausen *et al.* 2014). This may indeed be the case for (i) taxa with small and uncomplicated genomes, (ii) economically important species, or (iii) taxa closely related to species with an available reference genome. But biodiversity is vast, and possibly most evolutionary and ecological questions can be addressed without the need for whole genome data. Therefore, unless whole genome sequencing becomes a cheaper (and bioinformatically straight forward) option than reduced genome sequencing, it is likely that genotyping-by-sequencing will remain the molecular method of choice for: (1) taxa with large or completely unexplored genomes; and (2) for studies interested in population level variation across tens or hundreds of individuals. However, independently of how and how much genomic data is acquired, what will stand as truly important will be the evolutionary questions being asked.

6.2. Landscape genomics of tropical mountains: from evolutionary questions to conservation implications

The central aim of this thesis was to address, from a microevolutionary perspective, a long-standing question in biodiversity distribution: why are tropical mountains so species-rich? In Chapter 2 I reviewed the physical and phylogeographical history of the Mexican highlands as a way of introducing my study system - the timberline-alpine grasslands of the Transmexican Volcanic Belt (TMVB). These tropical mountains are a biodiversity hotspot surrounded by some of the most populated metropolitan areas of the world. They are also an interesting setting to test landscape genetic hypotheses, because they are an archipelago of sky-islands longitudinally distributed around the same tropical latitude (19ºN). The Quaternary origin of its highest stratovolcanoes (Ferrari *et al.* 2012) is normally considered to complicate the interpretation of phylogeographic patterns (e.g. Bryson *et al.* 2012a). This is because these newly arisen stratovolcanoes make it difficult to interpret if divergence times are due to topographic or climatic changes. However, landscape analyses allow for the incorporation of spatially explicit hypotheses on the effect of glacial cycles and volcanism, thus ameliorating this confounding effect. Also of importance, assuming niche conservatism (as found in McCormack *et al.* 2010), the recent origin of the highest stratovolcanoes provides an 'age limit' for the TMVB alpine grasslands. This is important because for species of lower altitudes (e.g. Bryson *et al.* 2012b; Parra-Olea *et al.* 2012), and for taxa of other tropical regions (e.g. Fjeldså & Bowie 2008; Smith *et al.* 2014), divergence times among sampling sites that are too deep (e.g. 5-11 Myr) do not allow for the explicit testing of landscape

as a driver of diversification. As detailed in Chapter 5, explaining the 'biodiversity excess' of tropical mountains has shifted from an entirely macroecological perspective (e.g. Kessler 2002; Kluge *et al.* 2006) to analyses looking to integrate historical evolutionary variables (Fjeldså & Bowie 2008; Smith *et al.* 2014). As a result, the effect of tropical mountains promoting long-term persistence of populations as well as diversification has emerged as a crucial factor to explain why these areas are biodiversity hot spots (Fjeldså *et al.* 2012). Chapter 5, contributes to closing a knowledge gap within the micro-macroevolutionary spectrum on which these processes are expected to occur. It does so by providing empirical evidence of the joint effect of long-term population persistence and population differentiation by isolation. Specifically, by using ddRAD data for alpine plants and landscape explicit analyses, I showed that (1) the TMVB has the physical characteristics to allow for glacial/interglacial *in situ* persistence of alpine grasslands, and (2) that the shape and altitude of the landscape surrounding the highest stratovolcanoes promotes population differentiation by restricting gene flow (even during the glacial stages) among island-like areas of suitable habitat.

Explaining the origin of montane biodiversity and endemism is important for the understanding of biodiversity itself, but it also has a relevant conservation consequence: protecting species of tropical mountains is not only guarding the currently observed taxa, but also ensuring the existence of biodiversity in areas where stable environmental conditions and further diversification are more likely to occur (Kruckeberg & Rabinowitz 1985; Fjeldså *et al.* 2012). To this statement I would add that, for the same reasons, it is necessary to promote conservation not only at the species, but also at the genetic

level. This is actually one of the agreements of the Convention on Biological Diversity of which Mexico forms part (United Nations 1992). In this way, allowing for the continuity of the evolutionary processes instead of protecting current species as steady entities, has become an important task for conservation biology (Crandall *et al.* 2000). The ways of achieving this include: (1) defining and integrating evolutionary-significant units into conservation targets (Moritz 1994); (2) creating natural protected areas enclosing regions that served as refugia during the Pleistocene glaciations (Avise 2008) and, more recently; (3) calling for special protection to long-term climatically stable regions within biodiversity hotspots (Carnaval *et al.* 2009, 2014; Fjeldså *et al.* 2012).

6.3. Future research

This thesis perhaps raises, or opens the path to, more questions than it has answered. Lines of further research include additional genetic and spatial analyses of the ddRAD data as well as an examination of the relationship between mountains more prone to population differentiation and regional peaks of species diversity. Firstly, the genetic point is perhaps the aspect offering richer immediate opportunities. Using these same datasets, or along with other published genomic resources and new analytical tools, it should be possible to: (i) undertake analyses regarding diversification times and the history of private alleles; (ii) examine patterns of speciation *and* gene flow or speciation *with* gene flow; (iii) examine if putative paralogous loci had a driving role on differentiation; and (iv) in general perform finer-tuned analyses focusing on the recovered loci, for example looking for signals of selection. Among these options,

I consider the further examination of paralogous loci particularly interesting. As I discussed in Chapter 4 (now published as Mastretta-Yanes *et al.* 2014b), genotyping-by-sequencing methods open the possibility of exploring genomic differentiation in contrast to the classical study of orthologous loci. Gene duplication has similar (or faster) rates than point mutations (Lynch & Conery 2000; Lynch 2002), and can generate ecological relevant variation (Moore & Purugganan 2005; Warren *et al.* 2014) as well as post-zygotic barriers to gene flow (Bikard *et al.* 2009). Exploring phenomena like these could be the true value of gathering genomic data from natural populations, instead of only producing more loci to perform classic population genetic analyses with more data. Secondly, from the spatial perspective, there remain sources of information to be explored. For example, a geological map of the TMVB could be incorporated into the analyses, as well as the alternative measures of topographic isolation that I discussed at the end of Chapter 5. Also related to landscape analyses, it is possible to test for isolation by environment (Wang & Bradburd 2014), or to evaluate topographic and environmental variables in a joint multivariate analyses (Wang 2013). Finally, it would also be interesting to examine whether sky-islands biodiversity accumulates following the predictions of neutral theory in macroecology (Hubbell 2001), similarly to studies in European mountains (Taberlet *et al.* 2012; Abellán & Svenning 2014) and islands (Papadopoulou *et al.* 2011). Such an analysis would test whether biodiversity is essentially structured as a fractal and thus that stochastic processes (migration, genetic/ecological drift, and mutation/speciation) act in an analogous way at all taxonomic scales, down to the level of haplotypes (Vellend 2003; Vellend & Geber 2005). Testing this for the TMVB may soon become possible because detailed species lists

(Steinmann *et al.* in prep) and further phylogeographic data (Uscanga *et al.*, in prep) of subalpine–alpine taxa are being performed with a geographical sampling useful for comparative analyses.

As final words, it is clear that studying natural populations with the joint analysis of genomic, landscape, evolutionary and ecological processes is the way forward to understand biodiversity and diversification in space and time. Much has been learned in the last two decades. This thesis represents a contribution to the understanding of the Mexican highlands, and of tropical mountains in general, as well as a methodological resource for further molecular ecology studies. But to me, the intricate interactions of genomes, environment, gene flow and isolation on generating biodiversity remains a gate open for wonder.

6.4. References

Abellán P, Svenning J-C (2014) Refugia within refugia – patterns in endemism and genetic divergence are linked to Late Quaternary climate stability in the Iberian Peninsula. *Biological Journal of the Linnean Society*, **113**, 13–28.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, advance online publication.

Avise JC (1994) *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.

Avise JC (2008) Three ambitious (and rather unorthodox) assignments for the field of biodiversity genetics. *Proceedings of the National Academy of Sciences*, **105**, 11564–11570.

Avise JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography*, **36**, 3–15.

Bikard D, Patel D, Metté CL *et al.* (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*, **323**, 623–626.

Bryson RW, García-Vázquez UO, Riddle BR (2012a) Relative roles of Neogene vicariance and Quaternary climate change on the historical diversification of bunchgrass lizards (*Sceloporus scalaris* group) in Mexico. *Molecular Phylogenetics and Evolution*, **62**, 447–457.

Bryson RW, García-Vázquez UO, Riddle BR (2012b) Diversification in the Mexican horned lizard *Phrynosoma orbiculare* across a dynamic landscape. *Molecular Phylogenetics and Evolution*, **62**, 87–96.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009) Stability predicts genetic diversity in the brazilian atlantic forest hotspot. *Science*, **323**, 785 –789.

Catchen J, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Crandall KA, Bininda-Emonds ORP, Mace GM, Wayne RK (2000) Considering evolutionary processes in conservation biology. *Trends in Ecology & Evolution*, **15**, 290–295.

DaCosta JM, Sorenson MD (2014) Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE*, **9**, e106713.

Davey JW, Blaxter ML (2010) RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Ferrari L, Orozco-Esquivel T, Manea V, Manea M (2012) The dynamic history of the Trans-Mexican Volcanic Belt and the Mexico subduction zone. *Tectonophysics*, **522–523**, 122–149.

Fjeldså J, Bowie RCK (2008) New perspectives on the origin and diversification of Africa's forest avifauna. *African Journal of Ecology*, **46**, 235–247.

Fjeldså J, Bowie RCK, Rahbek C (2012) The role of mountain ranges in the diversification of birds. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 249–265.

Hearn J, Stone GN, Bunnefeld L *et al.* (2013) Likelihood-based inference of population history from low coverage de novo genome assemblies. *Molecular Ecology*, **1**, 198-21.

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

Hewitt GM (2004) The structure of biodiversity - insights from molecular phylogeography. *Frontiers in Zoology*, **1**, 4.

Hickerson MJ, Carstens BC, Cavender-Bares J *et al.* (2010) Phylogeography's past, present, and future: 10 years after. *Molecular Phylogenetics and Evolution*, **54**, 291–301.

Huang H, Knowles LL (2014) Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Systematic Biology*, syu046.

Hubbell SP (2001) *The Unified Neutral Theory Of Biodiversity And Biogeography (MPB-32)*. Princeton University Press.

Kessler M (2002) The elevational gradient of Andean plant endemism: varying influences of taxon-specific traits and topography at different taxonomic levels. *Journal of Biogeography*, **29**, 1159–1165.

Kluge J, Kessler M, Dunn RR (2006) What drives elevational patterns of diversity? A test of geometric constraints, climate and species pool effects for pteridophytes on an elevational gradient in Costa Rica. *Global Ecology and Biogeography*, **15**, 358–371.

Kruckeberg AR, Rabinowitz D (1985) Biological Aspects of Endemism in Higher Plants. *Annual Review of Ecology and Systematics*, **16**, 447–479.

Lynch M (2002) Gene duplication and evolution. *Science*, **297**, 945–947.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014a) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, advance online publication.

Mastretta-Yanes A, Zamudio S, Jorgensen TH *et al.* (2014b) Gene duplication, population genomics and species-level differentiation within a tropical mountain shrub. *Genome Biology and Evolution*, evu205.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

McCormack JE, Zellmer AJ, Knowles LL (2010) does niche divergence accompany allopatric divergence in *Aphelocoma* jays as predicted under ecological speciation? Insights from tests with niche models. *Evolution*, **64**, 1231–1244.

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology*, **8**, 122–128.

Moritz C (1994) Defining "Evolutionarily Significant Units" for conservation. *Trends in Ecology & Evolution*, **9**, 33–375.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

Papadopoulou A, Anastasiou I, Spagopoulou F *et al.* (2011) Testing the species–genetic diversity correlation in the Aegean archipelago: toward a haplotype-based macroecology? *The American Naturalist*, **178**, 241–255.

Parra-Olea G, Windfield JC, Velo-Antón G, Zamudio KR (2012) Isolation in habitat refugia promotes rapid diversification in a montane tropical salamander. *Journal of Biogeography*, **39**, 353–370.

Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ (2014) ABC inference of multi-population divergence with admixture from un-phased population genomic data. *Molecular Ecology*, advance online publication.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Slavov GT, DiFazio SP, Martin J *et al.* (2012) Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa. New Phytologist*, **196**, 713–725.

Smith BT, McCormack JE, Cuervo AM *et al.* (2014) The drivers of tropical speciation. *Nature*, advance online publication.

Steinmann, V. W., Ramírez-Amezcua, Y., Arredondo-Amezcua, L. y Hernández-Cárdenas, R. A. In preparation. Flora alpina del centro de México.

Taberlet P, Zimmermann NE, Englisch T *et al.* (2012) Genetic diversity in widespread species is not congruent with species richness in alpine plant communities. *Ecology Letters*.

Tin MMY, Rheindt FE, Cros E, Mikheyev AS (2014) Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype-calling accuracy. *Molecular Ecology Resources*, advance online publication.

United Nations (1992) Convention on biological diversity. In: *Treaty Series* , pp. 142–382. Río de Janeiro, Brasil.

Uscanga A, Mastretta-Yanes A, López, H, Emerson, B and Piñero D. In preparation. Filogeografía comparada de especies de alta montaña de la Faja Volcánica Transmexicana.

Vellend M (2003) Island biogeography of genes and species. *The American Naturalist*, **162**, 358–365.

Vellend M, Geber MA (2005) Connections between species diversity and genetic diversity. *Ecology Letters*, **8**, 767–781.

Wang IJ (2013) Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation. *Evolution*, **67**, 3403–3411.

Wang IJ, Bradburd GS (2014) Isolation by Environment. *Molecular Ecology*, advance online publication.

Warren IA, Ciborowski KL, Casadei E *et al.* (2014) Extensive local gene duplication and functional divergence among paralogs in Atlantic salmon. *Genome Biology and Evolution*, evu131.

APPENDIX I

Supporting Information for Chapter 3

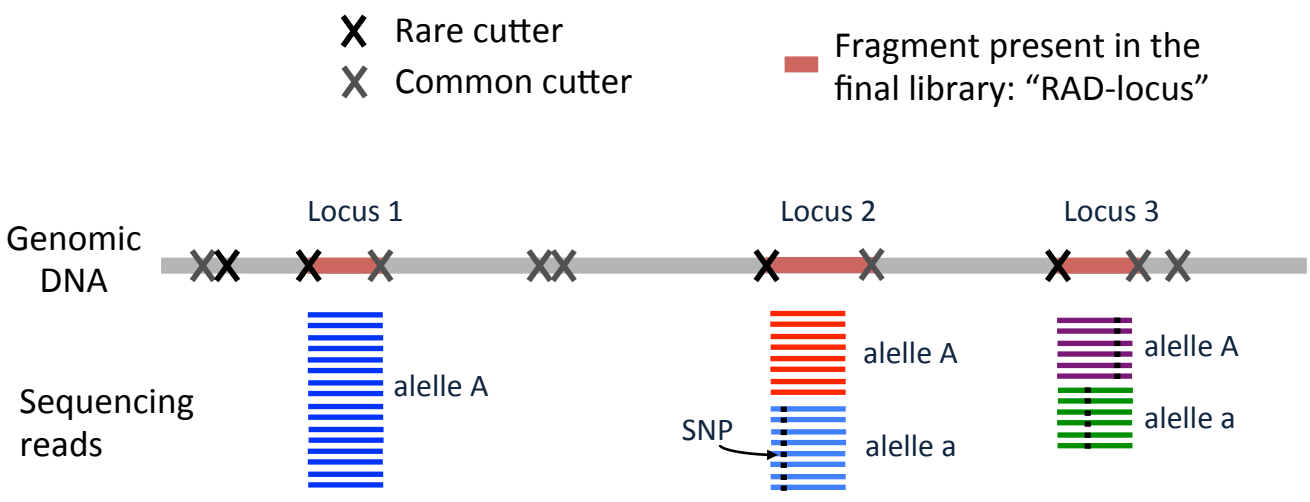
Insomnia Hunter, ticatla 2012

A scientist in his laboratory is not only a technician: he is also a child placed before natural phenomena which impress him like a fairy tale
- Marie Curie

Supporting Information 1. Schematic diagram of RAD data genotyping and differences between replicates

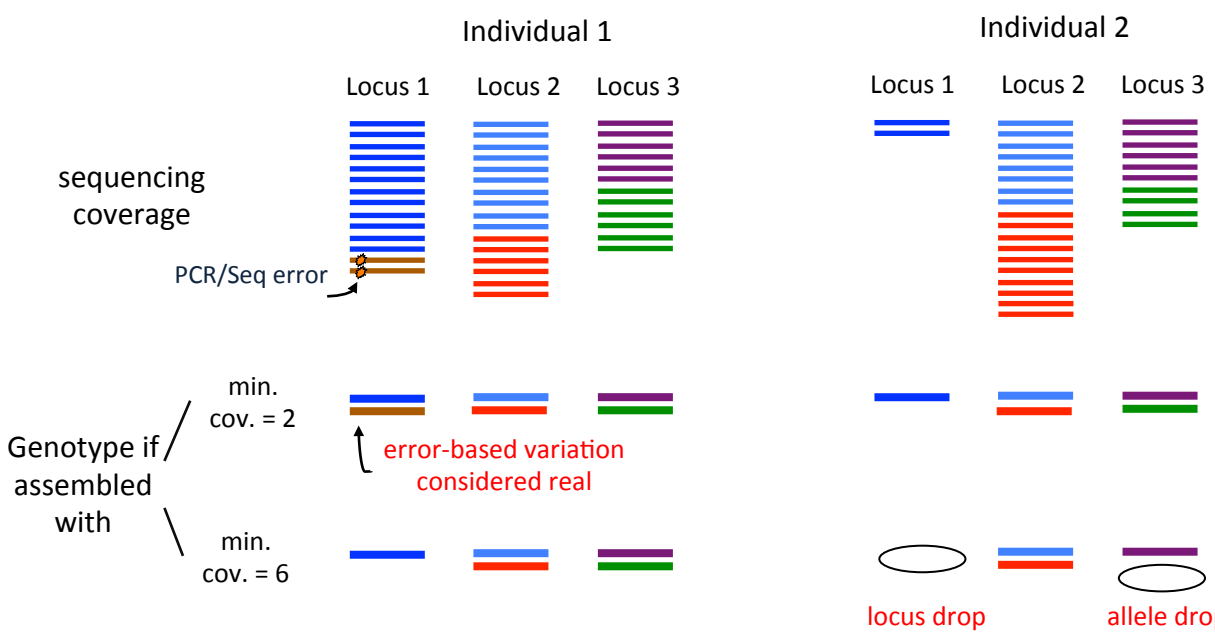
from Mastretta-Yanes *et al.* (2014). RAD sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference

(I) RAD-loci, alleles, SNPs and coverage



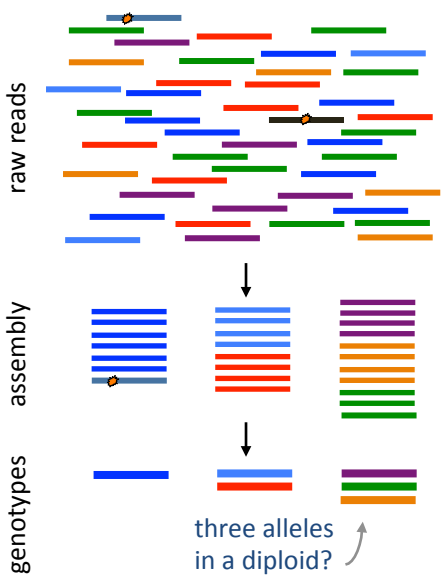
An example double digest RAD library: genomic DNA is digested with two restriction enzymes (a rare and a common cutter) and processed to create sequencing competent fragments. The RAD-loci present in the final library are the fragments kept after the size selection. A RAD-locus is thus a short DNA sequence. Each locus can have one or more alleles, which differ from each other by a small number of SNPs (black squares). Sequencing produces a number of reads per allele, which is referred to as coverage. The same principle applies to traditional RAD-seq libraries.

(II) The role of coverage



During assembly and genotyping, setting a threshold for minimal coverage (defined by researchers) allows to distinguish between PCR/sequencing error and real variation. If it is set too low it can lead to error-based variation being considered real. However, if it is set too high it can cause locus or allele dropout. Locus dropouts results in missing data, but allele dropout results in inferences of homozygosity, when the underlying state of the locus is heterozygous. See Table 1 for reasons that can lead to heterogeneous and low coverage, and for other sources of error.

(III) Loci, alleles, SNPs, error and *de novo* assembly



In the absence of a reference genome, loci are assembled by matching together similar sequences and considering them either different loci, or alleles from the same locus. This is done based on a given number of mismatches (which are defined by researchers).

The main challenges are that: 1) repetitive regions and paralogs can be erroneously assembled together, creating alleles that are not biologically meaningful; 2) PCR/sequencing can form error-based alleles; and 3) both 1 and 2 occur at the same time.

As a consequence, there can be genotypic differences between individuals that are not biologically true.

(IV) Differences between replicates and error rates

DNA replicates derived from the same sample should have the same genotype, and any differences can be considered error produced from any of several possible reasons (Table 1). The differences between replicates can be examined at the locus, allele and SNP levels. Consider 6 RAD-loci genotyped in 4 individuals, of which we have replicates for individual 1 and 2:

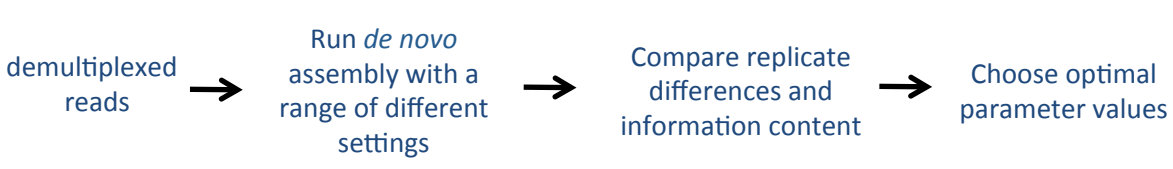
	Individual 1		Individual 2		Individual 3	Individual 4
	Replicate I	Replicate II	Replicate I	Replicate II		
Locus 1		**AA**		**aa**	**Aa**	**AA**
Locus 2	**Aa**	**Aa**	**aa**	**Aa**		**AA**
Locus 3	**AA**		**AA**	**AA**	**AA**	**AA**
Locus 4	**aa**	**aa**			**aa**	**aa**
Locus 5			**Ab**	**AA**	**aa**	
Locus 6		**Aa**	**Aa**	**Aa**	**Aa**	**AA**

If we look at the distribution of missing loci **per replicate pair**, we can see that for individual 1 loci 1 & 6 are missing in replicate I; locus 3 is missing for replicate II and locus 5 is missing in both replicates. This means that for the replicate pair of individual 1 the **number of missing loci** is 4 and that the **proportion of loci missing** relative to all loci in the population is 4/6. Note that of the four missing loci, only locus 5 was lost in both replicates (and therefore does not result in a genotypic difference between them), whilst loci 1, 3 and 6 were lost in one replicate or the other, but not in both. Therefore, the proportion of **missing loci where a locus was lost only in one of the replicates** is 3/4. If we estimate this same proportion but against the total number of loci found for all individuals we have 3/6, which is **the locus error rate**.

There is error at the allele level if the alleles of a locus are different for a replicate pair. This can be caused by allele dropout due to low coverage or assembly error. For example, for individual 2, locus 2 is homozygous for replicate I and heterozygous for replicate II, and in locus 5 there is a different allele not present in both replicates. If we count mismatches like these and divide by the number of loci present in both replicates we have that the **allele error rate** for individual 2 **is 2/4**. Since alleles of a RAD-locus can differ by more than one SNP (see diagram I), the same principle can be applied to estimate the **SNP error rate**.

(IV) Replicates can aid *de novo* assembly

Because DNA replicates derived from the same sample should have the same genotype, one can evaluate which parameter values of the assembly pipeline optimize for a high number of loci with less differences between replicate pairs.



Supporting Information 2

INDEX

I. Summary of ddRAD labwork, description of final libraries and sequencing output

Seventy-five specimens from the eight *B. alpina* populations (six to ten per population) plus three samples of each of the outgroups were used to construct ddRAD libraries with the reagents and conditions explained below. Individual DNA extracts were randomly divided into three groups, each of them corresponding to a pool of libraries (BERL1, BERL2, BERL3, Table 1). Each group was comprised of 27 *Berberis* sp. samples and five replicates for a total of 32 barcoded (sequence-tagged) individuals. For each of the groups, the five replicates consisted of four intra-library replicates and one inter-library replicate. Replicates had the same DNA source but were treated and barcoded independently. Replicates were chosen randomly but included at least one replicate per outgroup and population. Within each group of 32 barcoded samples all positions on the PCR plates were randomly selected (Table 1). The digestion, ligation and PCR steps were performed in the same plate for the three groups. Samples of the same group were then pooled and the size selection for all groups was performed on the same gel. The well position for each sample inside it's corresponding lane was randomly chosen. Each library (group of 32 individual samples) was sequenced in separate lane on an Illumina HiSeq2000 with a single read run, 100bp long at the Lausanne Genomic Technologies Facility, Switzerland.

Library preparation

For library preparation we followed a modified version of the Parchman et al.,

(2012) double digest RAD protocol. For adapter and PCR primer sequences and full protocol see section II of this Supplementary Material. In summary, the three library preparations consisted of the following steps: (1) Phenol-chloroform wash and ethanol precipitation of DNA extractions. DNA concentrations after the wash were standardized to approximately 45 ng/μL, with the exception of some samples where concentration was <10 ng/μL. (2) Digestion of each DNA sample with EcoRI (HF) and MseI at 37°C for eight hours, followed by inactivation of restriction enzymes at 65°C for 20 minutes. (3) Adapter ligation was performed in the same well from the digestion reaction using T4 DNA ligase at 16^{O}C for six hours. A general (non-sample specific) MseI adaptor was added to all samples in the ligation master mix, followed by the addition of a sample-specific EcoRI adaptor for each DNA sample. For sample-specific EcoR1 adaptors a unique 7bp long barcode + protective base (C) was used for each of the 96 barcoded EcoRI adapters. This adaptor could have been reused if using different Illumina PCR indexing primers as in the dual indexing method of Peterson et al., (2012), although for this experiment we only used one index. (4) Digestion-ligation products were diluted with 189 μL of 0.1x TE. (5) Amplification of adapters + barcodes ligated-fragments using Illumina PCR primers. To ameliorate stochastic differences in PCR production of fragments across reactions, the following reaction procedure was performed individually for each restriction-ligation product, and combined at a later stage (see step 8). Amplification reactions were performed with Phusion Taq, Phusion PCR buffer, dNTP, $MgCl_2$, DMSO and a PCR primer mix of ILLPCR1 and ILLPCR2-bar04 under the following conditions: 98 °C for 30 seconds; 30 cycles of: 98 °C for 20 seconds, 60° C for 30 seconds, 72° C for 40 seconds; final extension at 72° C for 10 minutes. (6) Addition of primers and

dNTPs for a final thermal cycle to reduce the concentration of single-stranded or heteroduplex PCR products. For this step, a reaction mix containing the Phusion PCR Buffer, dNTPs and the same PCR primer mix of the previous step (but excluding Phusion Taq and $MgCl_2$) was added to each of the previous reactions and cycled at 98° C for 3 minutes, 60° C for 2 minutes and 72° C for 12 minutes. (7) Electrophoresis of 2 µL of the reaction from step 6 in a 1.5% agarose gel, run at 100 V for 30 minutes to confirm reaction success. (8) Pooling of reactions within each library (BERL1, BERL2 and BERL3) into a single 1.5 ml microcentrifuge tube each which was then evaporated to half the volume. (9) Selection of a size range between 350-900 bp by manual excision from a 1.5% agarose gel run at 100 V for 1.45 hours. Purification of the gel extracts was performed with the MiniElute Qiagen gel extraction kit using one column per gel lane. The three libraries were run in the same gel, using 9 adjacent wells (40µL each) per library and separating each library with empty wells and DNA ladder. The final elutions of columns belonging to the same library were pooled together for a final ethanol precipitation. (10) Measurement of library concentration using Qubit fluorometer and submission to the Fragment Analyzer Automated CE System to evaluate the desired concentration and range of the fragments selected.

We used enzymes from New England Biolabs: EcoRI-HF (R3101S), MseI (R0525S), T4 DNA Ligase (M0202S), Phusion Taq (M0530S) and their correspondent buffers.

Table 1. Samples and barcodes used in the preparation of three ddRAD libraries for Illumina sequencing.

Sample DNA ID*	Library	Type in Library	Barcode	Position in plate
IzB10	BERL1	sample	TCAATATC	A1
OutBsHd115	BERL1	sample	GAATAGTC	B1
PeB05	BERL1	sample	TTGACTCC	C1
PeB09	BERL1	sample	TCTTCTGC	D1
IzB08	BERL1	sample	TTCAACCC	E1
TlB02	BERL1	sample	TTGAGGAC	F1
TlB07	BERL1	sample	AATCAGTC	G1
AnB03	BERL1	sample	GGCATATC	H1
PeB07	BERL1	sample	ACCGCCTC	A2
MaB06	BERL1	sample	GATTGATC	B2
IzB05	BERL1	intralane_replicate_04	AACTGCGC	C2
AnB02	BERL1	sample	TGATCGCC	D2
AjB02	BERL1	sample	GGCAAGGC	E2
MaB09	BERL1	sample	TCGCAAGC	F2
MaB21	BERL1	sample	TCCGGAAC	G2
OutBsHd112	BERL1	sample	ATACCGCC	H2
AnB05	BERL1	sample	ACTTGAAC	A3
IzB01	BERL1	sample	TATGCAGC	B3
TlB20	BERL1	sample	GAAGCGCC	C3
MaB21	BERL1	intralane_replicate_03	GAGGTAGC	D3
ToB02	BERL1	sample	CCGCTACC	E3
AjB21	BERL1	sample	CAAGACCC	F3
PeB17	BERL1	sample	CTCTCAGC	G3
MaB07	BERL1	sample	AATCTCAC	H3
TlB01	BERL1	sample	GCAGGATC	A4
AjB21	BERL1	intralane_replicate_01	GGTAGGTC	B4
IzB09	BERL1	sample	CATCGTCC	C4
PeB01	BERL1	INTERlane replicate	TTCAGAGC	D4
IzB05	BERL1	sample	CTGCTGAC	E4
PeB03	BERL1	sample	AGAGATTC	F4
OutBsHd112	BERL1	intralane_replicate_02	CGCAATTC	G4
AnB01	BERL1	sample	CGCTTGAC	H4
AnB09	BERL2	sample_22	CCGTTCAC	A5
AjB12	BERL2	sample_12	GCCGTCAC	B5
ZaB01	BERL2	sample_05	TTAGGCGC	C5
PeB06	BERL2	sample_01	CGGTTAGC	D5
IzB06	BERL2	sample_09	AGACGGAC	E5
TlB03	BERL2	sample_19	TAGCATCC	F5
OutBtAl212	BERL2	sample_13	TTCCTGCC	G5
ToB24	BERL2	sample_08	AATGATGC	H5
AjB20	BERL2	sample_04	AGGAGGCC	A6
IzB06	BERL2	intralane_replicate_04	TTATCCTC	B6

IzB03	BERL2	sample_06	ACTCTAGC	C6
AnB04	BERL2	intralane_replicate_02	GGCCATCC	D6
PeB06	BERL2	intralane_replicate_01	CAGAGTTC	E6
AnB08	BERL2	sample_23	ATCATCAC	F6
PeB16	BERL2	sample_07	GAACTTGC	G6
TlB10	BERL2	sample_20	CGCGGAGC	H6
MaB04	BERL2	sample_15	TGCCAGAC	A7
AjB01	BERL2	sample_03	TCTCTTAC	B7
ZaB06	BERL2	sample_10	GGTCGACC	C7
ToB07	BERL2	sample_14	GCTCTCCC	D7
ToB04	BERL2	sample_21	GGATATAC	E7
ZaB06	BERL2	intralane_replicate_03	GGACTCAC	F7
MaB25	BERL2	sample_25	TCTATCGC	G7
AnB04	BERL2	sample_02	GACGGTAC	H7
PeB01	BERL2	sample_17	GTTCATAC	A8
TlB19	BERL2	sample_27	ACTACGAC	B8
TlB09	BERL2	sample_26	AGCTTCTC	C8
ToB08	BERL2	sample_16	ACCGAGGC	D8
PeB01	BERL2	INTERlane replicate	TATACTAC	E8
AjB10	BERL2	sample_18	GGTATTGC	F8
AnB07	BERL2	sample_24	CCGTCTTC	G8
ToB23	BERL2	sample_11	CTGGAATC	H8
PeB01	BERL3	INTERlane replicate	TTCCGCAC	A9
MaB22	BERL3	sample_03	CAATCATC	B9
ToB03	BERL3	intralane_replicate_02	AAGCGAGC	C9
OutBtAl216	BERL3	sample_16	GAATGCCC	D9
AjB11	BERL3	sample_06	CGGAAGAC	E9
ToB25	BERL3	sample_09	AGGAATGC	F9
AjB18	BERL3	sample_18	CGGTATCC	G9
IzB02	BERL3	sample_20	GGAGTACC	H9
AnB06	BERL3	sample_05	CTAGTCTC	A10
ToB05	BERL3	sample_12	ATGACGGC	B10
TlB04	BERL3	sample_14	TAGGACTC	C10
MaB02	BERL3	sample_01	GCAACTTC	D10
IzB04	BERL3	sample_08	GCGTCGCC	E10
ZaB05	BERL3	sample_17	AATGGCTC	F10
OutBsHd113	BERL3	sample_19	TCAACGGC	G10
ZaB03	BERL3	sample_07	GTATCGGC	H10
MaB08	BERL3	intralane_replicate_04	ATGGCAAC	A11
TlB08	BERL3	sample_13	TTCGGTCC	B11
ToB22	BERL3	sample_23	CGTACGGC	C11
ZaB04	BERL3	sample_10	TCAAGCAC	D11
OutBtAl214	BERL3	sample_22	CATTATTC	E11
MaB08	BERL3	sample_11	AACTCGAC	F11
PeB04	BERL3	sample_26	CCTGGACC	G11
AjB19	BERL3	sample_27	CTGGCTGC	H11

OutBtAl216	BERL3	intralane_replicate_01	CTTACCTC	A12
PeB08	BERL3	sample_25	CTACCTTC	B12
AjB09	BERL3	sample_02	GTCCTCTC	C12
MaB03	BERL3	sample_21	TGGTTCCC	D12
IzB07	BERL3	sample_04	ACCTACCC	E12
ZaB02	BERL3	sample_15	CTATGAAC	F12
ToB03	BERL3	sample_24	AAGGAACC	G12
TlB08	BERL3	intralane_replicate_03	ACGCAGAC	H12

Sample IDs starting with Out correspond to the outgroups *B. trifolia* (OutBt) and *B. pallida* (OutBs). In the rest of the samples, the first two letters correspond to the Population ID as follows: El Zamorano (Za), Cofre de Perote (Pe), La Malinche (Ma), Cerro Tlaloc (Tl), Iztaccihuatl (Iz), Ajusco (Aj), Nevado de Toluca (To) and Cerro San Andres (An).

Description of final libraries and sequencing output

The final concentrations of BERL1, BERL2 and BERL3 were 23, 18.3 and 12.8 ng/µL, respectively. The fragment size distributions shown by the Fragment Analyzer showed a curve from ~ 150 bp to ~2000 bp, with a peak at 301, 305 and 318 bp in BERL1, BERL2 and BERL3 respectively. The number of raw reads after sequencing was of 105,261,642 for BERL1, 102,970,822 for BERL2 and 100,398,355 for BERL3 (Fig. 1, 2). Fifteen out of the 96 sequenced samples had less than one third of the median number or reads and did pass the filters of the downstream analyses. Among them was the interlibrary replicate sequenced in lane BERL1.

The sequencing error rate calculated based on the PhiX that was spiked into the libraries mix was 0.0012, 0.0028 and 0.0041 at cycles 35, 75 and 100 for BERL1; 0.0013, 0.0031 and 0.0045 for BERL2 and 0.0036, 0.0043, 0.0056 for BERL3.

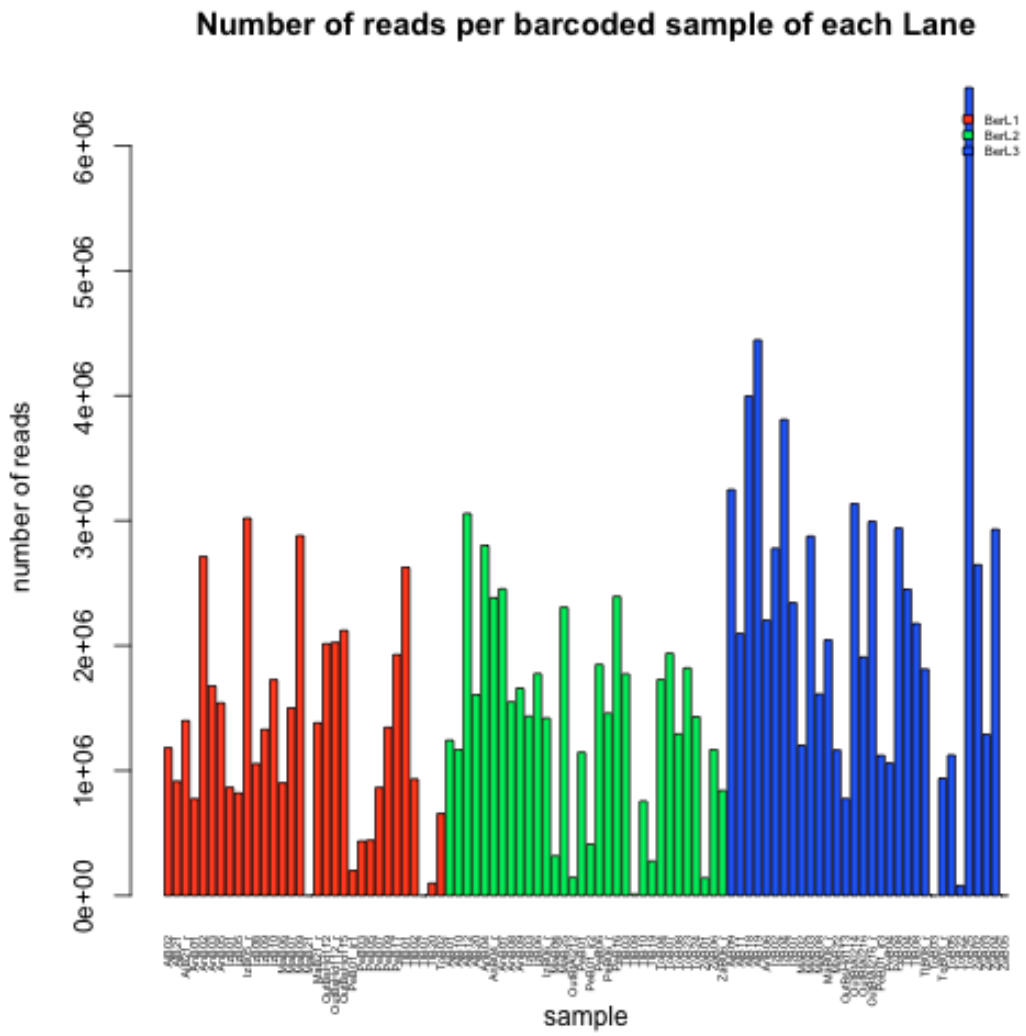


Figure S1.1. Number or raw reads per barcoded sample for each sequencing lane.

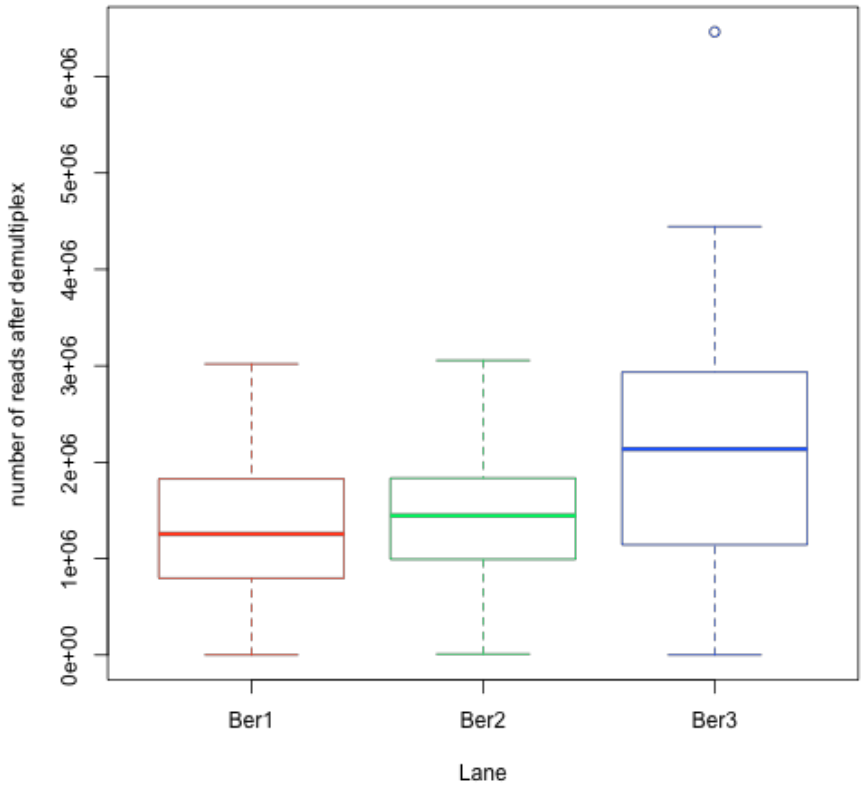


Figure S1.2. Reads per sequencing lane after demultiplexing for each sequencing lane.

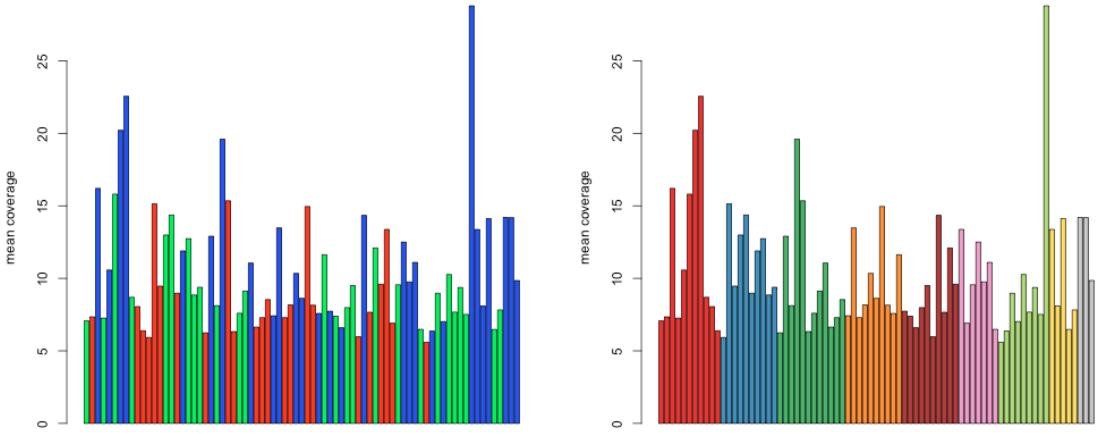


Figure S1.3. Mean coverage per sample after processing the data with *Stacks* optimal profile settings (see results). Left: color key corresponding to sequencing lanes as in Fig. S1.1. Right: color key corresponding to geographic origin of samples as in Fig. 1 (main text).

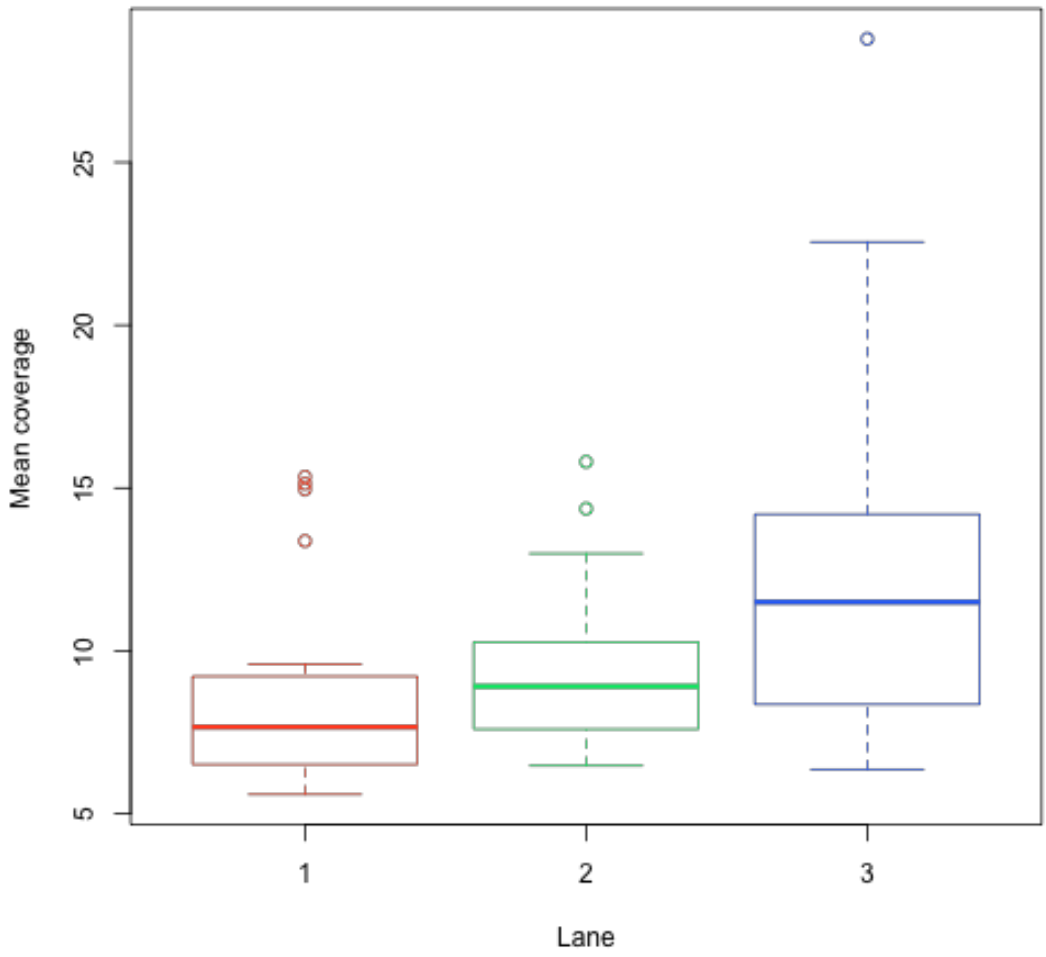


Figure S1.4. Mean coverage per sample after processing the data with *Stacks* optimal settings (see results) for each sequencing lane. Color key as in Fig. S1.1.

References

Parchman, T.L., Gompert, Z., Mudge, J., Schilkey, F.D., Benkman, C.W., and Buerkle, C.A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol. Ecol. *21*, 2991–3005.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE *7*, e37135.

II. Modified double digest RAD (ddRAD) sequencing protocol

Note: we present here the protocol as we followed it to prepare the *Berberis* libraries, but we made changes that improved it in further experiments. We advice to contact the authors for the most updated version of the protocol.

December 2012

Modifications added by A. Brelsford and A. Mastretta-Yanes based on protocol developed by Parchman, T.L., Gompert, Z., Mudge, J., Schilkey, F.D., Benkman, C.W., and Buerkle, C.A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. Molecular Ecology 21, 2991–3005.

Summary of changes from Parchman et al. protocol:

- Added dual-index barcoding to allow multiplexing >96 samples per library
- Modified restriction and ligation mixes to maintain consistent buffer concentration across both steps
- Addition of primers and dNTPs for a final thermal cycle, in order to reduce production of single-stranded or heteroduplex PCR products

Glossary

Adapter: fully or partially double-stranded product of annealing two oligos. Adapters are ligated to genomic DNA at restriction enzyme cut sites in order to add barcodes and common PCR priming sequences.

Barcode: short DNA sequence downstream of the sequencing primer annealing region of an adapter. Used to resolve products of different ligation reactions (usually separate individuals) after sequencing pooled libraries.

Fragment: section of genomic DNA resulting from restriction enzyme cleavage.

Index: short DNA sequence introduced during PCR amplification of the final library that uniquely identifies products of that PCR reaction. Used combinatorialy with Adapter P1 barcodes to resolve multiplexed sample pools.

Library: a collection of sequencing-competent fragments.

Notice that the dual indexing involves a barcode and an index, while other protocols use a single sequence-tag.

Note on starting DNA material

DNA should ideally be at a minimum concentration of 20 ng/μL and a maximum concentration of 150 ng/μL, but lower concentrations (up to 5 ng/μl) may still work. It is advisable to homogenize sample's concentration before digestion if the variation is orders of magnitude larger.

DNA can be extracted using either a phenol chloroform protocol or a Qiagen extraction kit. Some extractions can carry a salt excess or inhibitors for enzyme activity (e.g. some terpenoids in plant DNA extractions). If such is the case, it is advisable to perform a phenol chloroform DNA cleaning following these steps:

1. To 100 µl of eluted DNA, add 0.5 µl of 20% SDS and 100 µl phenol-chloroform (Sigma Aldrich P2069-100ML)
2. Mix well (vortex gently)
3. Centrifuge at room temperature for 5 min at 14,000 rpm.
4. Pipette the aqueous phase (upper phase, aprox. 80 µl, it is better to leave some DNA than to pipette the organic phase) to a new labeled tube.
5. Discard original tube
6. Add 1/10 volume Na acetate 3M pH 4.8 or 5.2 (i.e. 8 µl for 80 µl DNA solution in this example)
7. Add 2 volumes ethanol 100% (storage -20°C) (i.e. 176 µl in this example)
   Total volume: 264ul, possible with 264 ng. If concentration is below this (1ng/µl), you must add a carrier: glycogen or linear acrylamide.
8. Vortex gently
9. Put on dry ice for 30 min. or over night at -20°C
10. Centrifuge at 4°C for 30 min at 14,000 rpm.
11. Discard the supernatant.
12. Wash with 500ul ethanol 70% (storage 4°C)
13. Centrifuge at 4°C for 5 min at 14,000 rpm.
14. Discard the supernatant
15. Quick spin
16. Pipette out the last drop of ethanol
17. Speed Vac for 3 or 5-7 min at room temperature.
18. Resuspend in 25ul of Tris 10 mM pH 7.5 or 8.0

Enzimes
We used New England Biolabs enzimes: EcoRI-HF (R3101S), MseI (R0525S), T4 DNA Ligase (M0202S), Phusion Taq (M0530S) and their correspondent buffers.

0. Preparation of adapters and primers working solutions

Barcoded EcoRI adapters
Anneal EcoRI oligo pairs (Table 1) by mixing 1 µL of each oligo in a pair (100 µM stock) with 98 µL of water to make 100 µL of 1 pmole/µL (1 µM) of annealed, doubled stranded adaptor stock. Heat to 95ºC for 5 minutes and bring to 20ºC with a ramp of 0.1 ºC/s to slowly cool down. Once they are ready it is possible to freeze it for later use. Keep the set of adaptors organized in plate format that is convenient for later use in setting up reactions.

MseI adapter
Mix 100 µL of the MseI-adap1-bar and MseI-adap2-bar oligos (Table 1, 100 µM stock) with 800 µL of water to make 1000 µL of 10 pmole/µL (10 µM) stock. Heat to 95ºC for 5 minutes and bring to 20ºC with a ramp of 0.1 ºC/s to slowly cool down. Freeze for later use.

PCR primers
Mix 50 µL of the ILLPCR1 and ILLPCR2-bar_n (Table 1) with 900 µL of water to make a working solution (5 µM of each oligo). The dual-indexing barcode is incorporated in the ILLPCR2-bar_n oligo, so **this step must be repeated for each dual-indexing barcode** (mixing each uniquely barcoded version of ILLPCR2 with ILLPCR1, which will be the same oligo in all working solutions). This step

is necessary only if more than the amount of barcoded EcoRI adapters (in this case 96) are going to be used.

Note: If using only 2 indexed primers (i.e. to pool 96x2=192 samples) Illumina recommends to use the ILLPCR2_ind06 and ILLPCR2_ind12. If three primers, use 4, 6, 12.

1. Restriction Digest:

1. Prepare master mix I (see below, 3 μL prepared per sample), mix by vortexing, and centrifuge. We have found that making 1.2x per sample is sufficient to avoid running out due to high viscosity and/or pipetting error. Work on ice all times.

MASTER MIX I: DIGESTION	
SbfI-MseI	**Vol (μl) 1x**
10X T4 Buffer	0.9
1 M NaCl	0.45
1 mg/mL BSA	0.45
H2O	0.85
MseI (10,000 U/ml)	0.1
EcoRI (HF) (20,000 U/ml)	0.25
Total mix volume per sample	3

2. Place 6 μL of sample DNA in each well of a plate.

3. Add 3 μL of the combined master mix I to each well. The total reaction volume should be 9 μL.
4. Cover and seal the plate, centrifuge and incubate at 37°C for 8 hours* on a thermal cycler with a heated lid. Heat kill the enzime with 20 mins at 65ºC. Keep at 4ºC afterwards.

* The digestion time can be reduced to 3 hrs, but if the genome size is large it is advisable to perform the reaction during a long time to ensure complete digestion.

2. Adaptor Ligation

1. Thaw MseI and EcoRI adaptors. These adaptors should already be annealed (step 0).

2. Prepare master mix II (see below, 1.6 μL prepared per sample), mix by vortexing, and centrifuge. As above, it is best to prepare an extra 20% (1.2x/sample).

MASTER MIX II: LIGATION	
EcoRI-MseI	**Vol (μl) 1x**
10x T4 Buffer	0.16
1M NaCl	0.13
1 mg/mL BSA	0.13
Water	0.0125
MseI adapter 10 uM	1

T4 DNA Ligase (400,000 U/ml)	0.1675
Total mix volume per sample	1.6

3. Add 1.6 μL to each well of the restriction digested DNA.

4. Add 1 μL of the EcoRI adaptor to each well (a unique barcoded adaptor for each DNA sample).

5. The total reaction volume should now be 11.6 μL. Cover and seal the plate, vortex, centrifuge and incubate at 16° C for 6 hours on a thermocycler.

6. Dilute the Restriction-Ligation reaction with 189 μL of water (or 0.1x TE for long-term storage). Store at 4° C for a month, or -20° C for longer.

3. PCR Amplification

This PCR step uses the Illumina PCR primers to amplify fragments that have our adapters + barcodes ligated onto the ends. To ameliorate stochastic differences in PCR production of fragments in reactions, we run two separate 10 μL reactions per restriction-ligation product (i.e. perform next two steps twice with the same samples), and later combine them. If your sequencing batch includes fewer than 32 individuals, run each PCR at double volume (20 μL) to produce sufficient library quantity.

1. Prepare master mix III (see below, 8 μL per sample, but remember to prepare 2 PCR reactions per sample), vortex and centrifuge. **If you are running the dual-indexing protocol, be sure to prepare separate master mixes for samples to be indexed with different Illumina barcodes- these will each require a different primer mix (see step 0).** Remember, if only 2 index primers will be used use the ILLPCR2-bar06 and ILLPCR2-bar12, if three primers, use 4, 6, 12.

MASTER MIX III: PCR

	Vol (μl) 1x
Water	4.875
Phusion Buffer	2
dNTP (25mM)	0.08
MgCl2 (50 mM)	0.2
PCR Primer Mix	0.67
Phusion Taq	0.1
DMSO	0.075
Total mix volume per sample	8

2. Add 8 μL of the combined master mix III to each well of a plate.

3. Add 2 μL of the diluted restriction-ligation mix.

4. Thermal cycler profile for this PCR: 98° C for 30s; 30 cycles of: 98° C for 20s, 60° C for 30s, 72° C for 40s; final extension at 72° C for 10 min.

5. Prepare master mix IV (see below, 1 μL per sample), **remember to account for dual-indexing primers; they need to be prepared in separate mixes.** It is not necessary to add more polymerase or $MgCl_2$ as there is still enough from the previous PCR. This step reduce production of single-stranded or heteroduplex PCR products.

MASTER MIX IV: PCR final cycle

	Vol (μl)	1x
Water	0.05	
Buffer (Phusion)	0.2	
PCR primer mix	0.67	
dNTP (25 mM)	0.08	
Total mix volume per sample	1	

6. Add 1 μL to each PCR product (keep cold), run thermocycler profile as follows: 98° C for 3 min, 60° C for 2 min, 72° C for 12 min.

Note: it is advisable to run all the reactions in the same thermocycler.

4. Confirm reaction success of each sample (optional)

Pool equal samples of the two PCR reactions into the same plate ("stack" the plates) and run each PCR product on a 1.5% agarose gel for 20-30 minutes. You should see a smear of PCR product from 150 bp to between 500 and 1000 bp, often with a bright band of primer dimer at 130 bp. Samples that failed to amplify, or amplified only the adapter dimer, can be excluded from the pool (except negative controls, those must be pooled).

5. Gel Purification and Size Selection

In this protocol we used an agarose gel extraction to undertake the size selection, but it can also be done using a Blue Pippin (Sage Science) or using different Agencourt AMPure XP ratios. Automated methods like those reduce the variance of the size selection and provide better results.

To perform the standard agarose gel extraction follow the steps below.

Agarose gel size selection

1. Pool PCR product from both replicates and all samples from into one tube (see note before regarding pooling samples with different indexes). Measure DNA concentration using the Qubit. Depending on the genome size, enzymes and number of samples you should expect a concentration between 8 and 40 ng/μl.
2. Use a SpeedVac (keeping the temperature low) to evaporate the pool of PCR to increase concentration and reduce the number of wells needed tin the gel. Usually a final volume of half the original works fine (do not go below this due to salts overconcentration), but if the original concentration is high the final sample may represent and overload for the gel. The final amount of DNA in the gel should not be larger than 240 ng/mm for a tick gel or 120 ng/mm for a standard.

As a guideline, 200 µl of PCR pool at 65 ng/µl + 40 µl LB can be run loading 50-80 µl in 3-4 wells of 18 mm width. More volume will require more wells.

3. Fill a gel rig with new, clean TBE buffer and prepare a 1.5 or 2% agarose gel. Run the pooled PCR product at 100 volts for 1.45-2 hours. Include a good ladder on multiple gel lanes so that a clear line can be visualized across the gel, leave an empty lane between the ladder and the library sample. Ethidium bromide in the gel will not interfere after gel purification. A good approach is to tape together several gel combs to allow for larger wells (e.g. tape 5 1.5mm combs to generate a single one of 18 mm width), and to load 50-80 µL of the pool into each well.
4. Cut the desired region out of the gel using the large end of sterile 1000 µl pipette tips or with a sterile razor. We have used the region from 350-900 bp. To minimize gel exposure to UV it is possible to perform the extraction with the UV off by first using it only to mark with a 10 µl pipette tip the bands of interest in the ladders. Then use a dark straight paper or ruler below the gel bead to create a guide using the ladders marks a reference.
5. Store the excised gel fragments in clean 1.5 or 2 ml colorless eppendorf tubes (ensure tube size will be enough for QG buffer and isopropanol volume added in next steps). Proceed to extraction purification or store at 4ºC until then.

Extraction purification

The following steps use the QIAquick Minielute Gel Extraction Kit with modifications in the incubation and centrifuge conditions.

6. Weigh the gel slice (tare an empty tube first, then weight the one from step 5. Add 3 volumes of Buffer QG to 1 volume of gel (100 mg gel ~ 100 µl). The maximum amount of gel slice per spin column is 400 mg. For >2% agarose gels, add 6 volumes Buffer QG.
7. Incubate at 22°C for 30 min or until the gel slice has completely dissolved. This enriches GC bonds. Vortex gently the tube every 2–3 min during incubation to help dissolve the gel.
8. After the gel slice has dissolved completely, check that the color of the mixture is yellow (similar to Buffer QG without dissolved agarose). If the color of the mixture is orange or violet, add 10 µl 3 M sodium acetate, pH 5.0, and mix. The color of the mixture will turn to yellow. Note: if your gel slice contained LB the mixture color may change due to the LB pigment and not because of a pH change, so it is not necessary to add sodium acetate.
9. Add 1 gel volume of isopropanol to the sample and mix by inverting.
10. Place a MinElute spin column in a provided 2 ml collection tube.
11. Apply sample to the MinElute column and centrifuge for 1 min at 10,000 rpm.
12. Discard flow-through and place the MinElute column back into the same collection tube. For sample volumes of more than 800 µl, simply load and spin again.
13. Add 500 µl Buffer QG to the MinElute column and centrifuge* for 1 min at 10,000 rpm.
14. Discard flow-through and place the MinElute column back into the same collection tube.
15. Add 750 µl Buffer PE to MinElute column. Let the column stand 2–5 min after addition of Buffer PE.
16. Centrifuge* for 1 min at 10,000 rpm.
17. Discard flow-through and place the MinElute column back into the same collection tube.
18. Centrifuge the column in a 2 ml collection tube (provided) for 1 min. Residual ethanol from Buffer PE will not be completely removed unless the flow-through is discarded before this additional centrifugation.

If more than one column was used to purify a gel extract from the same library, perform the following steps independently with each column in the same same eppendorf tube.

19. Place the MinElute column into a clean 1.5 ml eppendorf tube. To elute DNA, add 10 µl Buffer EB (10 mM Tris·Cl, pH 8.5) to the center of the MinElute membrane. (Ensure that the EB is dispensed directly onto the membrane for complete elution of bound DNA.)
20. Let the column stand for 1 min, and then centrifuge the column for 1 min.

* To increase the amount of DNA recovered it is advisable to increase the speed gradually. If the centrifuge does not has this option it can be done by first centrifuging at around 2,000 rpm for few seconds, then stopping it, centrifuging at around 5,000 rpm for few seconds, stopping again and finally centrifuging at the desired revolutions and time (10,000 rpm for 1 min in this case).

6. Preparing final template for Illumina sequencing

1. Use the Qubit to measure DNA concentration of the prepared library.
2. Perform an ethanol precipitation to increase concentration and remove excess salts. Note: the concentration of DNA in the precipitation solution (i.e. library solution + NaAc + 100% ethanol) should be a minimum of 1 ng/µl, otherwise it would not precipitate and will be lost.
   A. Add 1/10 volume 3 M sodium acetate 3M pH 4.8 or 5.2 (e.g. 2ul for 20 µl DNA solution)
   B. Add 2 volumes of 100% ethanol (molecular biology grade) stored at -20°C, chill in dry ice for 30 min or overnight in a -20ºC freezer.
   C. Centrifuge at max speed for 15 minutes, remove supernatant carefully.
   D. Add 200 µL 70% Ethanol (diluted from absolute, not technical)
   E. Centrifuge 10 minutes, remove supernatant
   F. Dry DNA Pellet
   G. Resuspend using 20-40 µL of Tris 10 mM or TE
3. Measure concentration again. A total concentration of >25 ng/µL is ideal for Illumina sequencing, but we can go as low as 2 ng/µL.
4. Make an aliquot of the library and submit it to Fragment Analyzer or Bioanalyzer. You should expect to see a curve with a peak in the middle of the range of the size selection. A peak around 130 bp indicates that there was primer dimer carry over. If the peak is small relatively to the library, it is possible to sequence the as it is, as they will represent a small percentage of the total reads.
5. If the Fragment Analyzer profile and concentration are the desired the library is now ready for sequencing. The library can be submitted for sequencing in a Illumina HiSeq2000 (or similar) system in a single or pair-end run. The index sequencing is done separately from the insert sequencing, and the index sequence is not effected by the insert length, so it is not necessary to run the pair end to get the indexes sequence. If you used this protocol with more than one index, then you will be asked by the sequencing facility to provide their ID so that they can demultiplex the reads by index. Then your pipeline will have to include a second demultiplexing step to separate the reads by individual.

Table 1. Adapters and primers sequences

Oligo sequences same as Parchman et al. protocol (all 5'-3' oriented):
EcoRI adapter 1: CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNC EcoRI adapter 2: AATTGNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT …where "NNNNNNN" is a unique 7bp sequence for each of 96 barcoded adapters. The barcodes were designed using the Python script at https://bioinf.eva.mpg.de/multiplex/ . Full barcode list in Table 2. ILLPCR1: A*A*TGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
New or modified oligos:
MseI-adap1-bar: TAAGATCGGAAGAGCACACGTCTGAACTCCAGTCA MseI-adap2-bar: CTGGAGTTCAGACGTGTGCTCTTCCGATCT ILLPCR2-bar04: C*A*AGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT ILLPCR2-bar06: C*A*AGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT ILLPCR2-bar12: C*A*AGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Table 2. Barcodes list included in the EcoRI adapters.

Barcode.ID	barcode
Barcode_1	TCAATATC
Barcode_2	GAATAGTC
Barcode_3	TTGACTCC
Barcode_4	TCTTCTGC
Barcode_5	TTCAACCC
Barcode_6	TTGAGGAC
Barcode_7	AATCAGTC
Barcode_8	GGCATATC
Barcode_9	ACCGCCTC
Barcode_10	GATTGATC
Barcode_11	AACTGCGC
Barcode_12	TGATCGCC
Barcode_13	GGCAAGGC
Barcode_14	TCGCAAGC
Barcode_15	TCCGGAAC
Barcode_16	ATACCGCC
Barcode_17	ACTTGAAC
Barcode_18	TATGCAGC
Barcode_19	GAAGCGCC
Barcode_20	GAGGTAGC
Barcode_21	CCGCTACC
Barcode_22	CAAGACCC

Barcode_23	CTCTCAGC
Barcode_24	AATCTCAC
Barcode_25	GCAGGATC
Barcode_26	GGTAGGTC
Barcode_27	CATCGTCC
Barcode_28	TTCAGAGC
Barcode_29	CTGCTGAC
Barcode_30	AGAGATTC
Barcode_31	CGCAATTC
Barcode_32	CGCTTGAC
Barcode_33	CCGTTCAC
Barcode_34	GCCGTCAC
Barcode_35	TTAGGCGC
Barcode_36	CGGTTAGC
Barcode_37	AGACGGAC
Barcode_38	TAGCATCC
Barcode_39	TTCCTGCC
Barcode_40	AATGATGC
Barcode_41	AGGAGGCC
Barcode_42	TTATCCTC
Barcode_43	ACTCTAGC
Barcode_44	GGCCATCC
Barcode_45	CAGAGTTC
Barcode_46	ATCATCAC
Barcode_47	GAACTTGC
Barcode_48	CGCGGAGC
Barcode_49	TGCCAGAC
Barcode_50	TCTCTTAC
Barcode_51	GGTCGACC
Barcode_52	GCTCTCCC
Barcode_53	GGATATAC
Barcode_54	GGACTCAC
Barcode_55	TCTATCGC
Barcode_56	GACGGTAC
Barcode_57	GTTCATAC
Barcode_58	ACTACGAC
Barcode_59	AGCTTCTC
Barcode_60	ACCGAGGC
Barcode_61	TATACTAC
Barcode_62	GGTATTGC
Barcode_63	CCGTCTTC
Barcode_64	CTGGAATC
Barcode_65	TTCCGCAC
Barcode_66	CAATCATC
Barcode_67	AAGCGAGC
Barcode_68	GAATGCCC

Barcode_69	CGGAAGAC
Barcode_70	AGGAATGC
Barcode_71	CGGTATCC
Barcode_72	GGAGTACC
Barcode_73	CTAGTCTC
Barcode_74	ATGACGGC
Barcode_75	TAGGACTC
Barcode_76	GCAACTTC
Barcode_77	GCGTCGCC
Barcode_78	AATGGCTC
Barcode_79	TCAACGGC
Barcode_80	GTATCGGC
Barcode_81	ATGGCAAC
Barcode_82	TTCGGTCC
Barcode_83	CGTACGGC
Barcode_84	TCAAGCAC
Barcode_85	CATTATTC
Barcode_86	AACTCGAC
Barcode_87	CCTGGACC
Barcode_88	CTGGCTGC
Barcode_89	CTTACCTC
Barcode_90	CTACCTTC
Barcode_91	GTCCTCTC
Barcode_92	TGGTTCCC
Barcode_93	ACCTACCC
Barcode_94	CTATGAAC
Barcode_95	AAGGAACC
Barcode_96	ACGCAGAC

In this set of 96, all of the barcodes are separated from each other by at least 3 substitutions. The last base (C) is common to all barcodes as it was used as protective base. Barcodes were generated using the script available at https://bioinf.eva.mpg.de/multiplex/ with an edit distance of 3, and excluded any potential barcodes that contained an EcoRI or MseI cut site.

III. Sequencing Quality Control Report for each lane

(digital version only)

UHTS-LGTF

Quality Control Report

Fri Jan 25 12:26:37 CET 2013

Summary

Run Date : 18.01.2013	Laboratory : Nadir Alvarez
Instrument : HiSeq 2000	Library name : BERL1
Run number : 0001	Ref. organism : Other
Flowcell : AD1KUMACXX	Loaded quantity : 7.5 pM/lib
Lane number : 3	Protocol : Custom
Run type : single read 100 cycles	

Notes :

- Post processing done using Illumina pipeline Casava 1.82
- Fastq file : contains passed and not passed filter reads ; in header « Y » stand for « sequence is filter OUT ». Quality score is encoded in ASCII -33 (Sanger).
- For quality control purpose, each lane is spiked with approx. 5% of phix genomique sequences. Your data result may contain trace of phix and ecoli genomique sequences (57.42 % + 3.96 %)

Basic Statistics

Measure	Value
Filename	BERL1_NoIndex_L003_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	102345053
Filtered Sequences	2916589
Sequence length	101
%GC	43

Sequence diversity

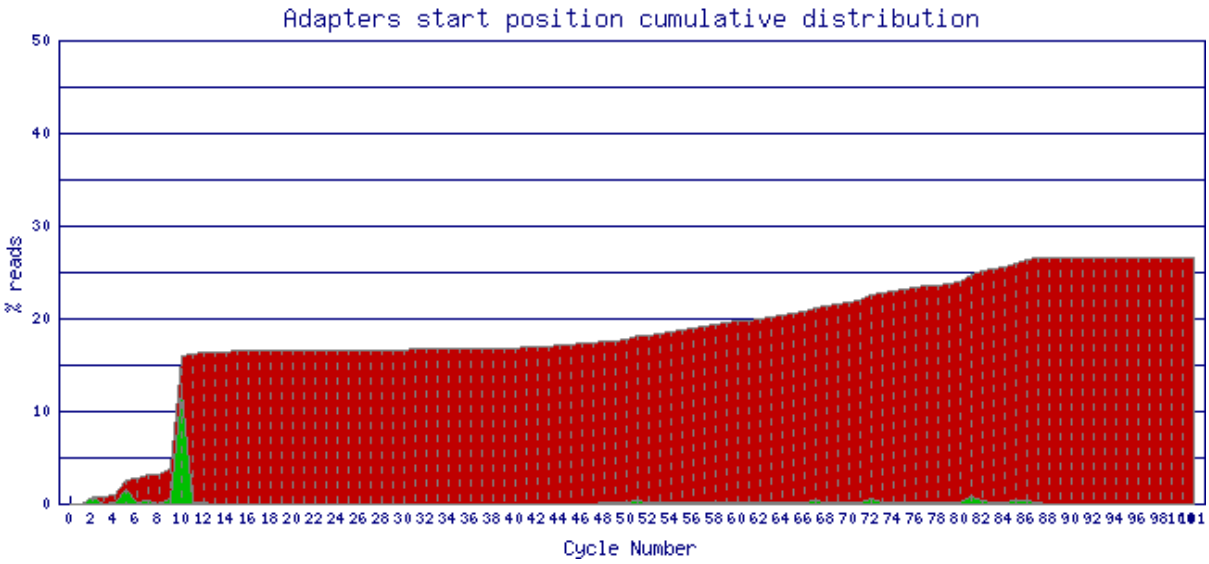
User Library [BERL1] :

- 0.43 % of PF reads
- diversity ratio (different/all):0.89

Sequence identification

Note : megablast versus embl_db on a subset of 100'000 PF reads

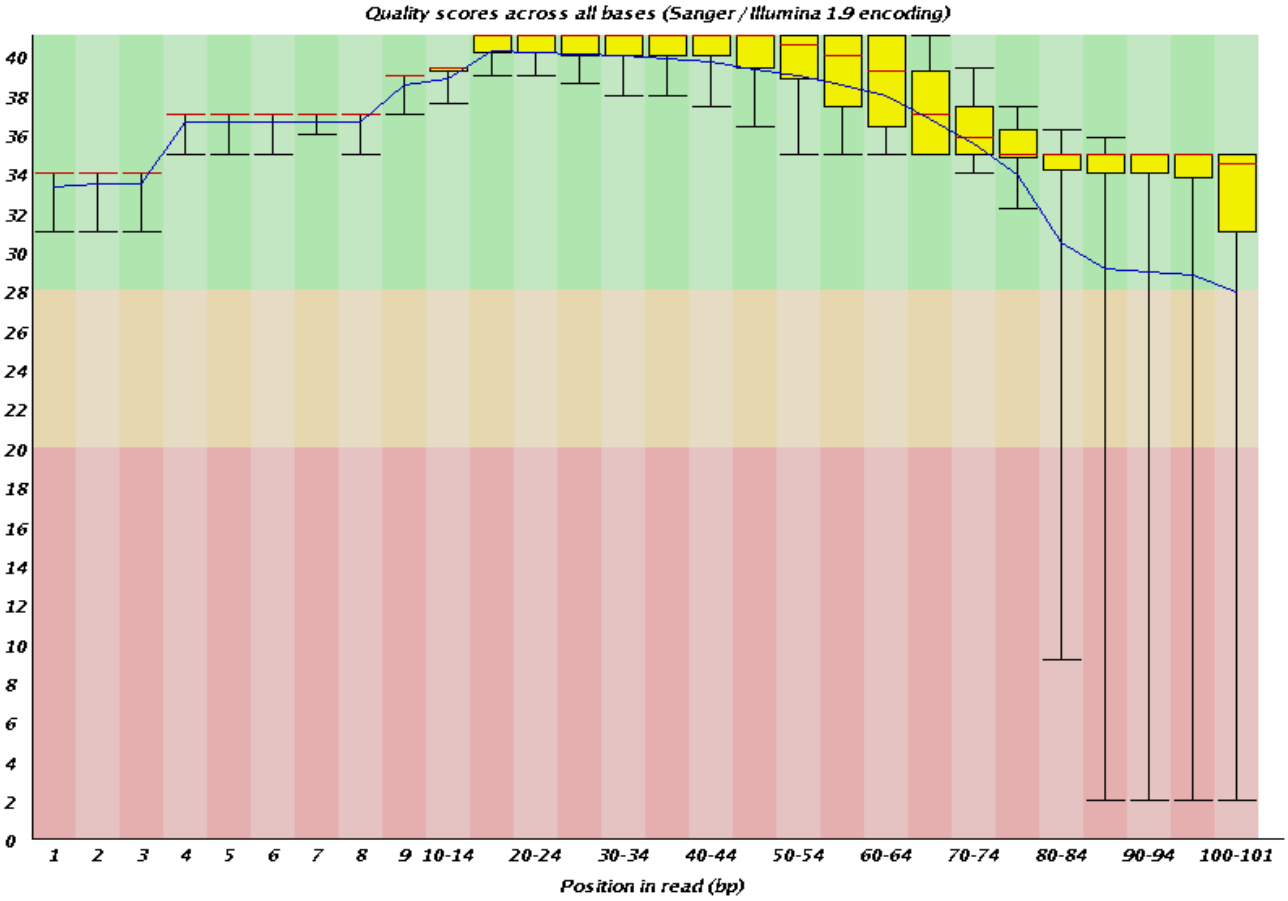
Adapters



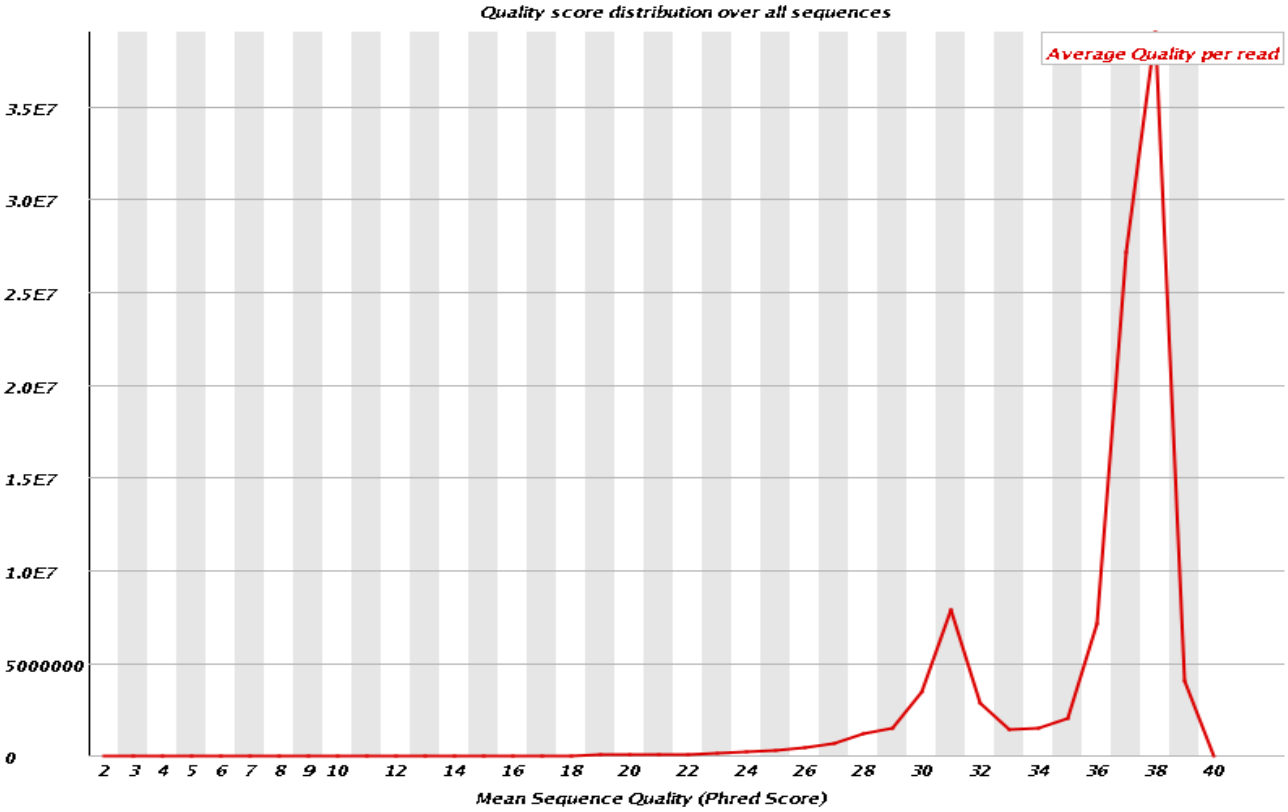
Non specifique sequences

ncRNA db	miRNA db	rRNA db	Repeat db	Mito and chloro db	cloning vector db	Adapter db	EMBL db
8.78 %	11.05 %	0.80 %	3.13 %	0.83 %	5.27 %	26.42 %	65.89 %

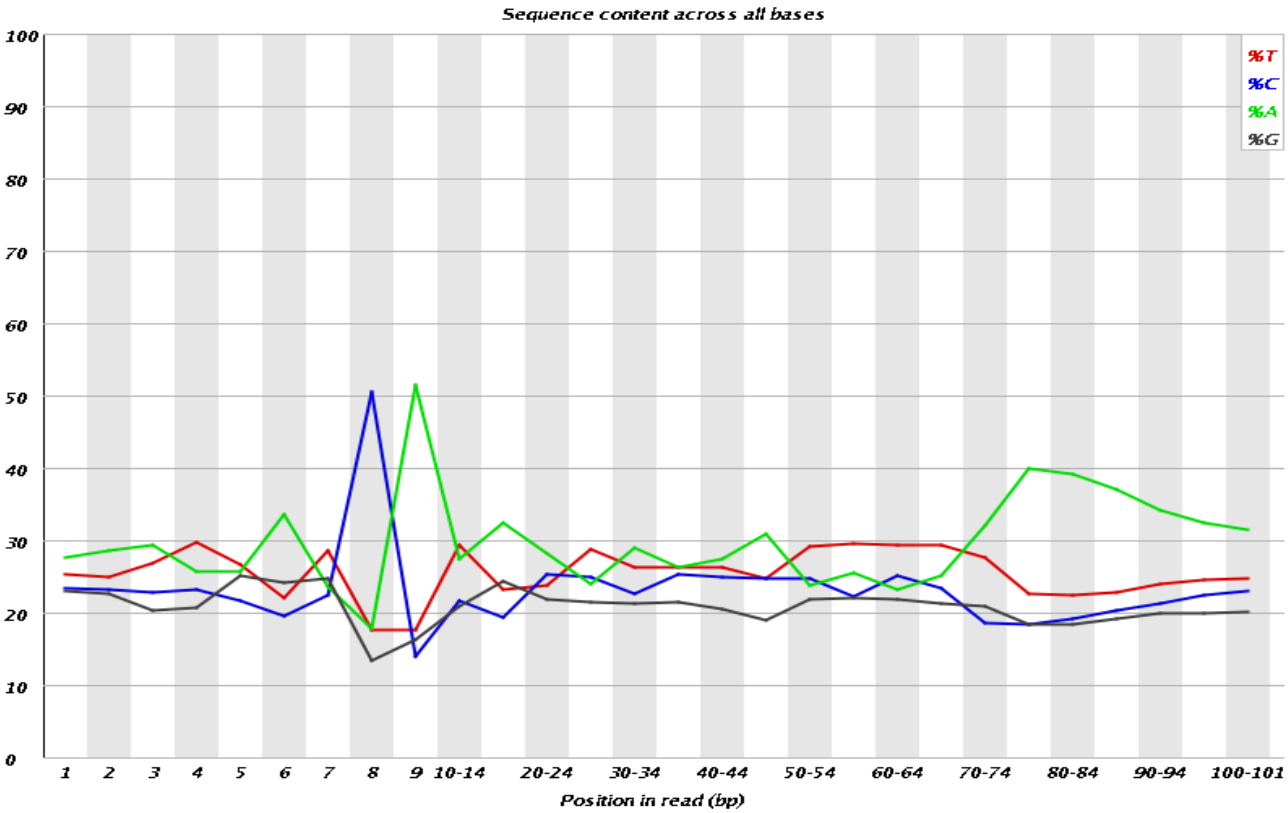
Per base sequence quality



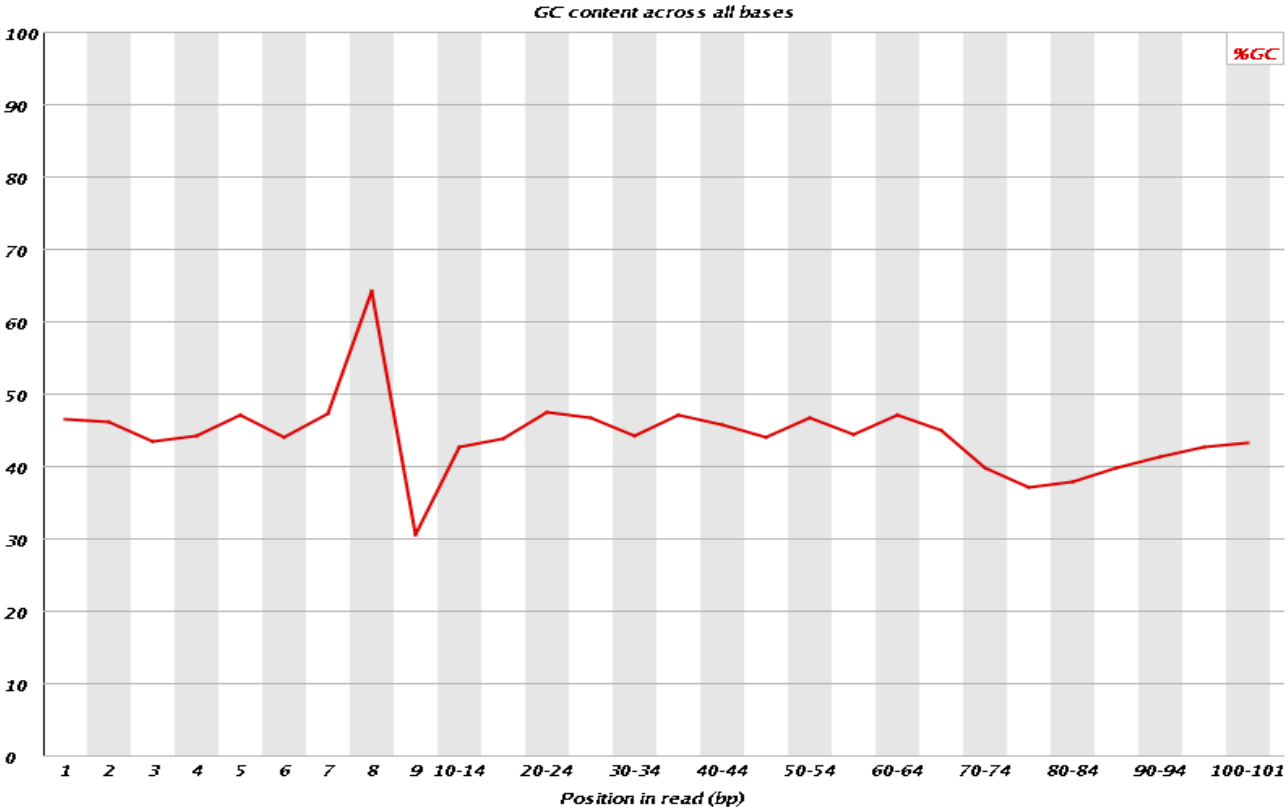
Per sequence quality scores



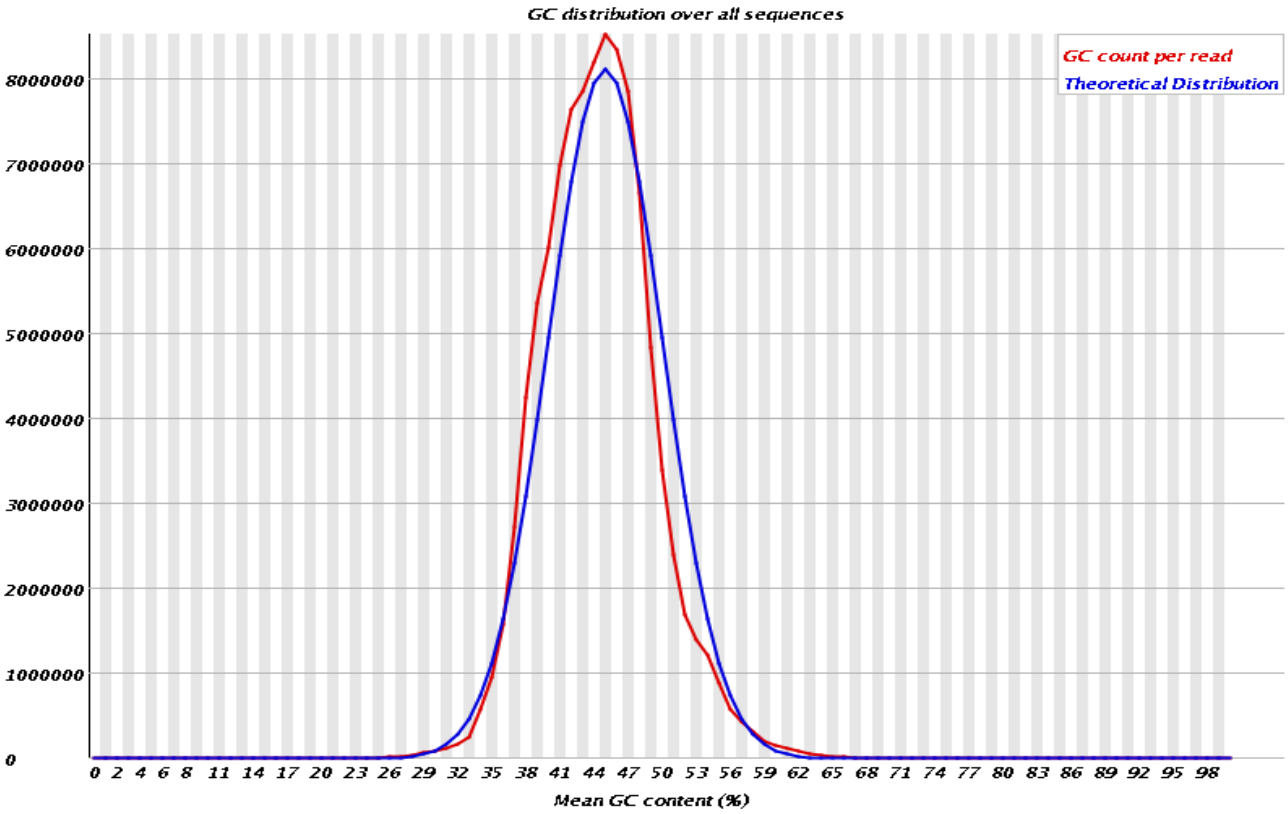
Per base sequence content



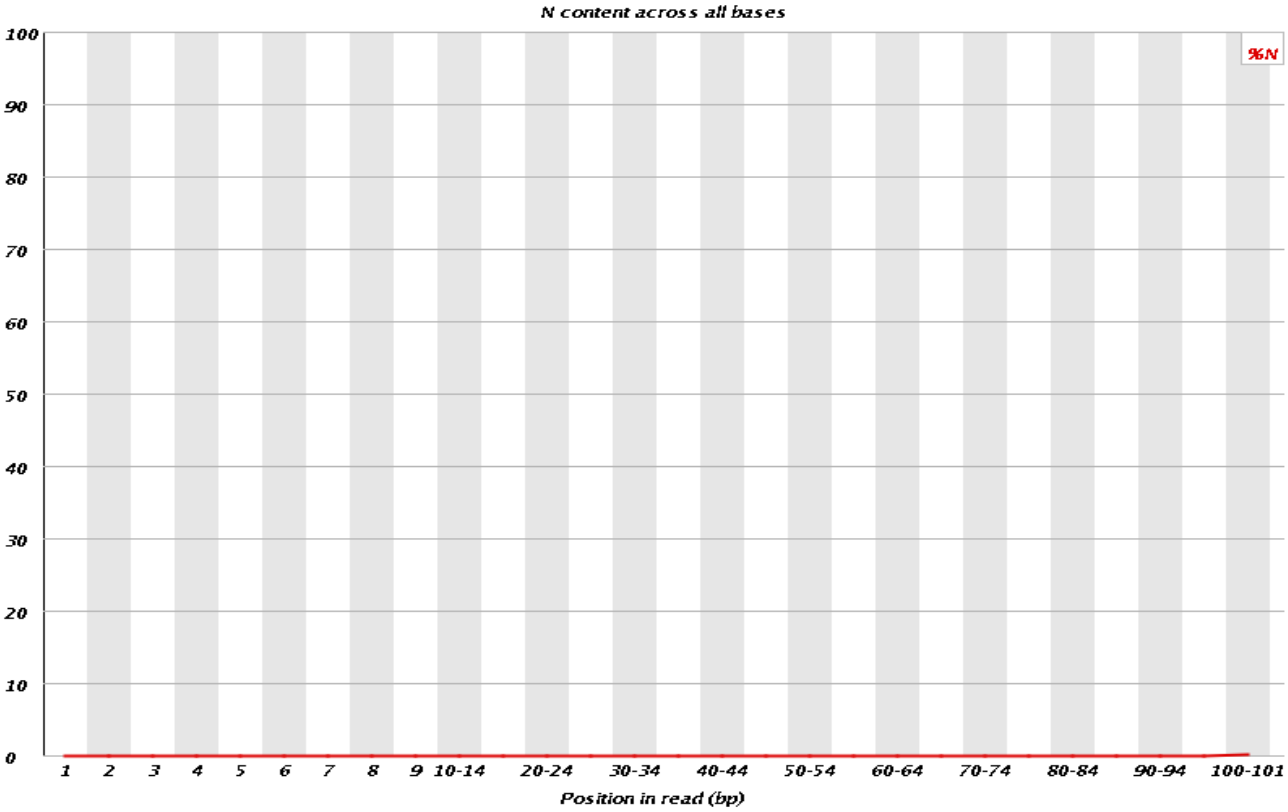
Per base GC content



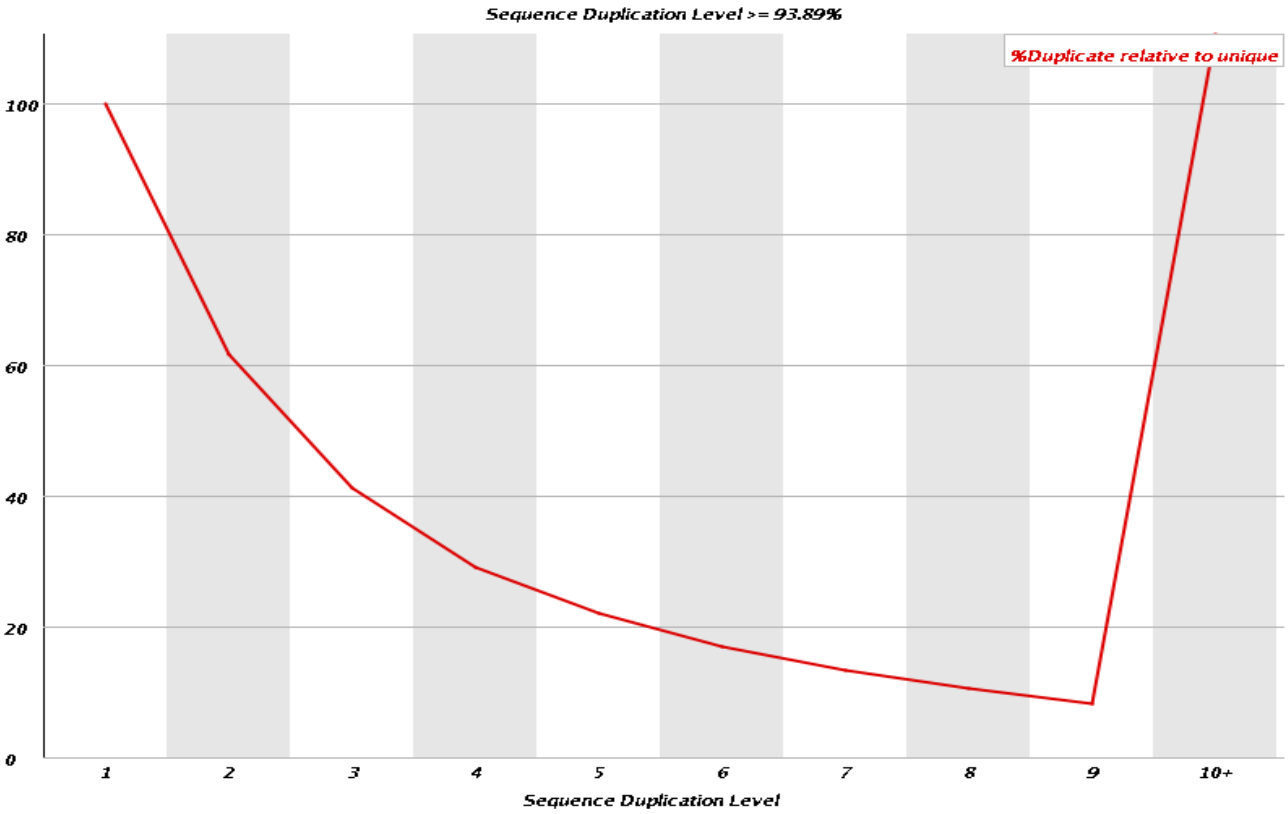
Per sequence GC content



Per base N content



Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CCGCTACCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	1253970	1.2252375305331074	TruSeq Adapter, Index 4 (100% over 41bp)
ACTTGAACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	919756	0.8986814438407688	TruSeq Adapter, Index 4 (100% over 41bp)
GCAGGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGT	806083	0.787613056392672	TruSeq Adapter, Index 4 (100% over 46bp)
AATCTCACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	714314	0.6979467781408056	TruSeq Adapter, Index 4 (100% over 41bp)
CTGCTGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	664327	0.6491051404311647	TruSeq Adapter, Index 4 (100% over 41bp)
TCGCAAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	623142	0.6088638207066052	TruSeq Adapter, Index 4 (100% over 41bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATG	567425	0.5544234756515295	TruSeq Adapter, Index 4 (100% over 49bp)
ATACCGCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	542281	0.5298556052337967	TruSeq Adapter, Index 4 (100% over 41bp)
AGAGATTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	541562	0.5291530798269263	TruSeq Adapter, Index 4 (100% over 41bp)
GCAGGATCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	497158	0.4857665177035963	TruSeq Adapter, Index 4 (100% over 41bp)

GATTGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGT	495064	0.4837204979511809	TruSeq Adapter, Index 4 (100% over 46bp)
AACTGCGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	454505	0.44409083456139303	TruSeq Adapter, Index 4 (100% over 41bp)
GAGGTAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	449290	0.43899532691628973	TruSeq Adapter, Index 4 (100% over 41bp)
CGCTTGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	436819	0.4268100774738961	TruSeq Adapter, Index 4 (100% over 41bp)
TGATCGCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	416600	0.40705435953020613	TruSeq Adapter, Index 4 (100% over 41bp)
TTGACTCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	401040	0.3918508889726209	TruSeq Adapter, Index 4 (100% over 41bp)
CTCTCAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	387950	0.3790608228030328	TruSeq Adapter, Index 4 (100% over 41bp)
CGCAATTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	341305	0.3334846091681637	TruSeq Adapter, Index 4 (100% over 41bp)
TATGCAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	319436	0.3121166980098198	TruSeq Adapter, Index 4 (100% over 41bp)
CAAGACCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	276472	0.2701371408738242	TruSeq Adapter, Index 4 (100% over 41bp)
TTCAGAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	259570	0.25362241983498707	TruSeq Adapter, Index 4 (100% over 41bp)

CATCGTCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	257972	0.2520610351337646	TruSeq Adapter, Index 4 (100% over 41bp)
TCAATATCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	242679	0.23711844675091426	TruSeq Adapter, Index 4 (100% over 41bp)
GGTAGGTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	200776	0.196175578706281	TruSeq Adapter, Index 4 (100% over 41bp)
GGTAGGTCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGT	169919	0.1660256114186584	TruSeq Adapter, Index 4 (97% over 46bp)
TTGAGGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	162903	0.15917037045259042	TruSeq Adapter, Index 4 (100% over 41bp)
GAATAGTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	145404	0.1420723285960876	TruSeq Adapter, Index 4 (100% over 41bp)
GATTGATCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	136414	0.1332883182932154	TruSeq Adapter, Index 4 (100% over 41bp)
AGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTA	131391	0.12838041131309005	TruSeq Adapter, Index 4 (100% over 47bp)
TCTTCTGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	129578	0.1266089529505642	TruSeq Adapter, Index 4 (100% over 41bp)
ACCGCCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	122115	0.1193169541863445	TruSeq Adapter, Index 4 (100% over 41bp)
GGCAAGGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	104234	0.10184566517347936	TruSeq Adapter, Index 4 (100% over 41bp)

Produced using part of FastQC (version 0.10.0)

UHTS-LGTF

Quality Control Report

Wed Mar 6 18:46:25 CET 2013

Summary

Run Date : 28.02.2013	Laboratory : Nadir Alvarez
Instrument : HiSeq 2000	Library name : BERL2
Run number : 0094	Ref. organism : Other
Flowcell : AD1WB0ACXX	Loaded quantity : 17 pM/lib
Lane number : 2	Protocol : Custom
Run type : single read 100 cycles	

Notes :

- Post processing done using Illumina pipeline Casava 1.82
- Fastq file : contains passed and not passed filter reads ; in header « Y » stand for « sequence is filter OUT ». Quality score is encoded in ASCII -33 (Sanger).
- For quality control purpose, each lane is spiked with approx. 5% of phix genomique sequences. Your data result may contain trace of phix and ecoli genomique sequences (55.62 % + 3.82 %)

Basic Statistics

Measure	Value
Filename	BERL2_NoIndex_L002_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	99483564
Filtered Sequences	3487258
Sequence length	101
%GC	44

Sequence diversity

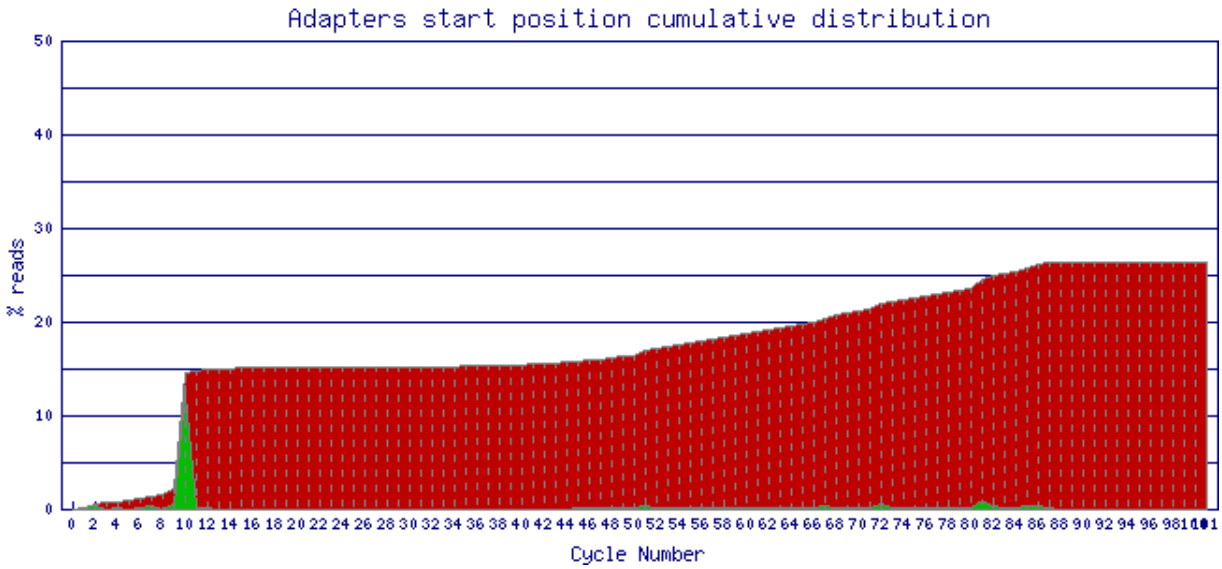
User Library [BERL2] :

- 3.92 % of PF reads
- diversity ratio (different/all):0.48

Sequence identification

Note : megablast versus embl_db on a subset of 100'000 PF reads

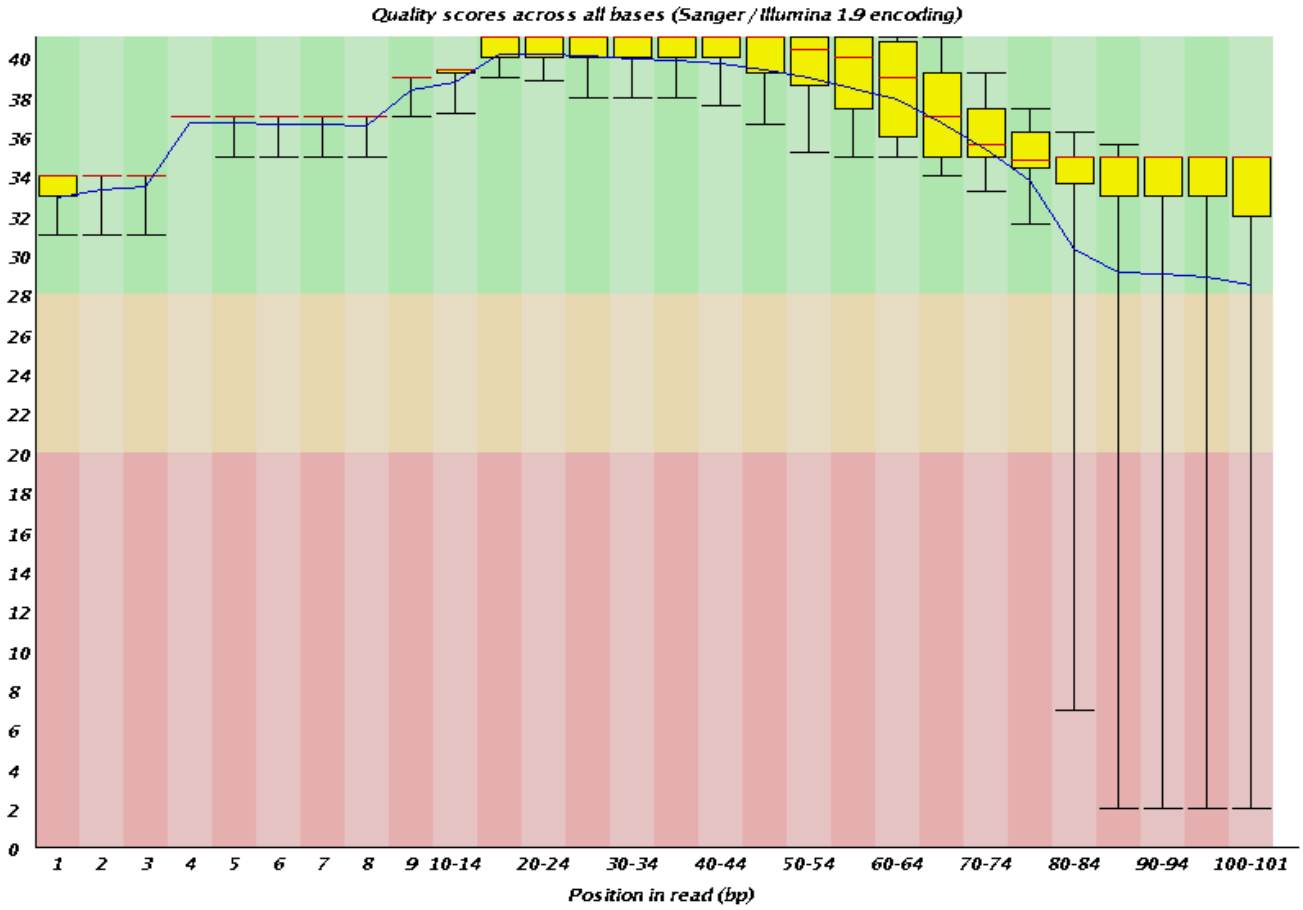
Adapters



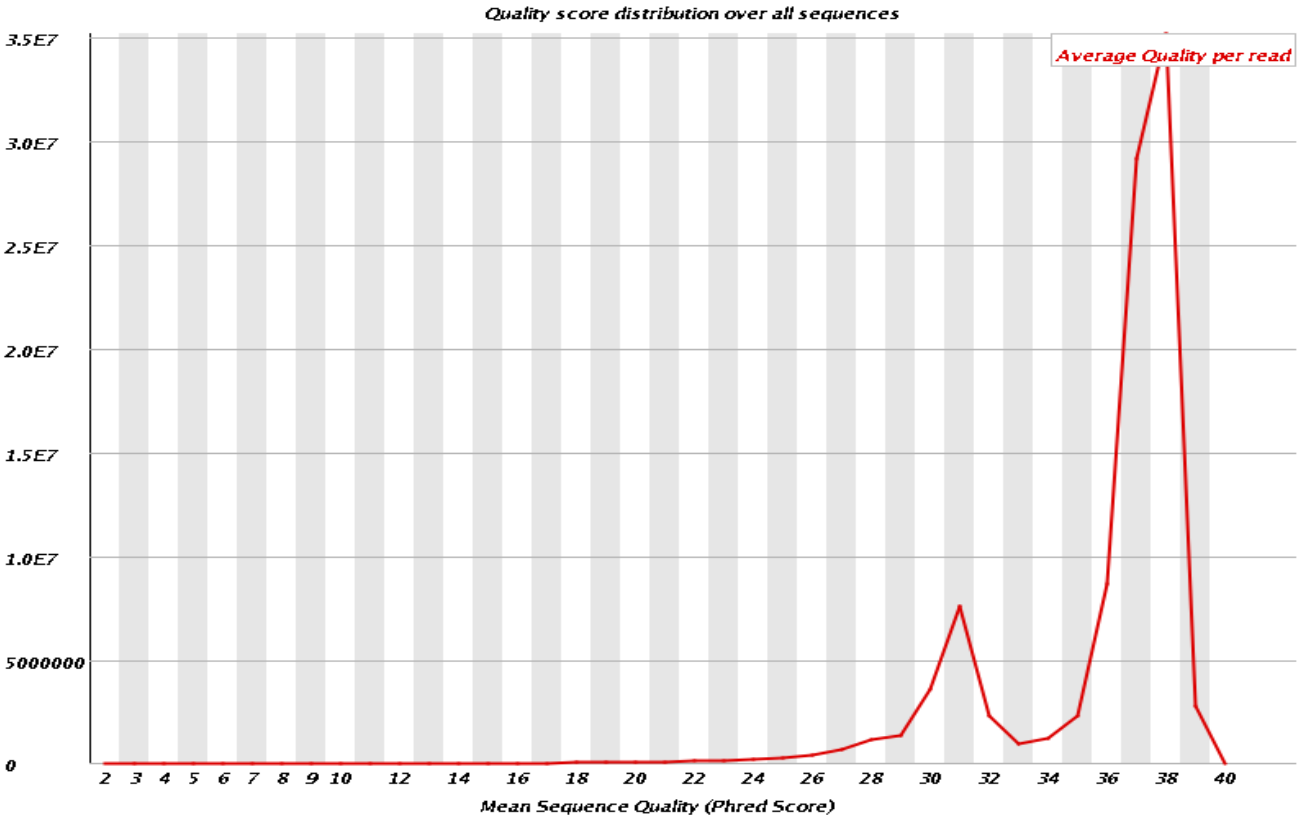
Non specifique sequences

ncRNA db	miRNA db	rRNA db	Repeat db	Mito and chloro db	cloning vector db	Adapter db	EMBL db
5.38 %	0.05 %	0.95 %	0.56 %	0.78 %	3.12 %	26.39 %	67.73 %

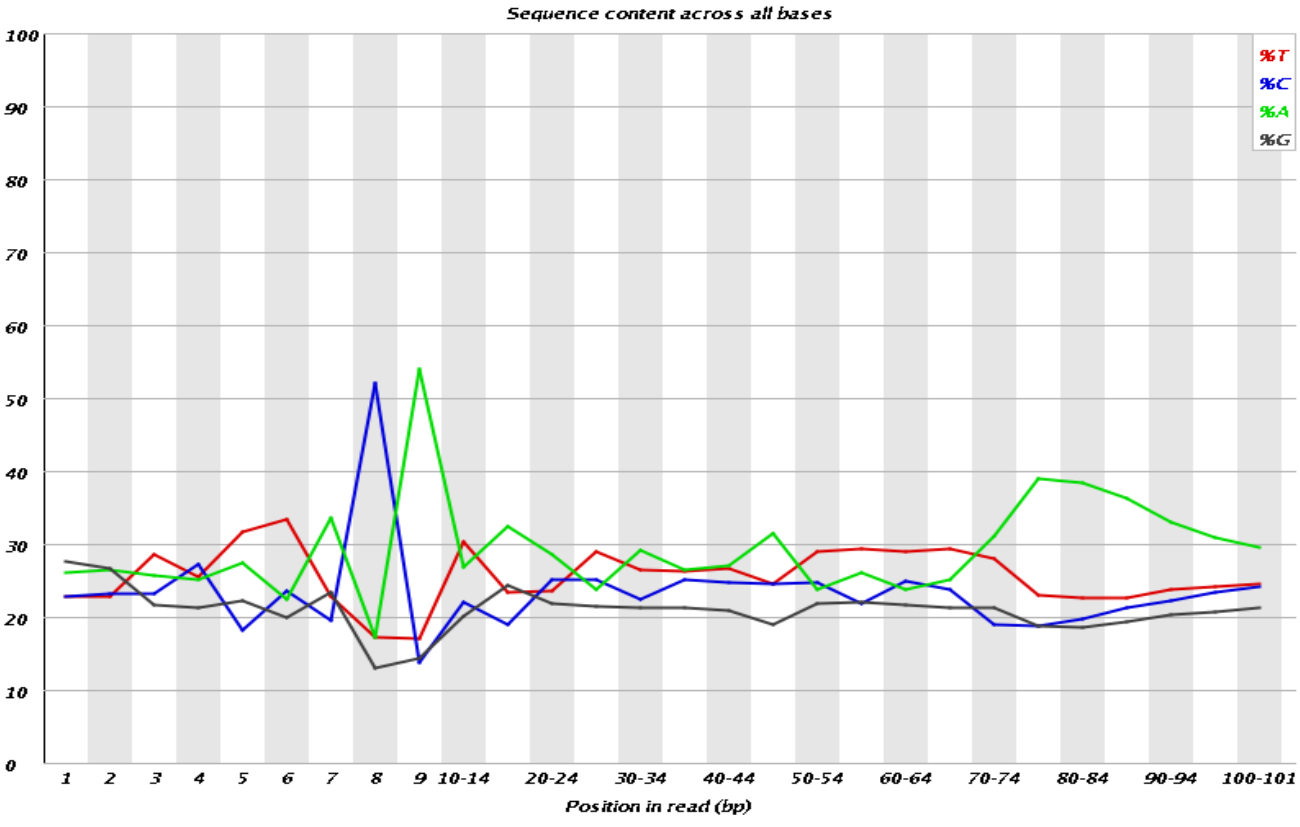
Per base sequence quality



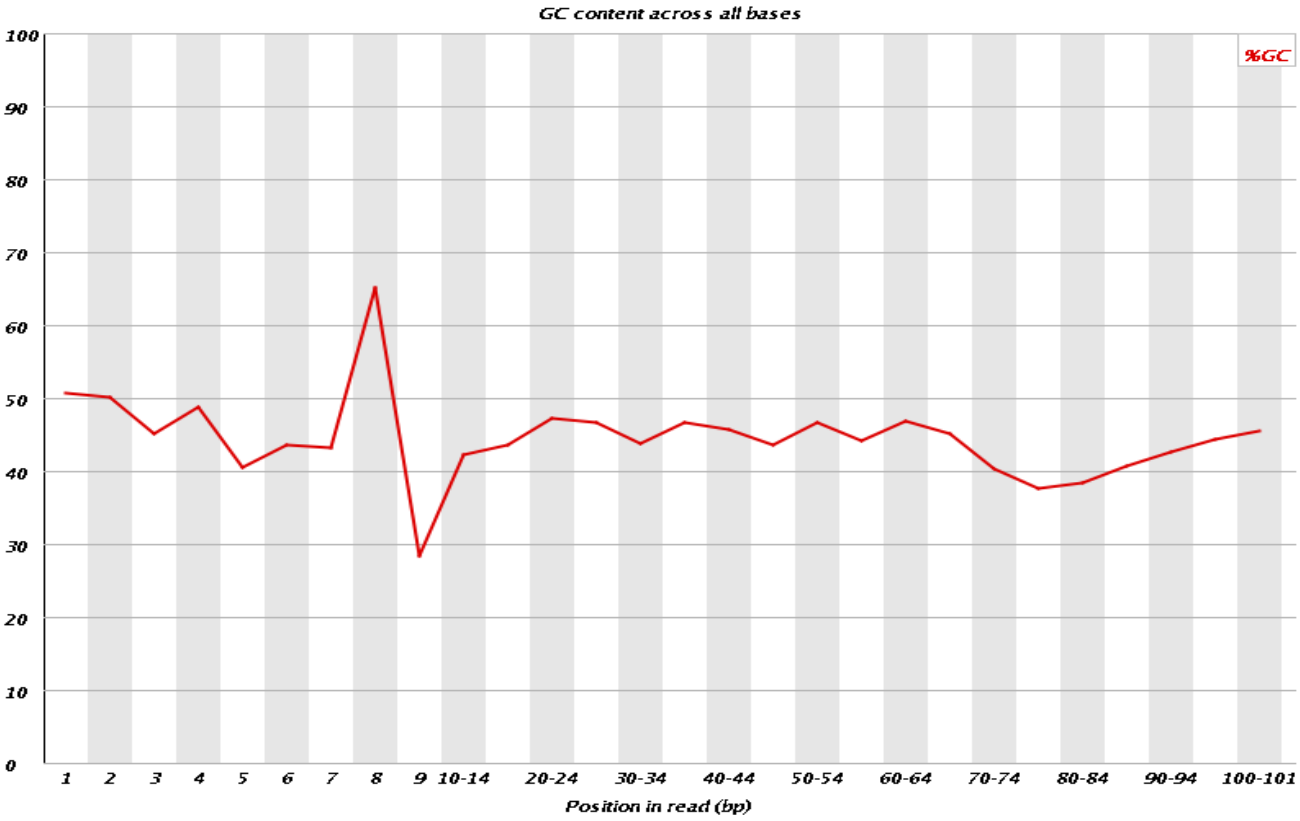
Per sequence quality scores



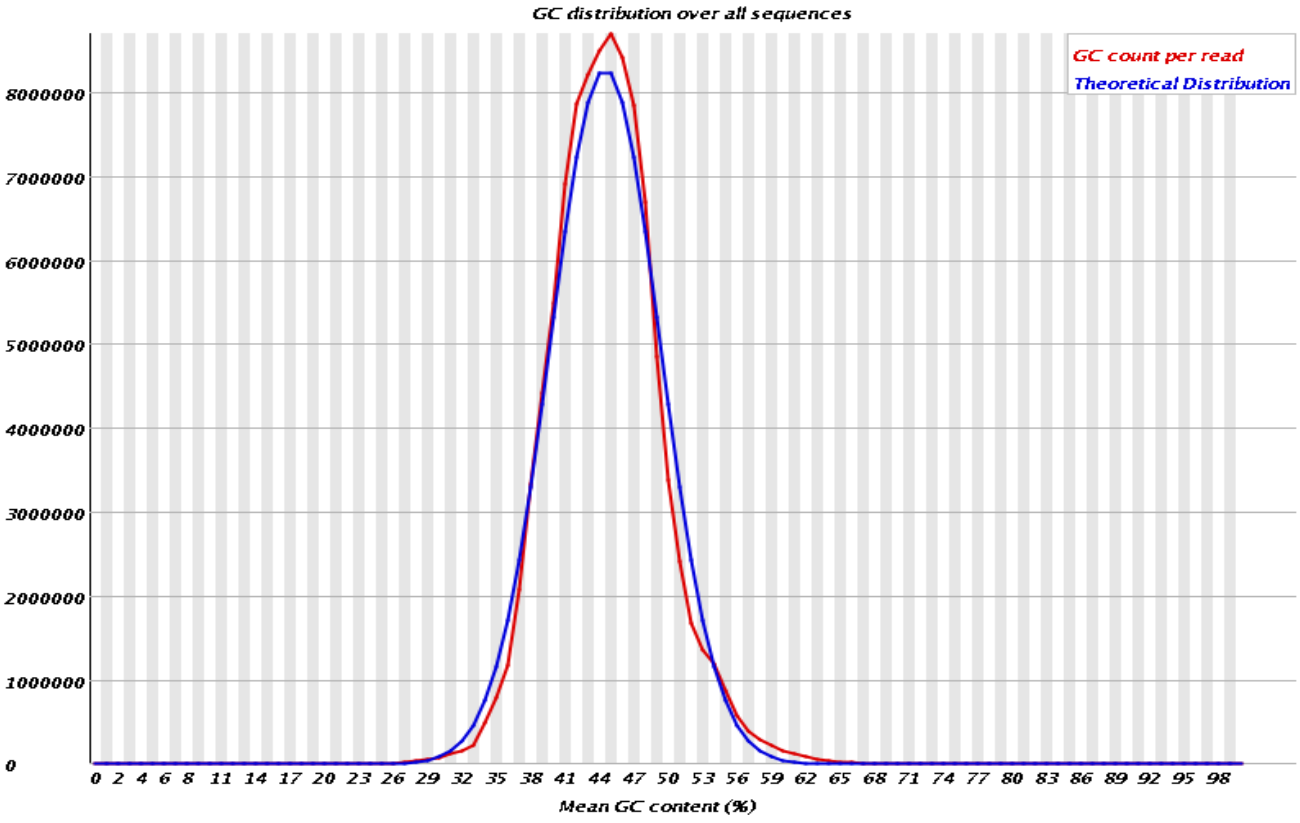
Per base sequence content



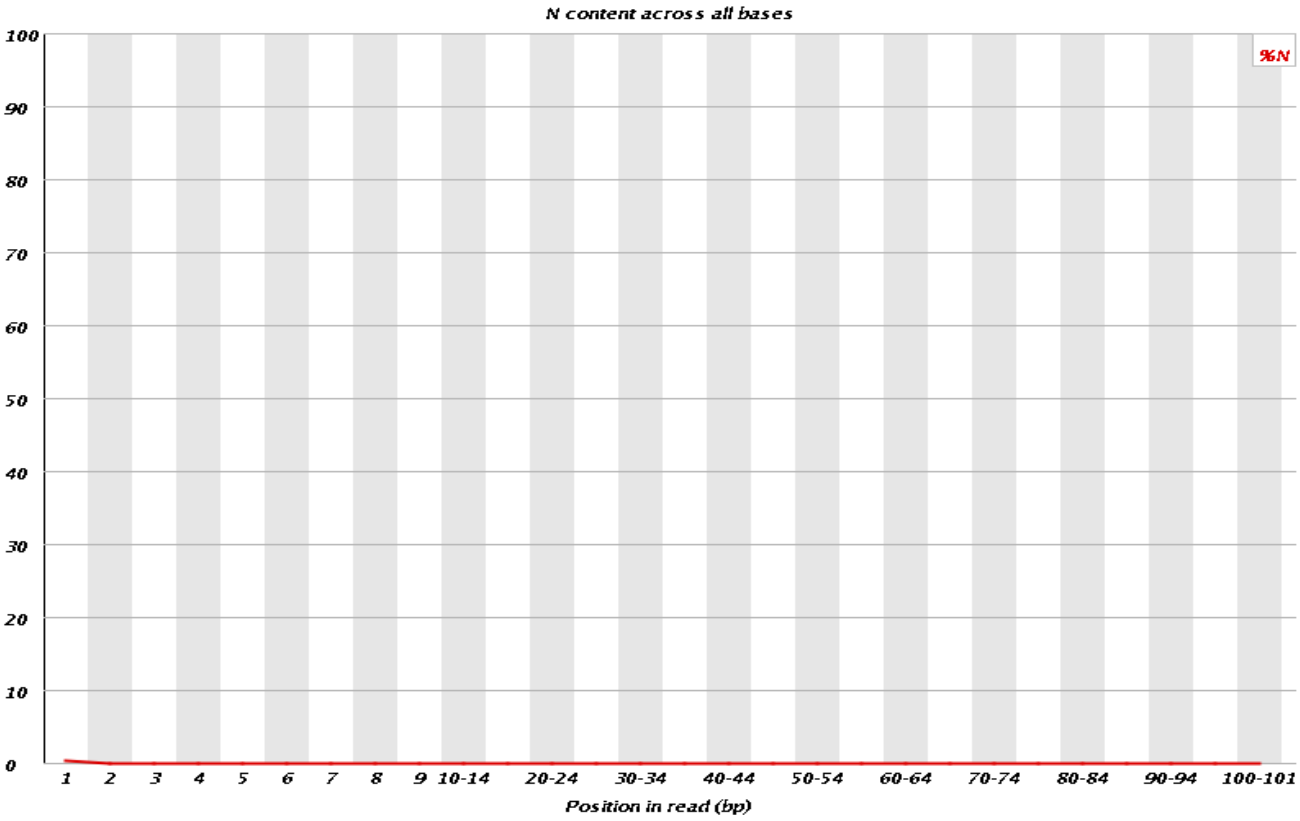
Per base GC content



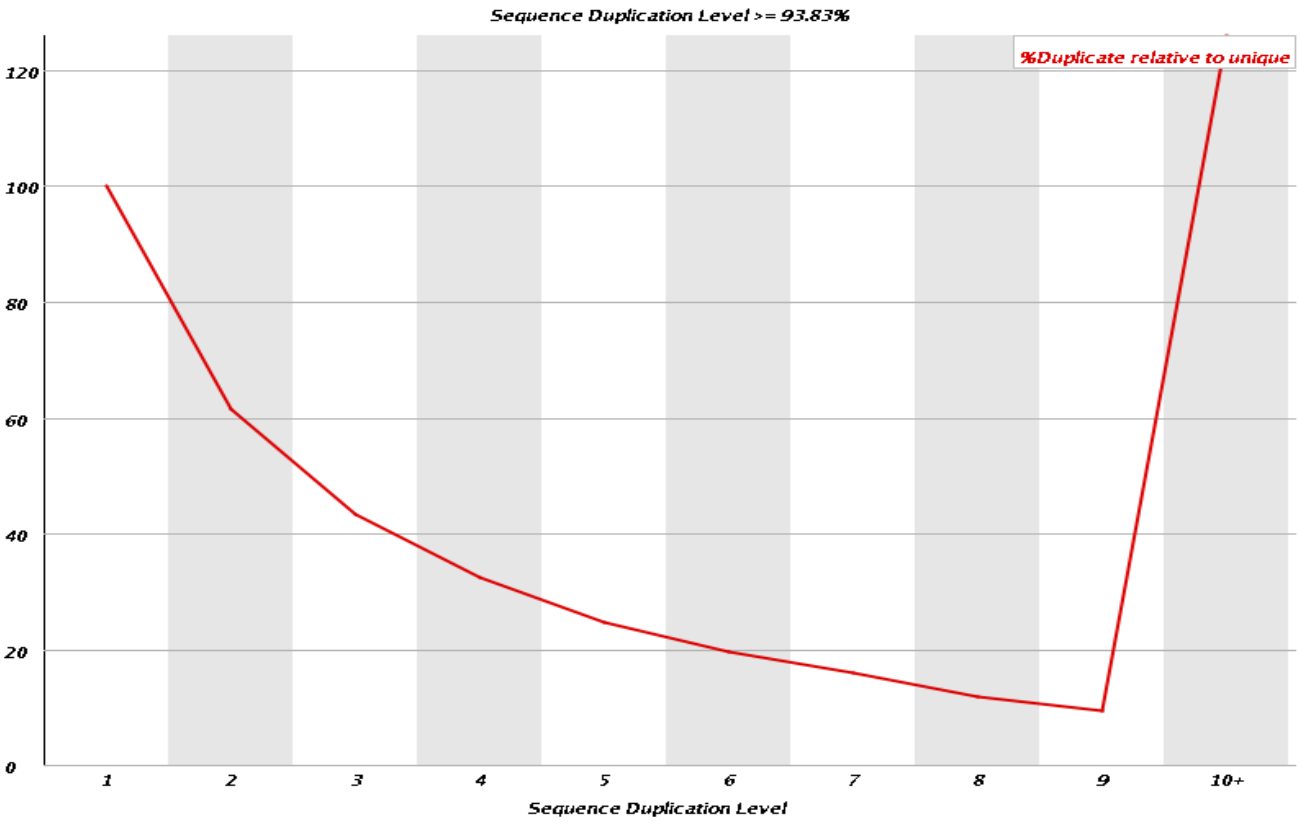
Per sequence GC content



Per base N content



Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGATATACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	774198	0.7782169927084639	TruSeq Adapter, Index 4 (100% over 41bp)
TATACTACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	753567	0.7574788936994657	TruSeq Adapter, Index 4 (100% over 41bp)
AGACGGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	682859	0.6864038365171558	TruSeq Adapter, Index 4 (100% over 41bp)
ACTACGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	582341	0.585364030584992	TruSeq Adapter, Index 4 (100% over 41bp)
GTTCATACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	567961	0.5709093815738246	TruSeq Adapter, Index 4 (100% over 41bp)
GGACTCACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	521290	0.523996104522351	TruSeq Adapter, Index 4 (100% over 41bp)
CCGTCTTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	497681	0.5002645462118748	TruSeq Adapter, Index 4 (100% over 41bp)
TGCCAGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	482777	0.4852831770281169	TruSeq Adapter, Index 4 (100% over 41bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATG	459437	0.46182201514211935	TruSeq Adapter, Index 4 (100% over 49bp)
CGCGGAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	454366	0.4567246907237862 4	TruSeq Adapter, Index 4 (100% over 41bp)

CAGAGTTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	419681	0.4218596350247364	TruSeq Adapter, Index 4 (100% over 41bp)
TTATCCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	416364	0.41852541591694487	TruSeq Adapter, Index 4 (100% over 41bp)
TCTATCGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	407106	0.4092193560737329	TruSeq Adapter, Index 4 (100% over 41bp)
GACGGTACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	391233	0.39326395664715025	TruSeq Adapter, Index 4 (100% over 41bp)
GCCGTCACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	389151	0.39117114863315516	TruSeq Adapter, Index 4 (100% over 41bp)
TAGCATCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	335679	0.3374215664408646	TruSeq Adapter, Index 4 (100% over 41bp)
GGTCGACCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	312773	0.31439665752224155	TruSeq Adapter, Index 4 (100% over 41bp)
CCGTTCACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	311472	0.3130889038112869	TruSeq Adapter, Index 4 (100% over 41bp)
CTGGAATCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	272654	0.27406939301048766	TruSeq Adapter, Index 4 (100% over 41bp)
TCTCTTACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	267066	0.2684523847577475	TruSeq Adapter, Index 4 (100% over 41bp)
CGGTTAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	246821	0.24810228954000885	TruSeq Adapter, Index 4 (100% over 41bp)

GAACTTGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	244091	0.24535811764845897	TruSeq Adapter, Index 4 (100% over 41bp)
CAGAGTTCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGT	236005	0.23723014185539232	TruSeq Adapter, Index 4 (97% over 46bp)
GGCCATCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	221512	0.22266190624212057	TruSeq Adapter, Index 4 (100% over 41bp)
ATCATCACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	221482	0.22263175050704856	TruSeq Adapter, Index 4 (100% over 41bp)
AGCTTCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	209435	0.210522212493312	TruSeq Adapter, Index 4 (100% over 41bp)
ACCGAGGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	178702	0.17962967229441038	TruSeq Adapter, Index 4 (100% over 41bp)
ACTCTAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	174087	0.17499071504917135	TruSeq Adapter, Index 4 (100% over 41bp)
GAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCT	168796	0.16967224857364377	TruSeq Adapter, Index 4 (100% over 50bp)
AGGAGGCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	149365	0.15014037896752472	TruSeq Adapter, Index 4 (100% over 41bp)
GCTCTCCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	138570	0.1392893402974586	TruSeq Adapter, Index 4 (100% over 41bp)
AATGATGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	104452	0.10499422799126899	TruSeq Adapter, Index 4 (100% over 41bp)

Produced using part of FastQC (version 0.10.0)

UHTS-LGTF

Quality Control Report

Mon Apr 15 09:14:53 CEST 2013

Summary

Run Date : 05.04.2013	Laboratory : Nadir Alvarez
Instrument : HiSeq 2000	Library name : BERL3
Run number : 0017	Ref. organism : Berberis (barberry)
Flowcell : BD21YUACXX	Loaded quantity : 16 pM/lib
Lane number : 4	Protocol : Custom
Run type : single read 100 cycles	

Notes :

- Post processing done using Illumina pipeline Casava 1.82
- Fastq file : contains passed and not passed filter reads ; in header « Y » stand for « sequence is filter OUT ». Quality score is encoded in ASCII -33 (Sanger).
- For quality control purpose, each lane is spiked with approx. 5% of phix genomique sequences. Your data result may contain trace of phix and ecoli genomique sequences (27.66 % + 1.86 %)

Basic Statistics

Measure	Value
Filename	BERL3_NoIndex_L004_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	96813840
Filtered Sequences	3584515
Sequence length	101
%GC	43

Sequence diversity

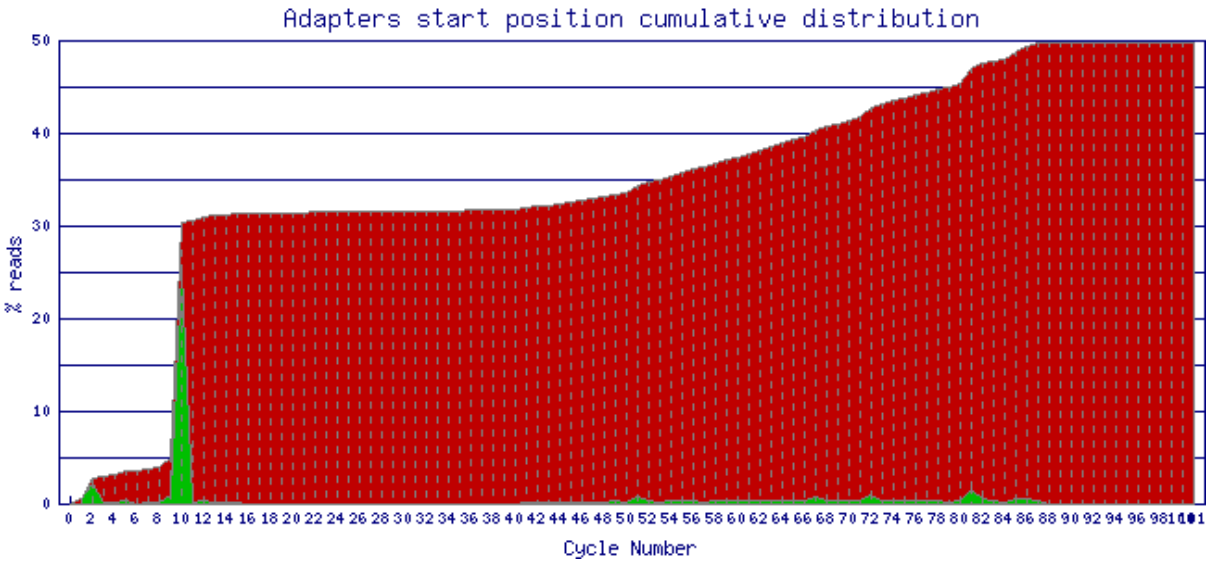
User Library [BERL3] :

- 11.42 % of PF reads
- diversity ratio (different/all):0.34

Sequence identification

Note : megablast versus embl_db on a subset of 100'000 PF reads

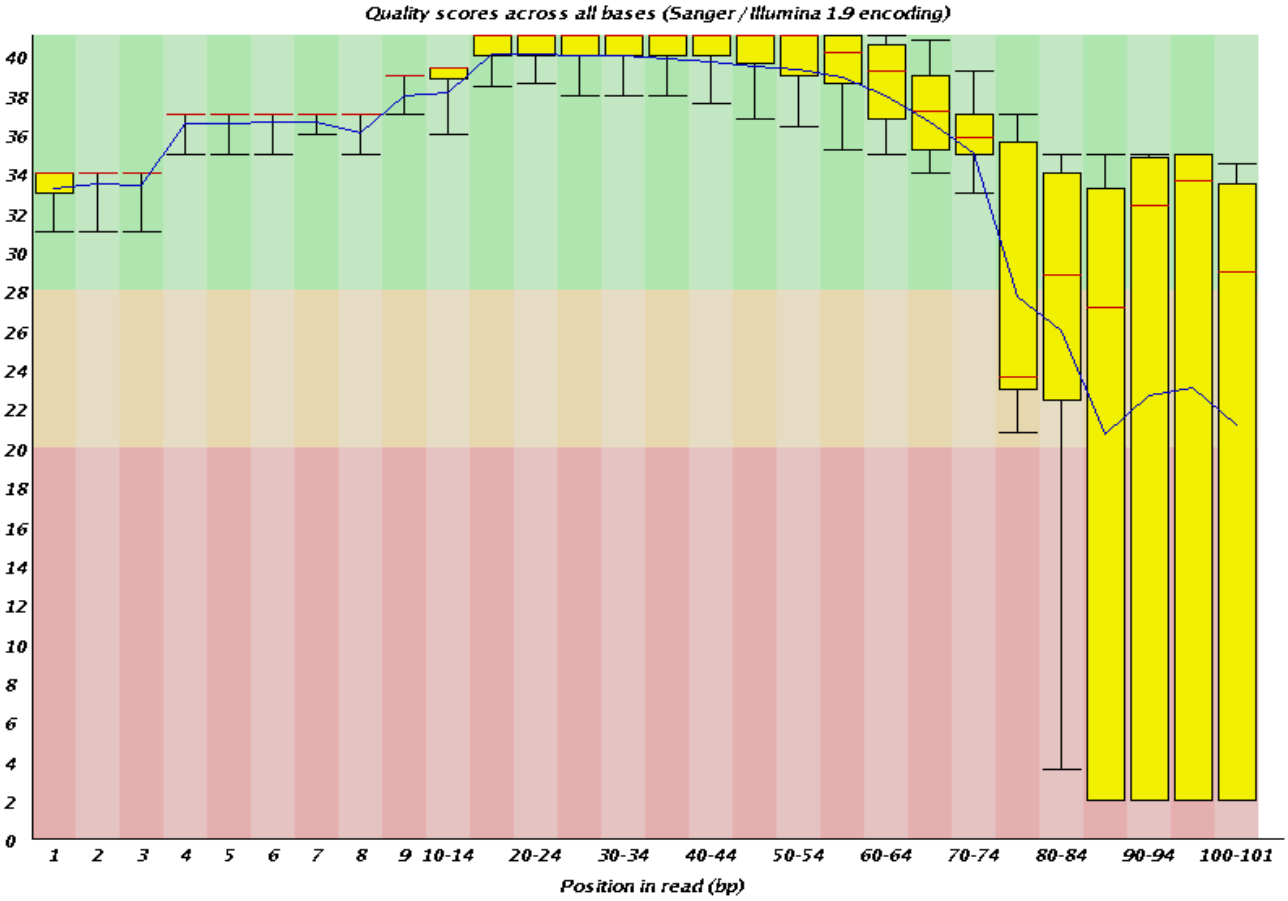
Adapters



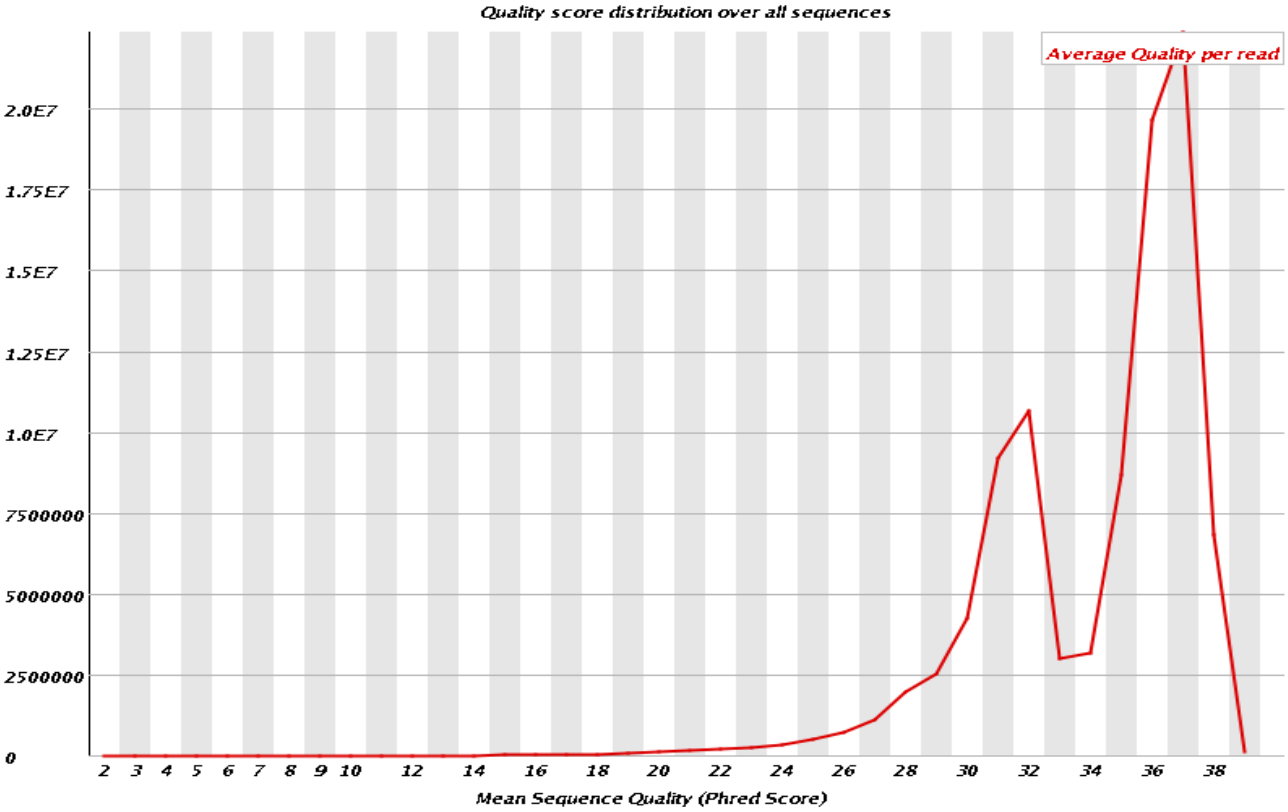
Non specifique sequences

ncRNA db	miRNA db	rRNA db	Repeat db	Mito and chloro db	cloning vector db	Adapter db	EMBL db
17.38 %	0.06 %	1.35 %	4.10 %	1.20 %	4.45 %	49.53 %	39.33 %

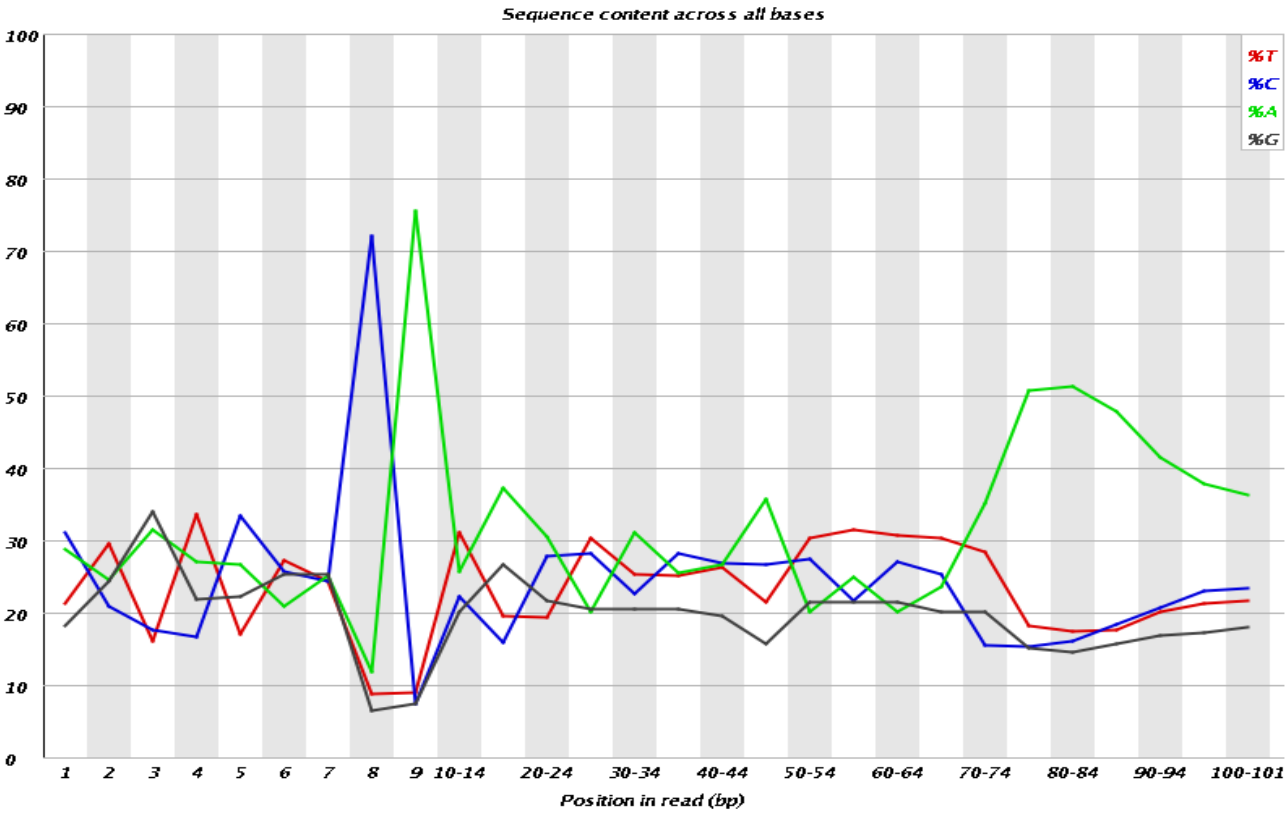
Per base sequence quality



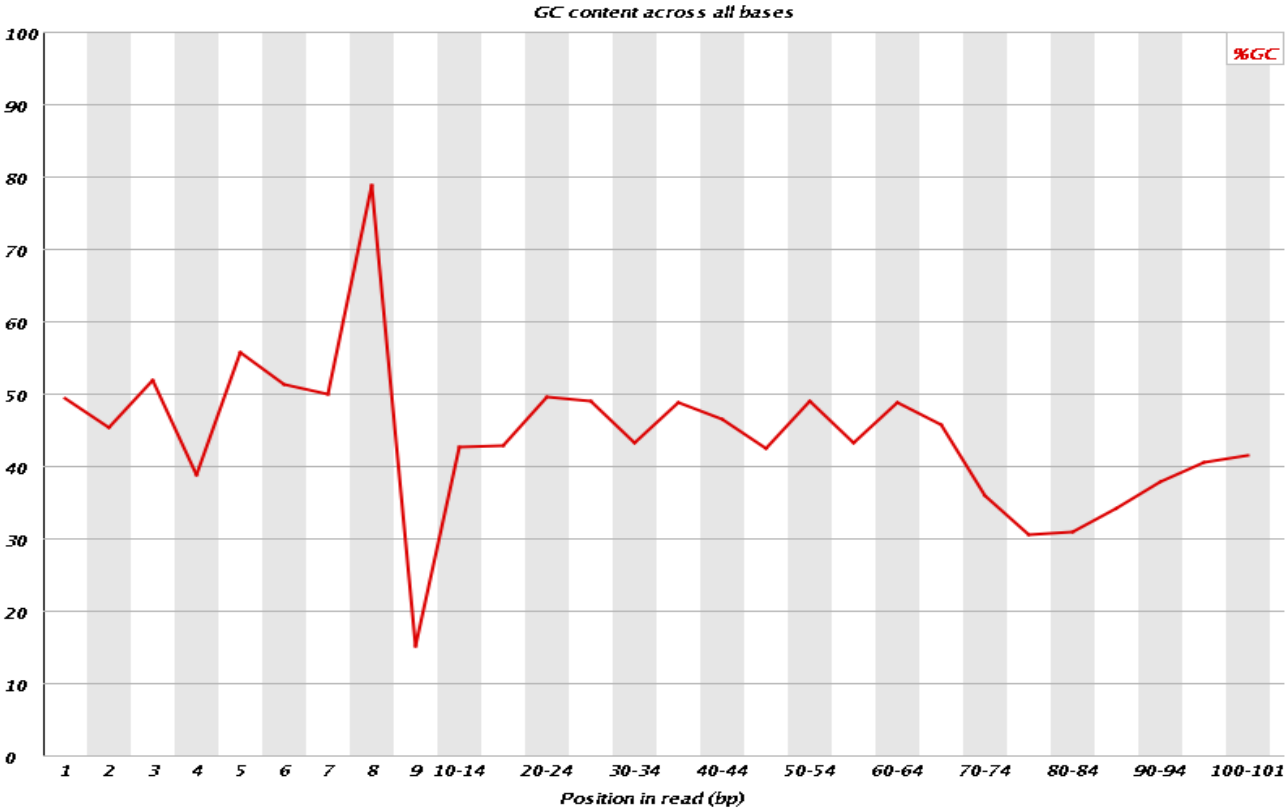
Per sequence quality scores



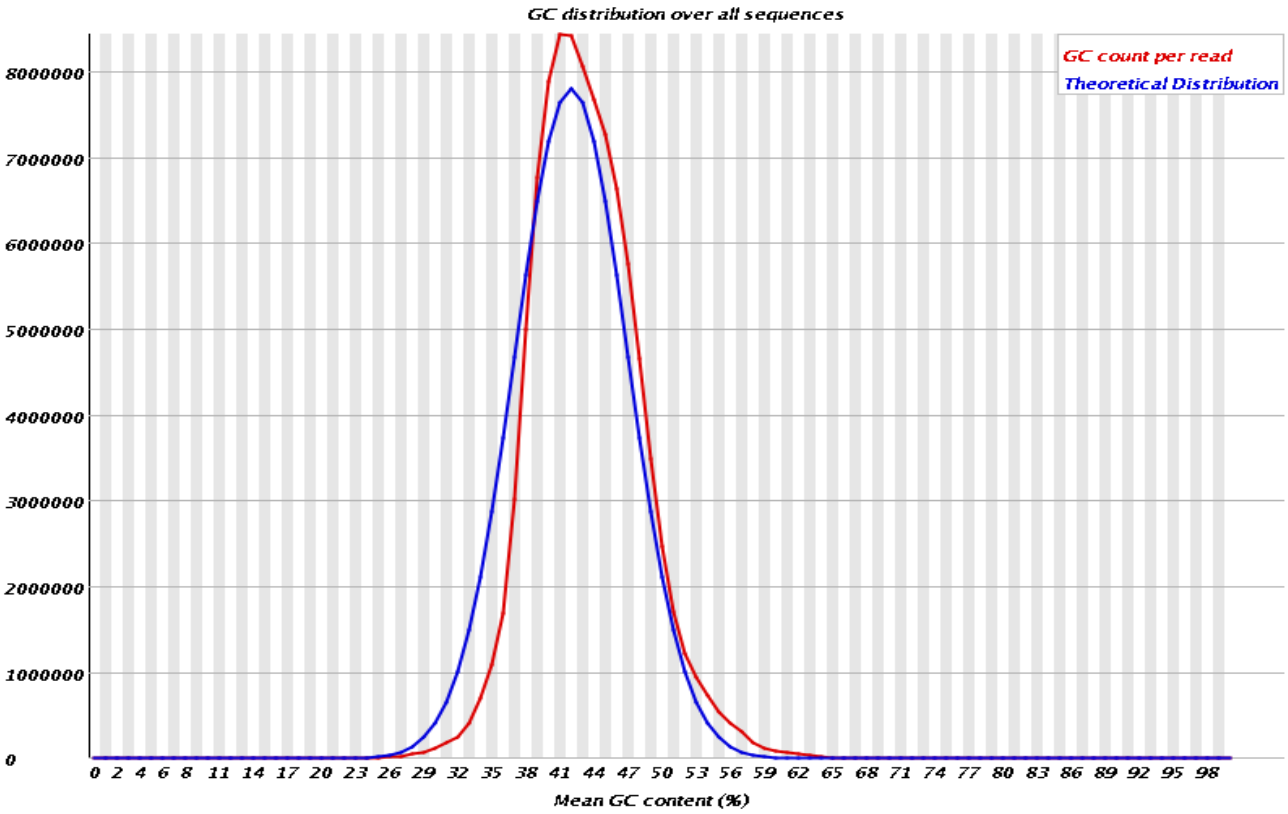
Per base sequence content



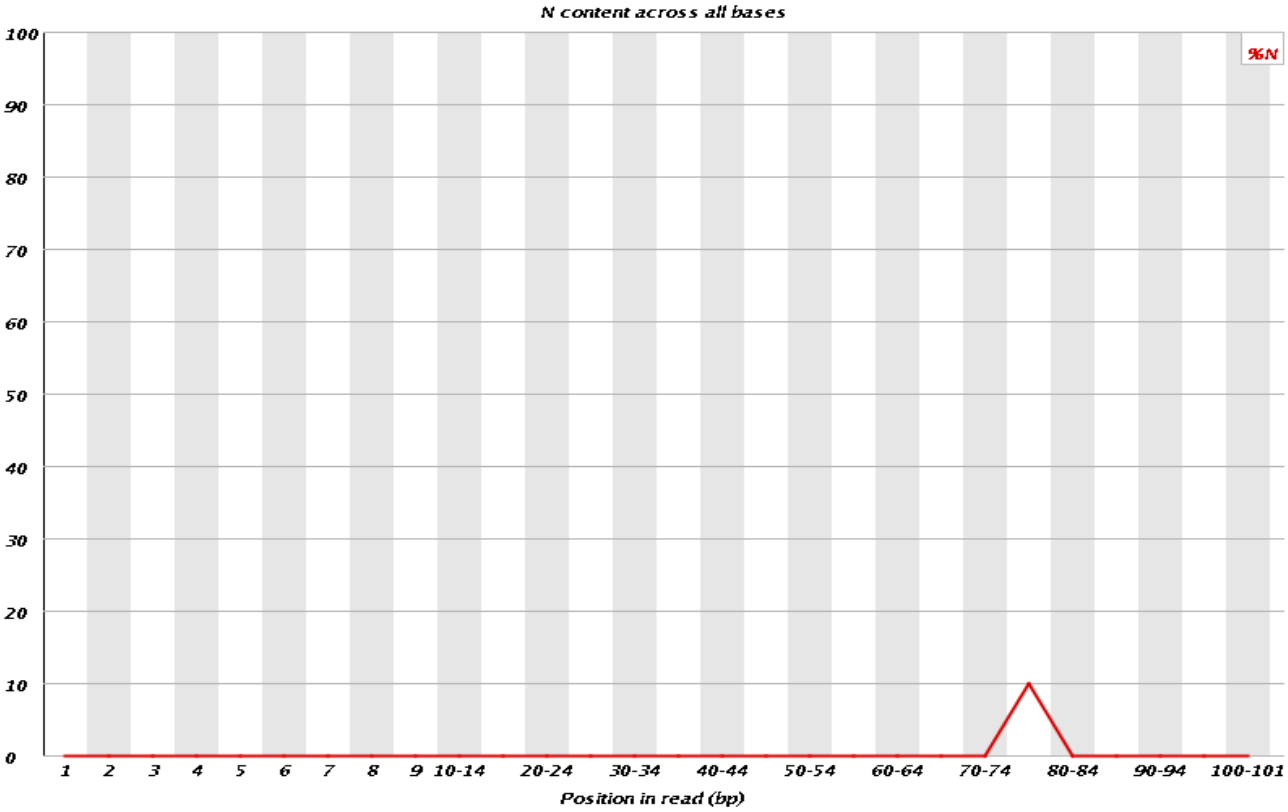
Per base GC content



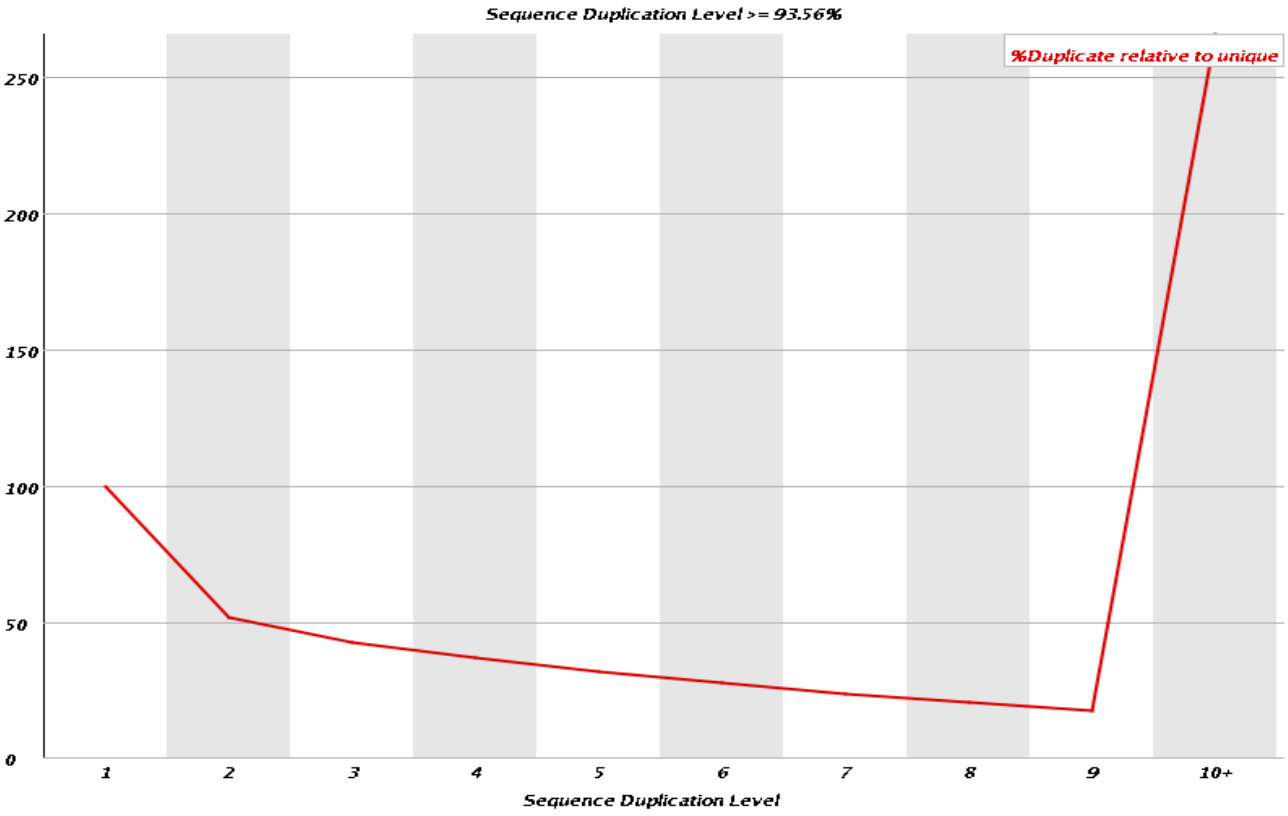
Per sequence GC content



Per base N content



Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AACTCGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	2125182	2.1951221023770984	TruSeq Adapter, Index 4 (100% over 41bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATG	1780121	1.8387050859670475	TruSeq Adapter, Index 4 (100% over 49bp)
CCTGGACCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	1437707	1.485022182778826	TruSeq Adapter, Index 4 (100% over 41bp)
TCAAGCACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	1357739	1.4024224222487198	TruSeq Adapter, Index 4 (100% over 41bp)
CAATCATCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	1150640	1.1885077588080382	TruSeq Adapter, Index 4 (100% over 41bp)
AAGCGAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	1149105	1.1869222416960221	TruSeq Adapter, Index 4 (100% over 41bp)
AGGAATGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	1026226	1.05999927283124	TruSeq Adapter, Index 4 (100% over 41bp)
TCAACGGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	956119	0.9875850394943533	TruSeq Adapter, Index 4 (100% over 41bp)
ATGACGGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	953969	0.9853642826273599	TruSeq Adapter, Index 4 (100% over 41bp)
GCAACTTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	944197	0.9752706844393323	TruSeq Adapter, Index 4 (100% over 41bp)

TTCCGCACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	900338	0.9299682772628376	TruSeq Adapter, Index 4 (100% over 41bp)
TAGGACTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	782530	0.8082831958736478	TruSeq Adapter, Index 4 (100% over 41bp)
CGTACGGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	748633	0.7732706398176128	TruSeq Adapter, Index 4 (100% over 41bp)
TGGTTCCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	742273	0.7667013311319952	TruSeq Adapter, Index 4 (100% over 41bp)
CTGGCTGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	714921	0.7384491721431564	TruSeq Adapter, Index 4 (100% over 41bp)
ATGGCAACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	586820	0.6061323463670071	TruSeq Adapter, Index 4 (100% over 41bp)
CGGTATCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	572634	0.5914794826855334	TruSeq Adapter, Index 4 (100% over 41bp)
CTATGAACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	569496	0.5882382105698938	TruSeq Adapter, Index 4 (100% over 41bp)
CGGAAGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	529215	0.5466315559841445	TruSeq Adapter, Index 4 (100% over 41bp)
CATTATTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	452248	0.4671315588762929	TruSeq Adapter, Index 4 (100% over 41bp)
CTAGTCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	435751	0.450091639790344	TruSeq Adapter, Index 4 (100% over 41bp)

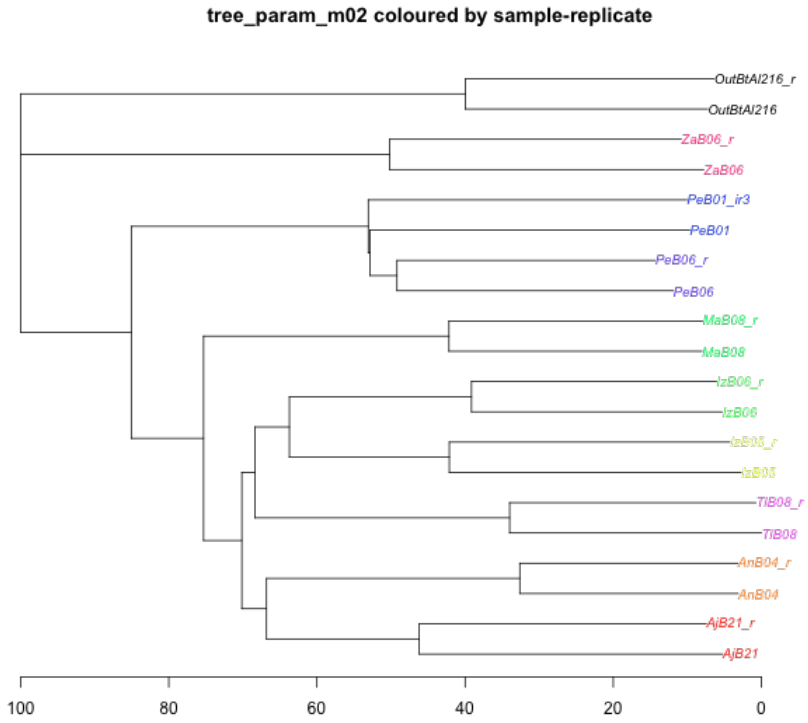
TTCGGTCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	419030	0.43282034882615955	TruSeq Adapter, Index 4 (100% over 41bp)
CTACCTTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	401083	0.41428270999270356	TruSeq Adapter, Index 4 (100% over 41bp)
ACGCAGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	358965	0.3707785994233882	TruSeq Adapter, Index 4 (100% over 41bp)
GTCCTCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	344405	0.3557394273380748	TruSeq Adapter, Index 4 (100% over 41bp)
GAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCT	319408	0.3299197718012218	TruSeq Adapter, Index 4 (100% over 50bp)
CTTACCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	280263	0.28948650316938157	TruSeq Adapter, Index 4 (100% over 41bp)
GCGTCGCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	270322	0.27921834316250654	TruSeq Adapter, Index 4 (100% over 41bp)
GGAGTACCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	264687	0.27339789435064243	TruSeq Adapter, Index 4 (100% over 41bp)
GAATGCCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	210081	0.21699480156969292	TruSeq Adapter, Index 4 (100% over 41bp)
ACCTACCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	181184	0.18714679636713097	TruSeq Adapter, Index 4 (100% over 41bp)
CGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGT	169779	0.17536645587035904	TruSeq Adapter, Index 4 (100% over 50bp)

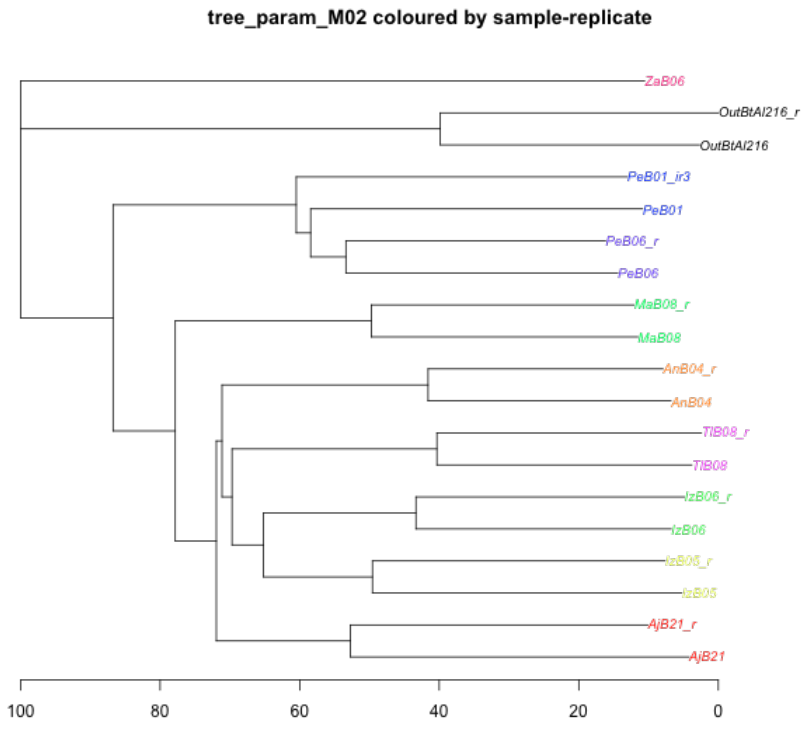
GTATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATG	156982	0.1621483044159802	TruSeq Adapter, Index 4 (97% over 49bp)
TCAACGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATG	126709	0.13087901481854247	TruSeq Adapter, Index 4 (100% over 46bp)
GTATCGGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAAT	115887	0.11970086095128547	TruSeq Adapter, Index 4 (100% over 41bp)
AAGCGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTC	110308	0.11393825510898029	TruSeq Adapter, Index 4 (100% over 44bp)

Produced using part of FastQC (version 0.10.0)

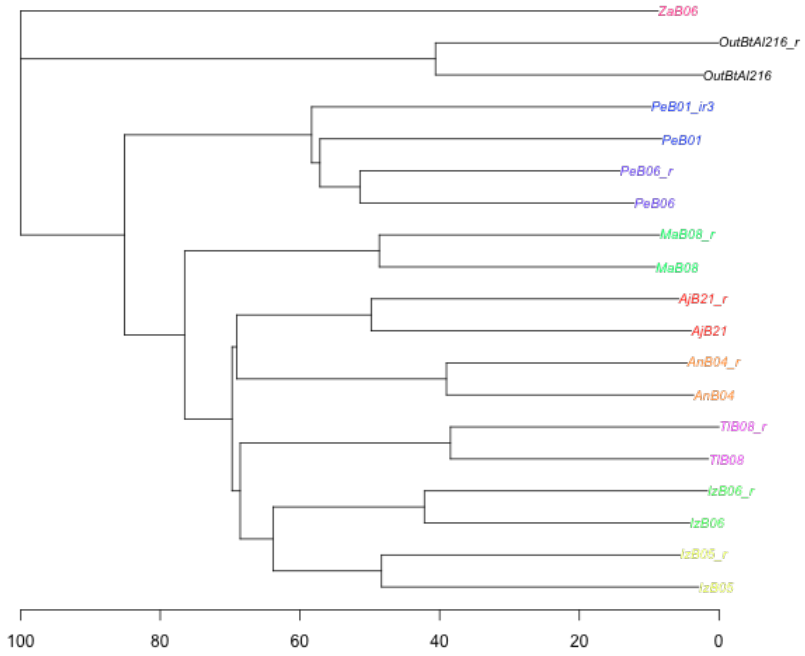
Supporting Information 3. Dendograms

A NJ dendogram was built based on the distance matrix between replicate pairs for each of the combination of Stacks parameters. In each run only the parameter shown in the title varied and the rest were set to m=3, M= 2, n=0, max_locus_stacks = 3 and N= M+2. Trees are scaled 0-100.

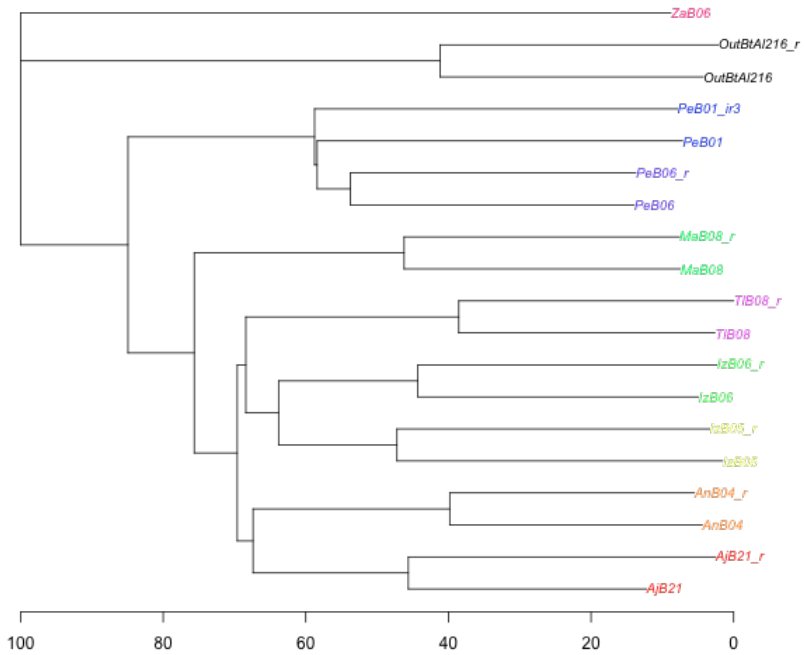




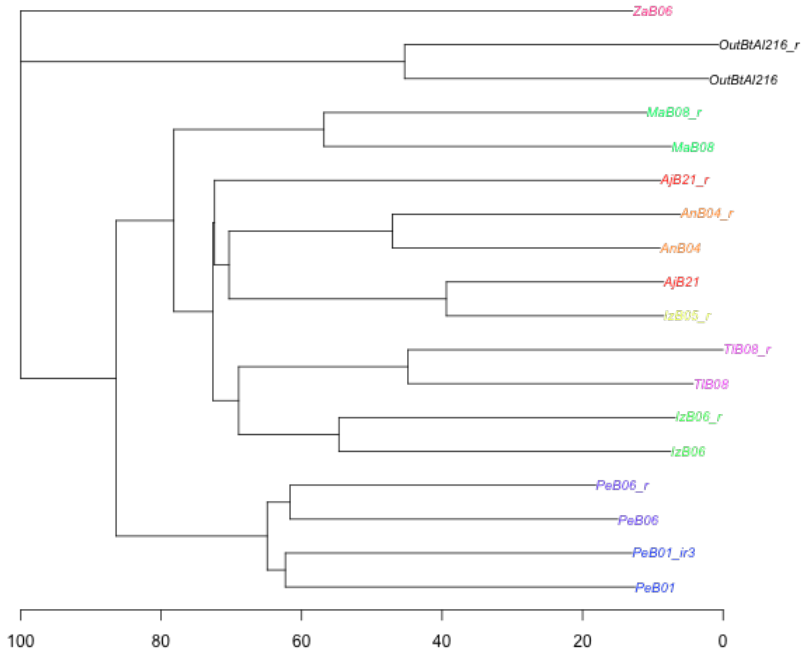
tree_param_m03 coloured by sample-replicate



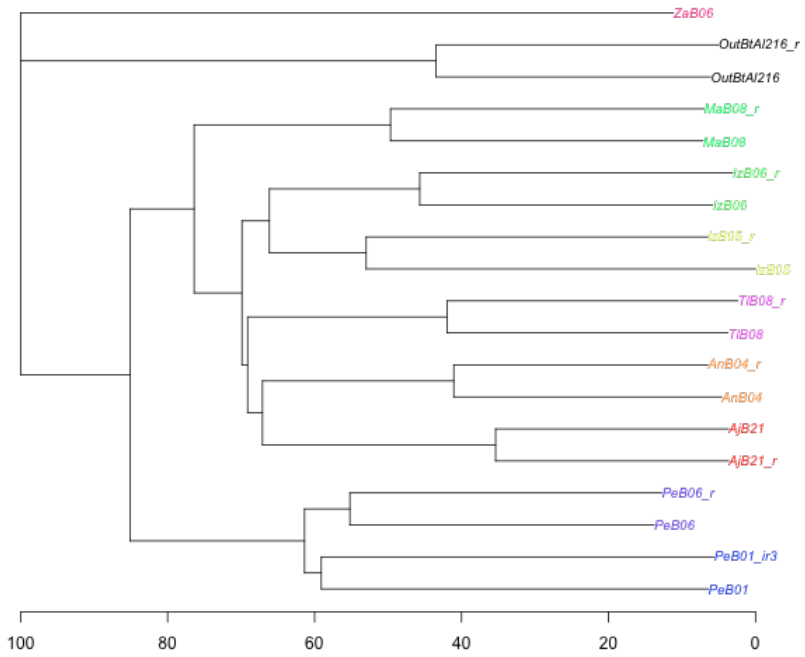
tree_param_M03 coloured by sample-replicate



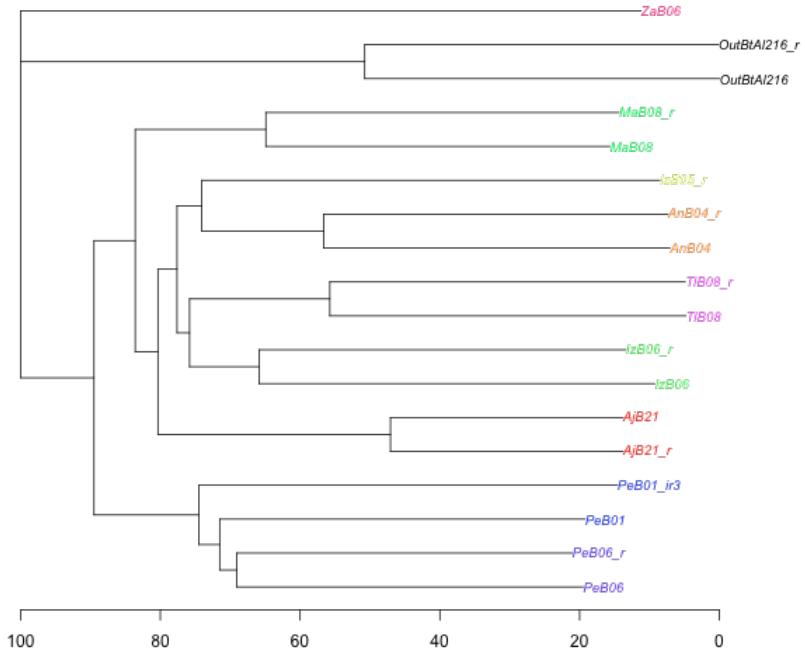
tree_param_m04 coloured by sample-replicate



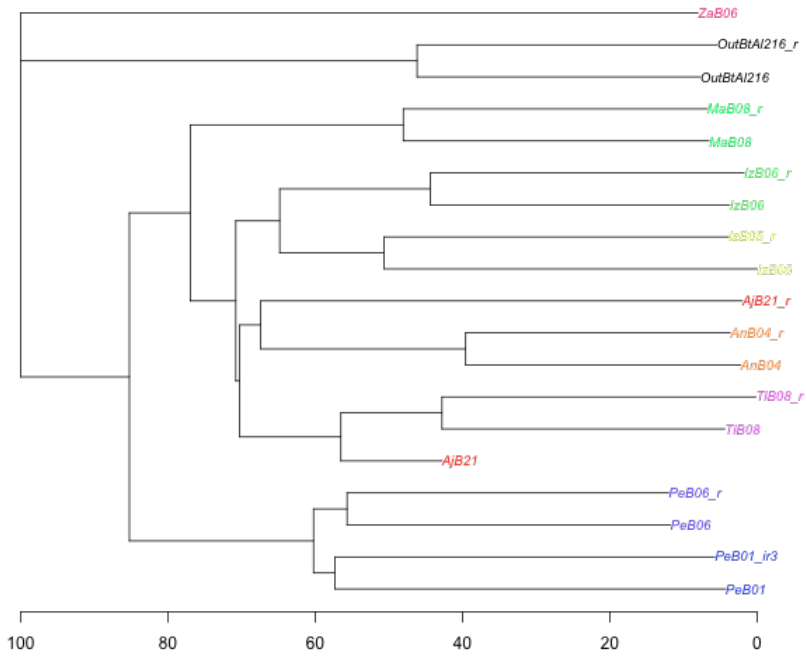
tree_param_M04 coloured by sample-replicate



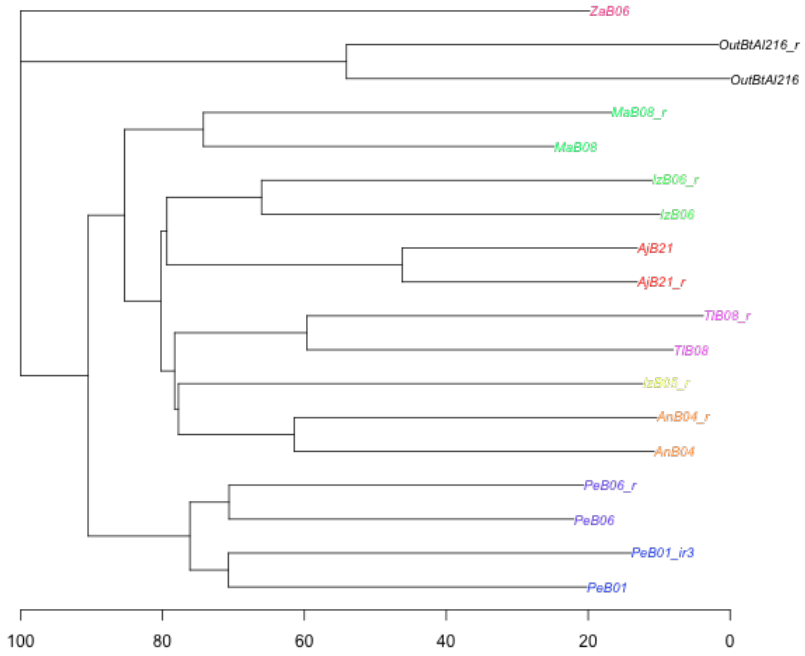
tree_param_m05 coloured by sample-replicate



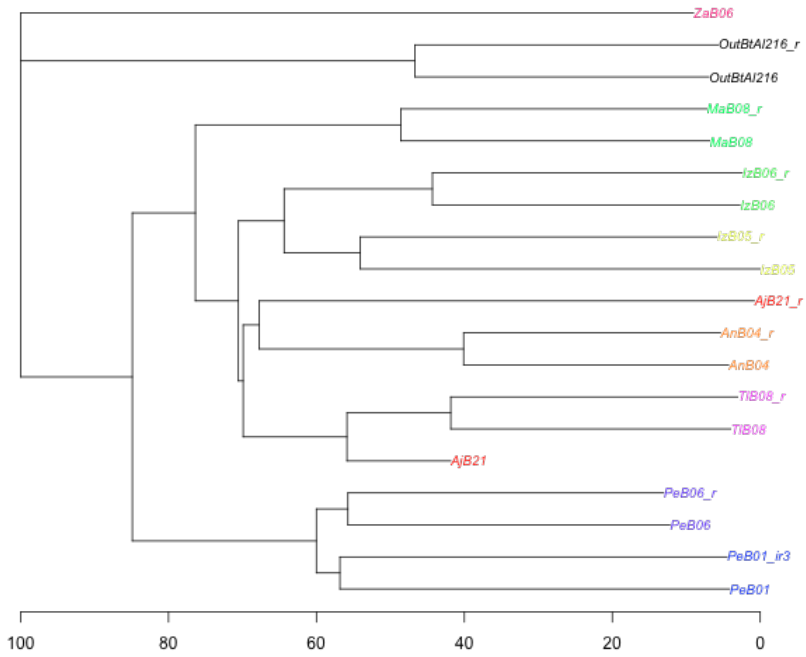
tree_param_M05 coloured by sample-replicate



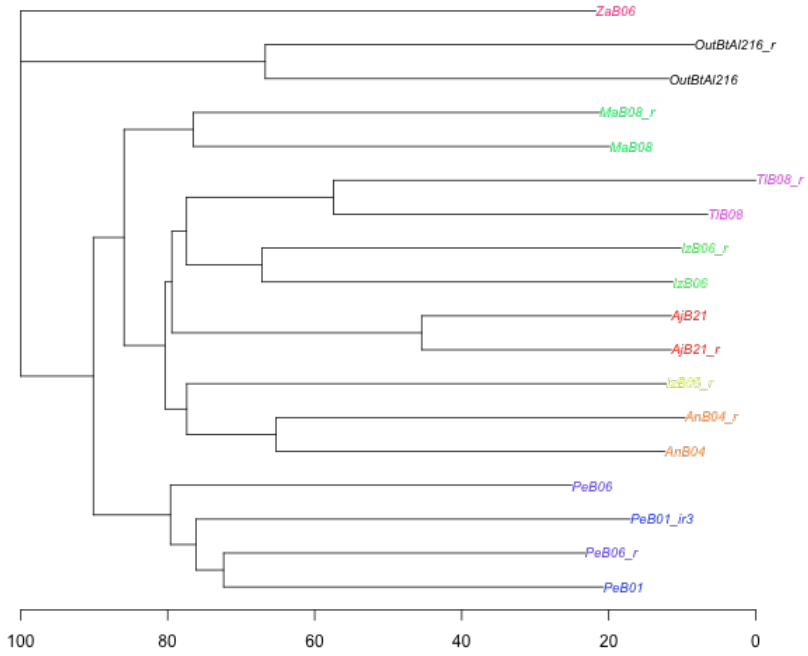
tree_param_m06 coloured by sample-replicate



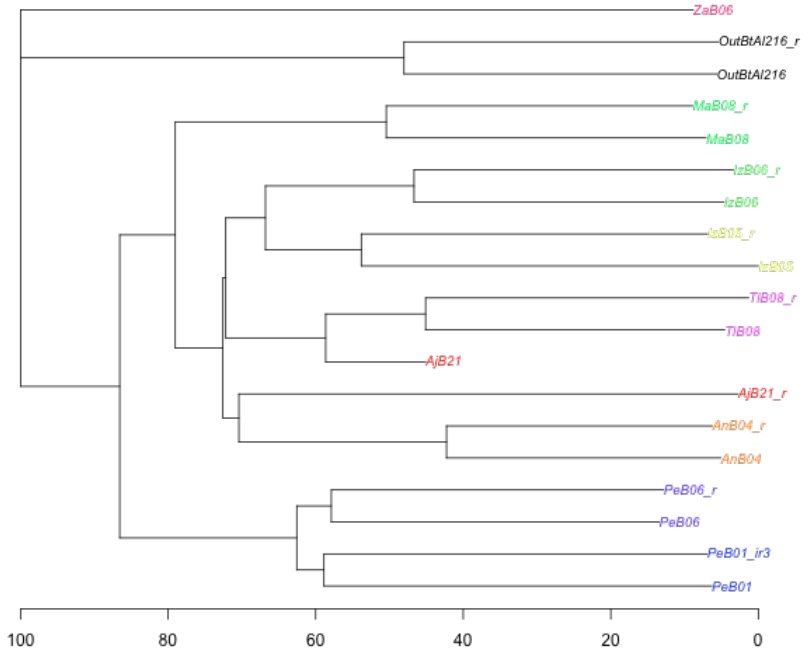
tree_param_M06 coloured by sample-replicate



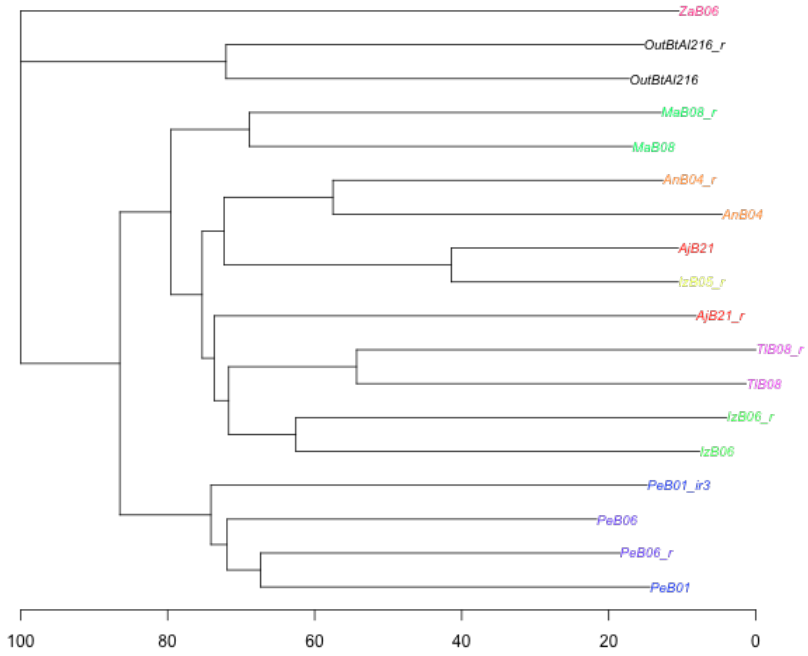
tree_param_m07 coloured by sample-replicate



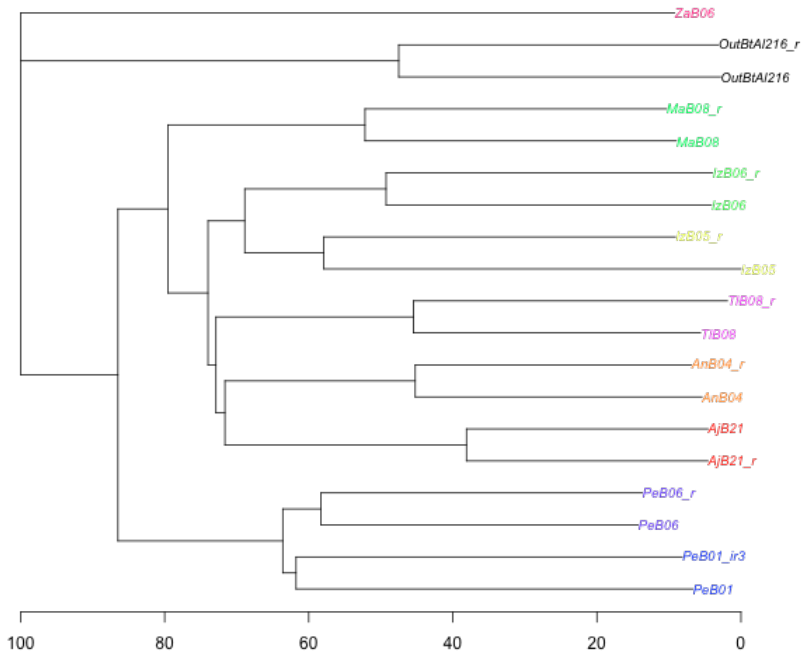
tree_param_M07 coloured by sample-replicate



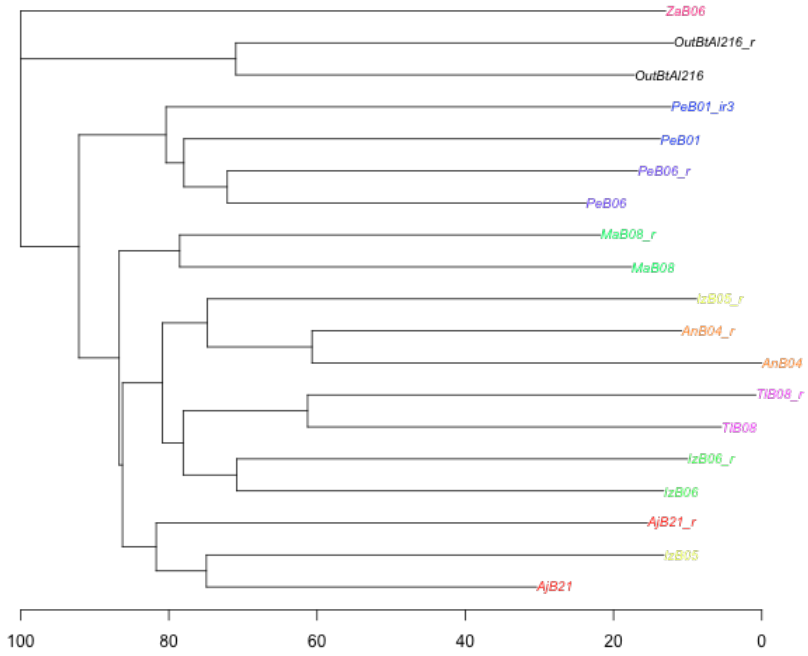
tree_param_m08 coloured by sample-replicate



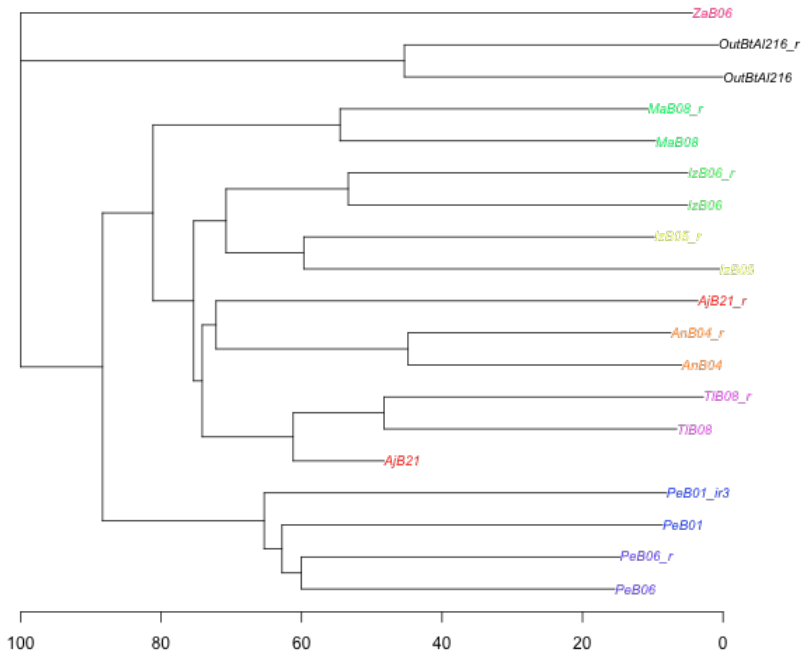
tree_param_M08 coloured by sample-replicate



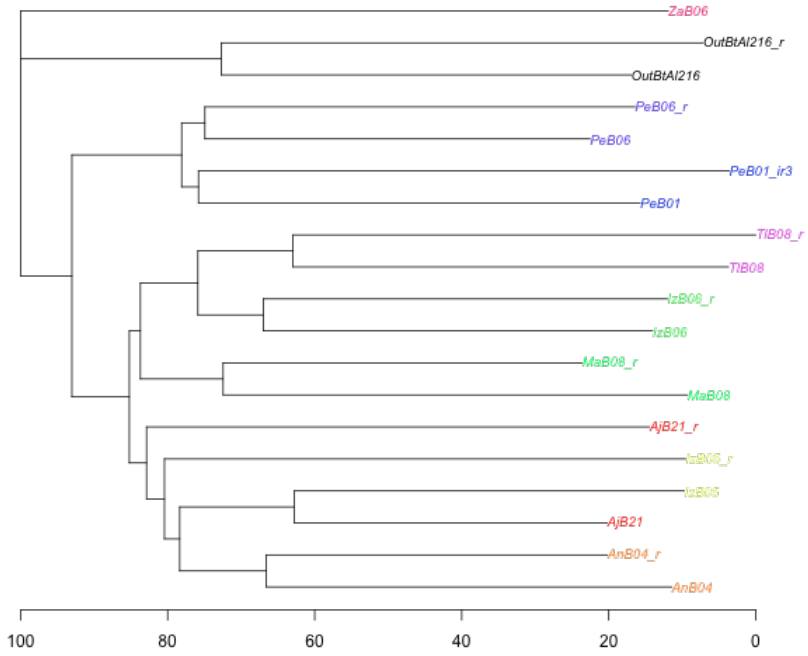
tree_param_m09 coloured by sample-replicate



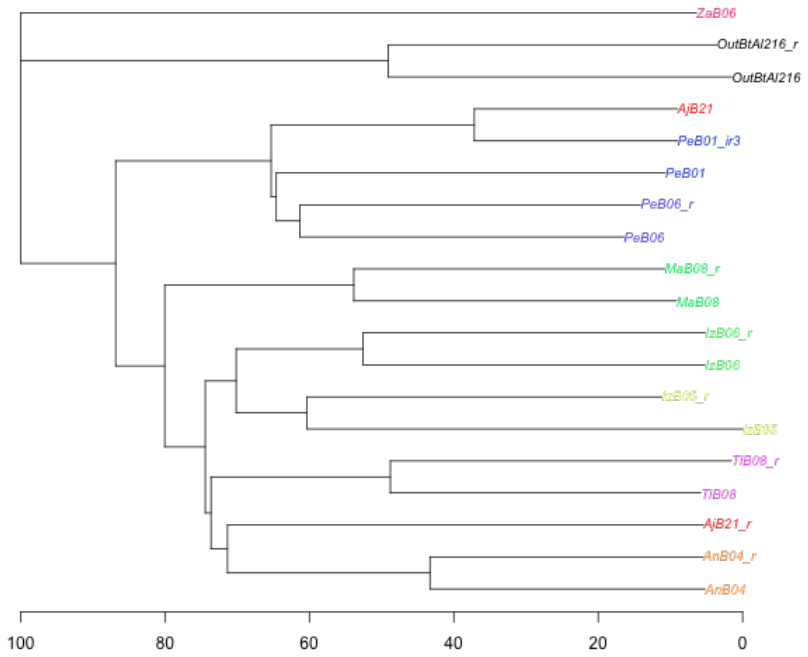
tree_param_M09 coloured by sample-replicate



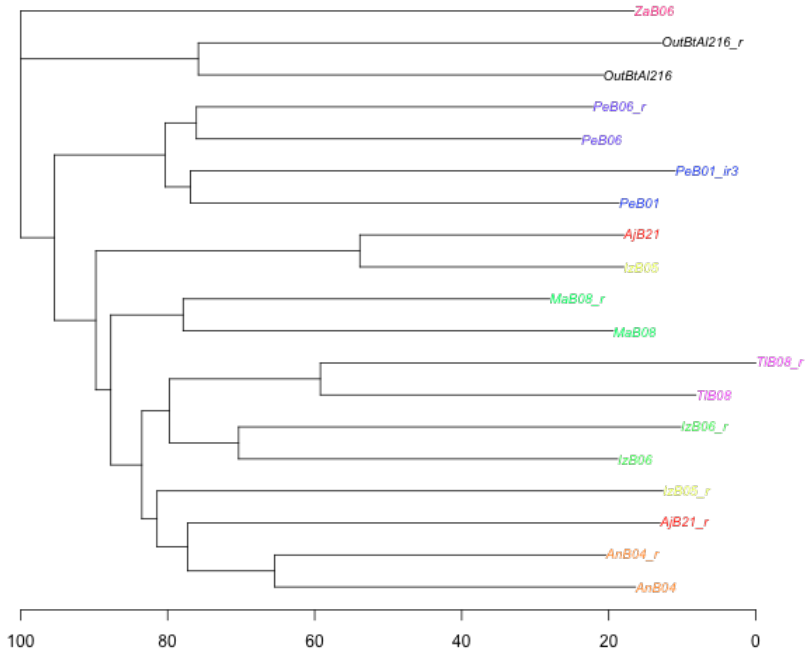
tree_param_m10 coloured by sample-replicate



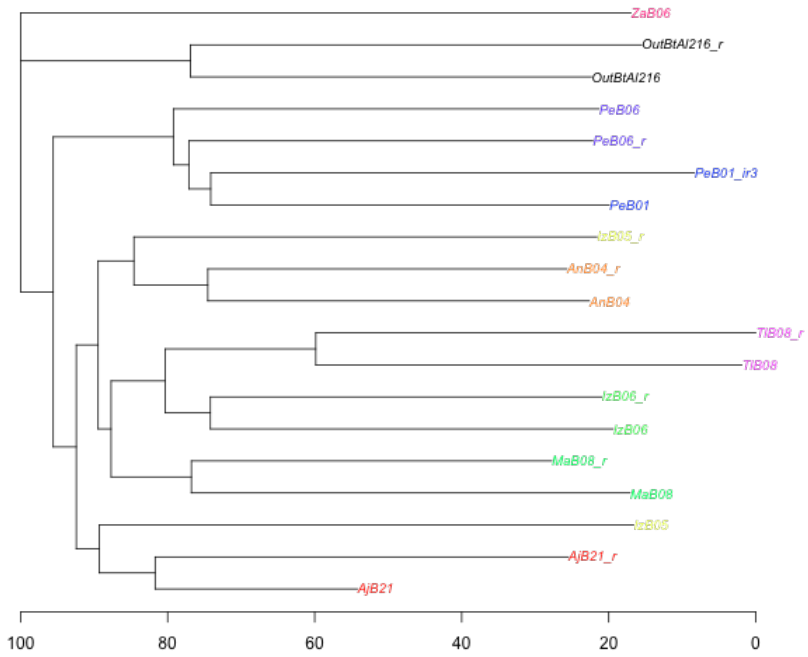
tree_param_M10 coloured by sample-replicate



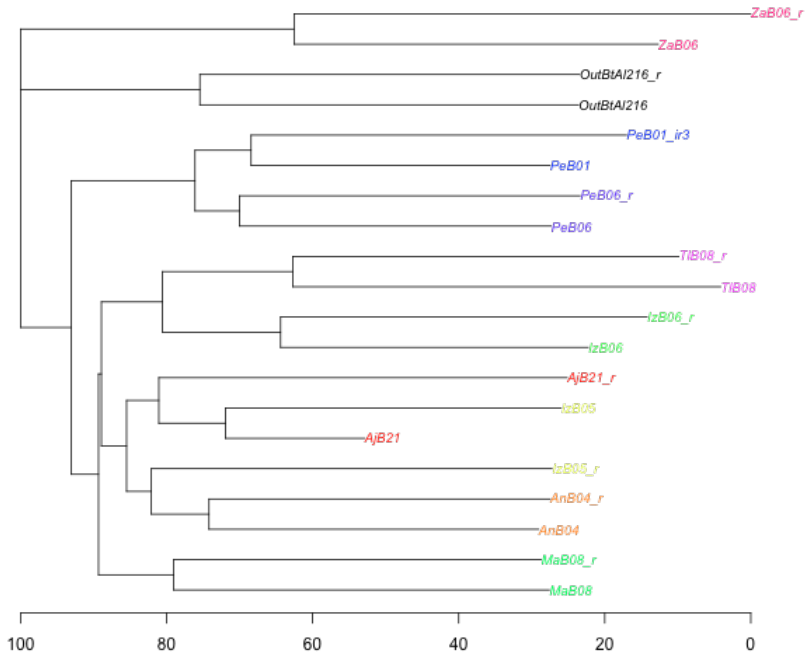
tree_param_m11 coloured by sample-replicate



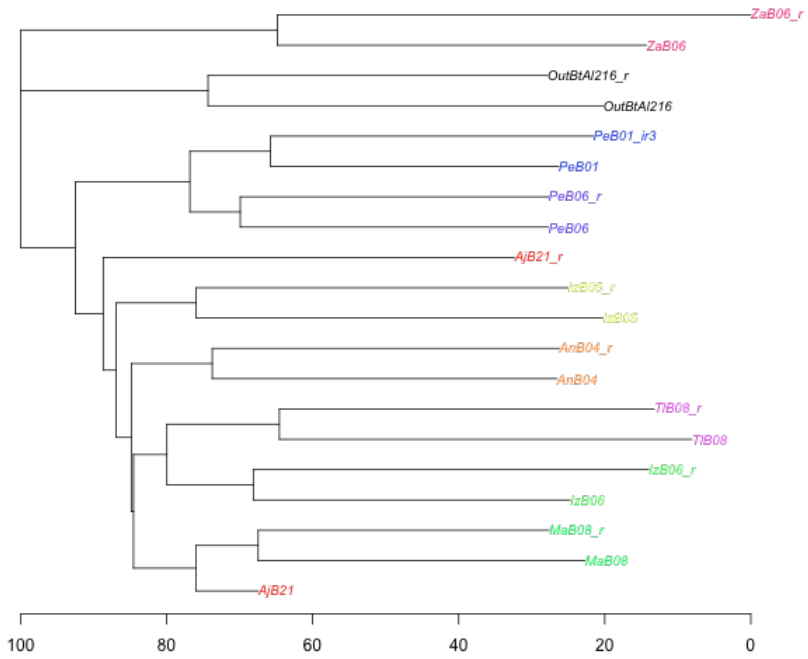
tree_param_m12 coloured by sample-replicate



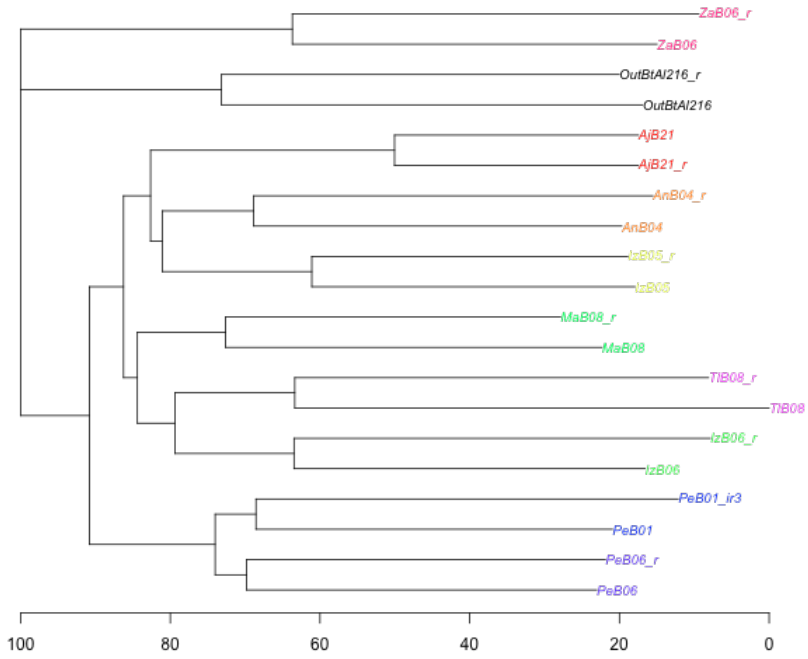
tree_param_m13 coloured by sample-replicate



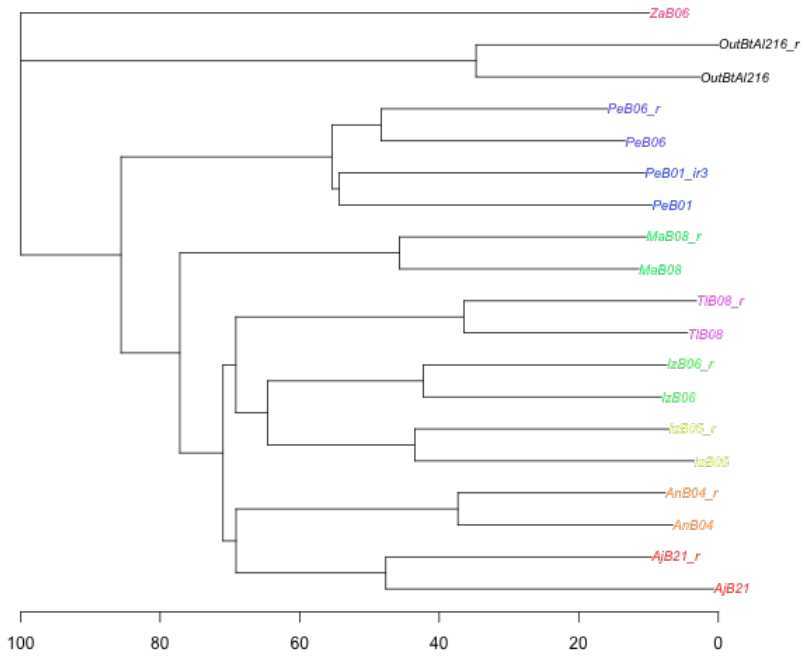
tree_param_m14 coloured by sample-replicate



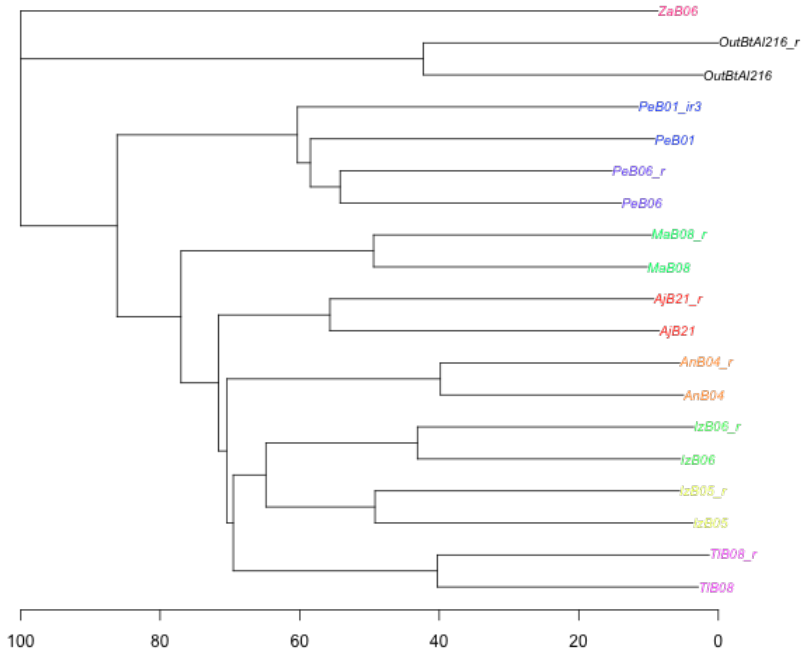
tree_param_m15 coloured by sample-replicate



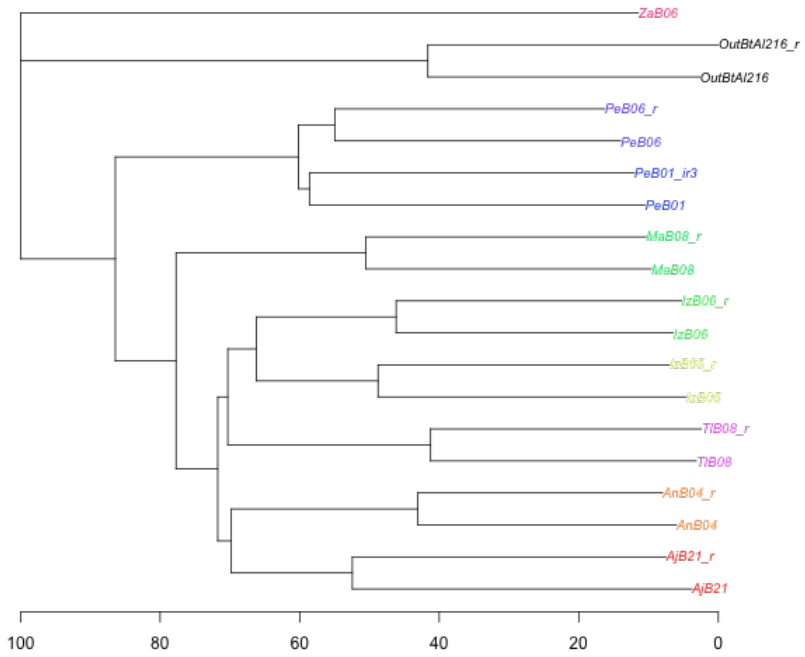
tree_param_maxloc2 coloured by sample-replicate



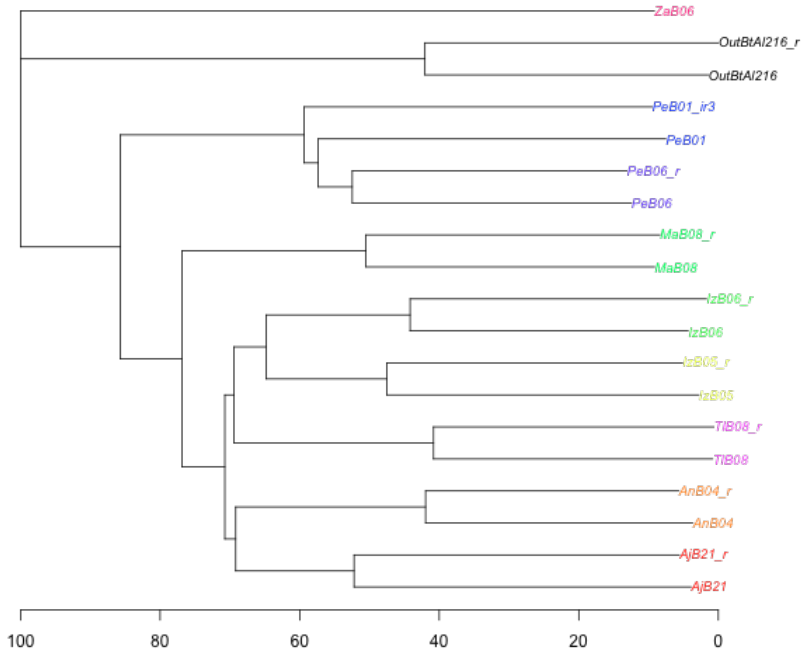
tree_param_maxloc3 coloured by sample-replicate



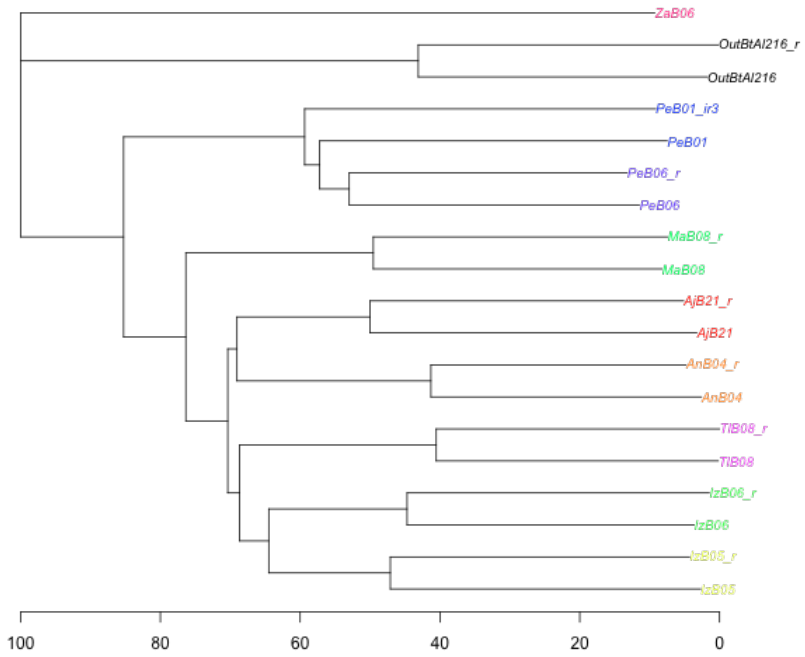
tree_param_maxloc4 coloured by sample-replicate



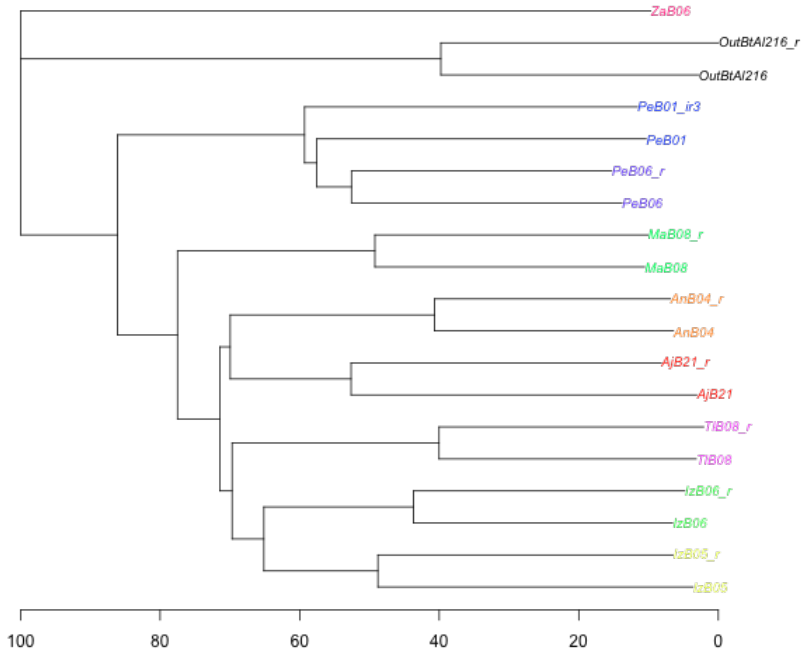
tree_param_maxloc5 coloured by sample-replicate



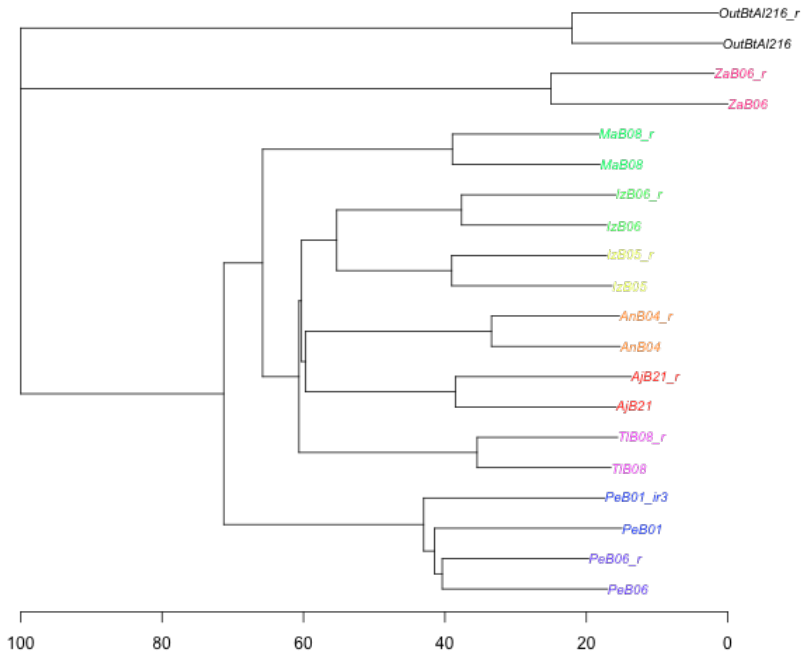
tree_param_maxloc6 coloured by sample-replicate



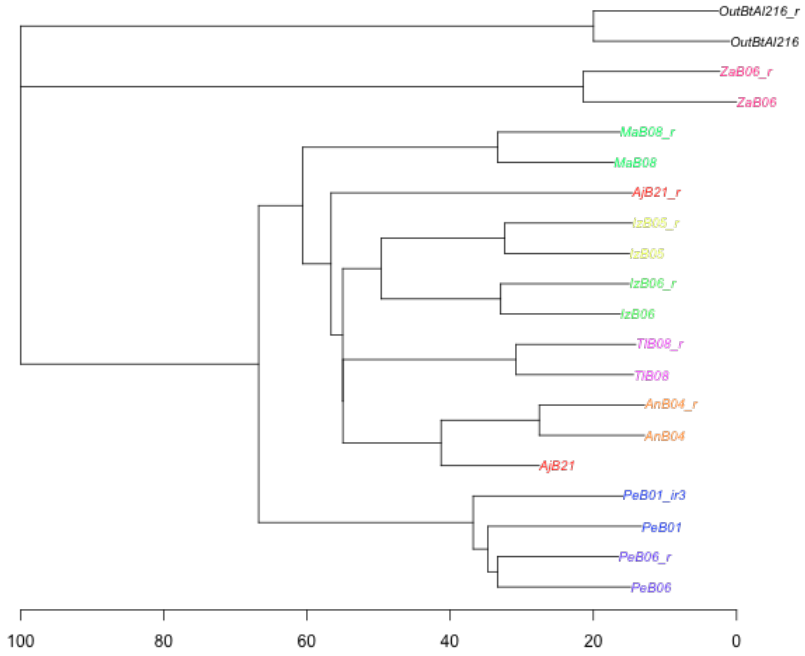
tree_param_n0 coloured by sample-replicate



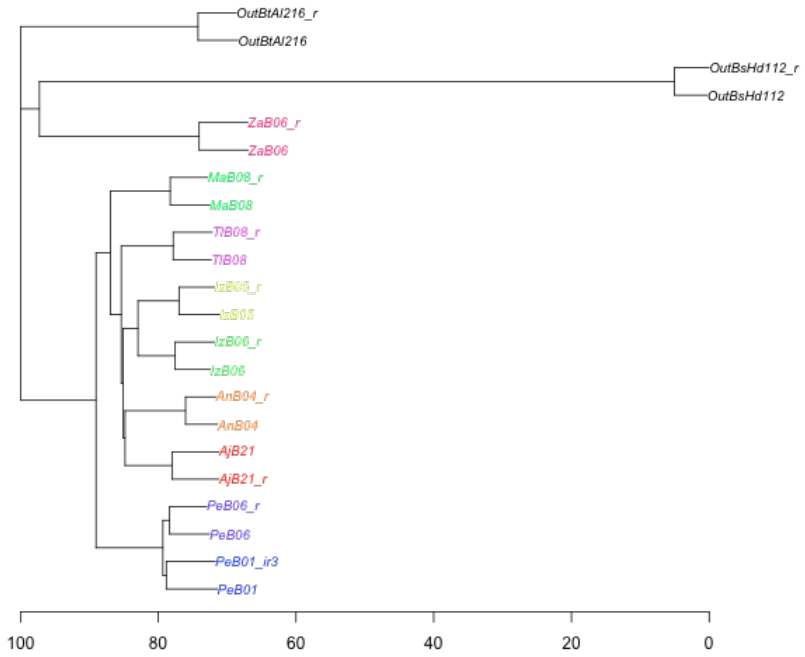
tree_param_n1 coloured by sample-replicate



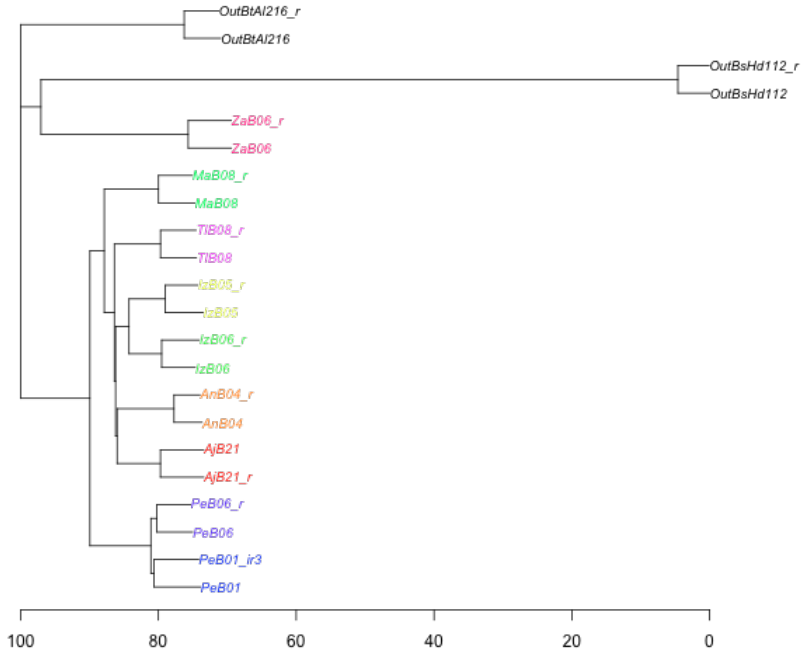
tree_param_n2 coloured by sample-replicate



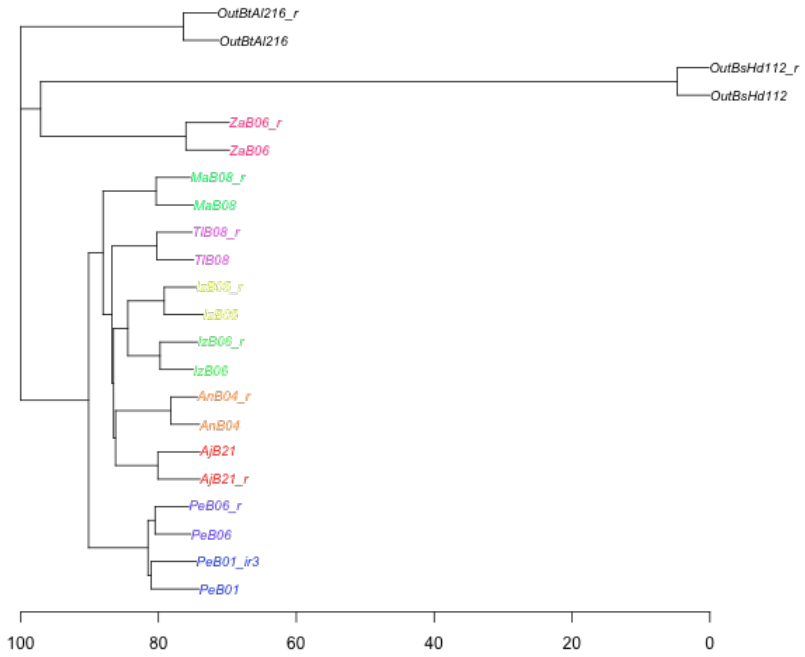
tree_param_n3 coloured by sample-replicate



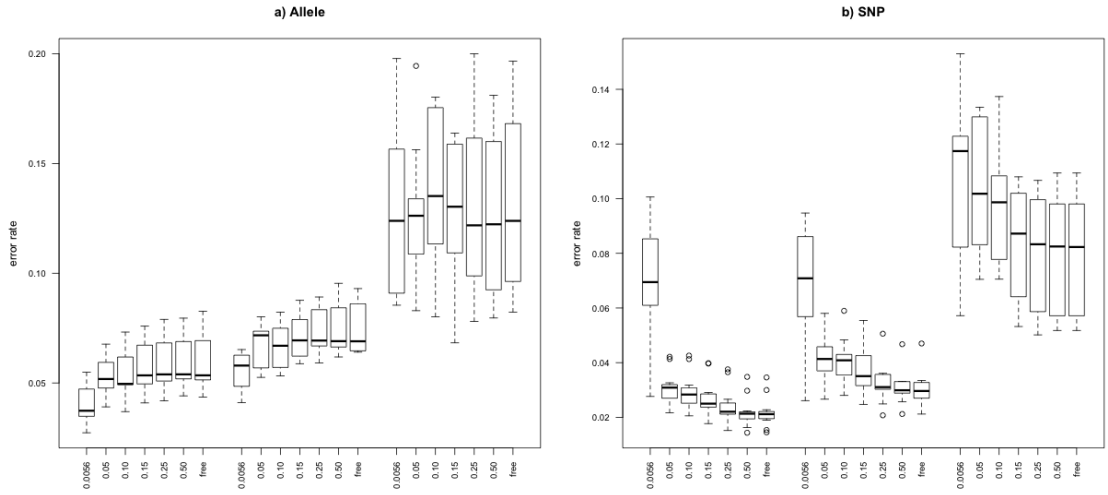
tree_param_n4 coloured by sample-replicate



tree_param_n5 coloured by sample-replicate

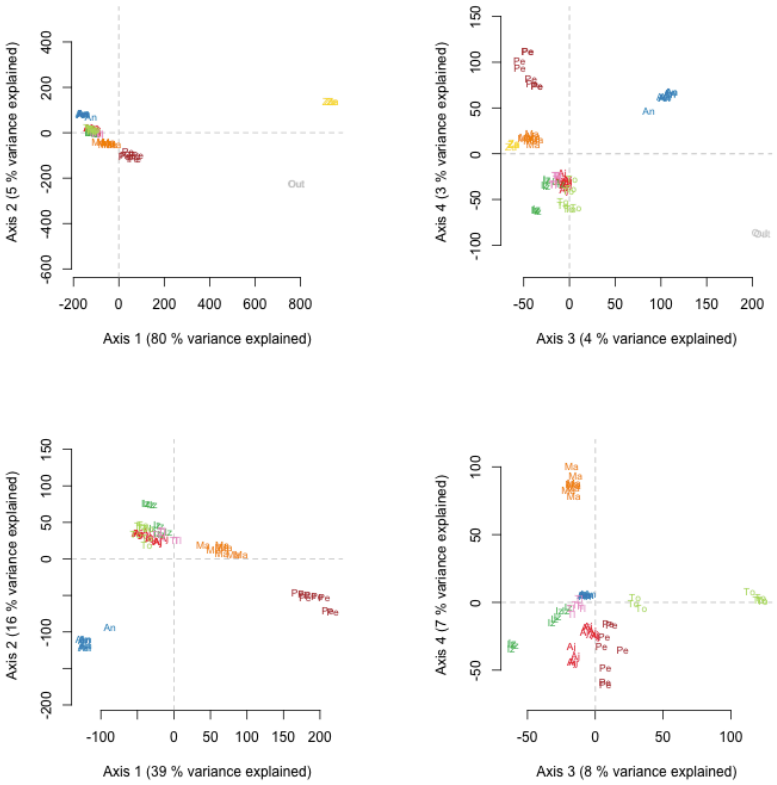


Supporting extra figures

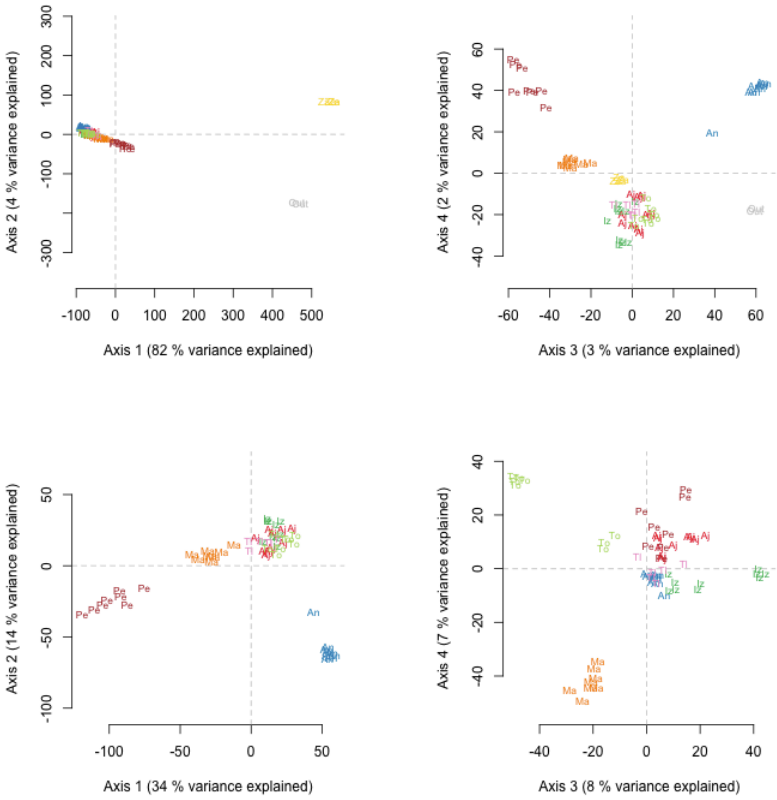


S4. Effect on a) the allele error rate and b) the SNP error rate of using a bounded SNP calling model with different values for the upper bounder (0.0056, 0.05, 0.10, 0.15, 0.25, 0.50) or using the default SNP calling model (free) for three values of -*m*: m=3 (left), m=4 (middle) and m=10 (right).

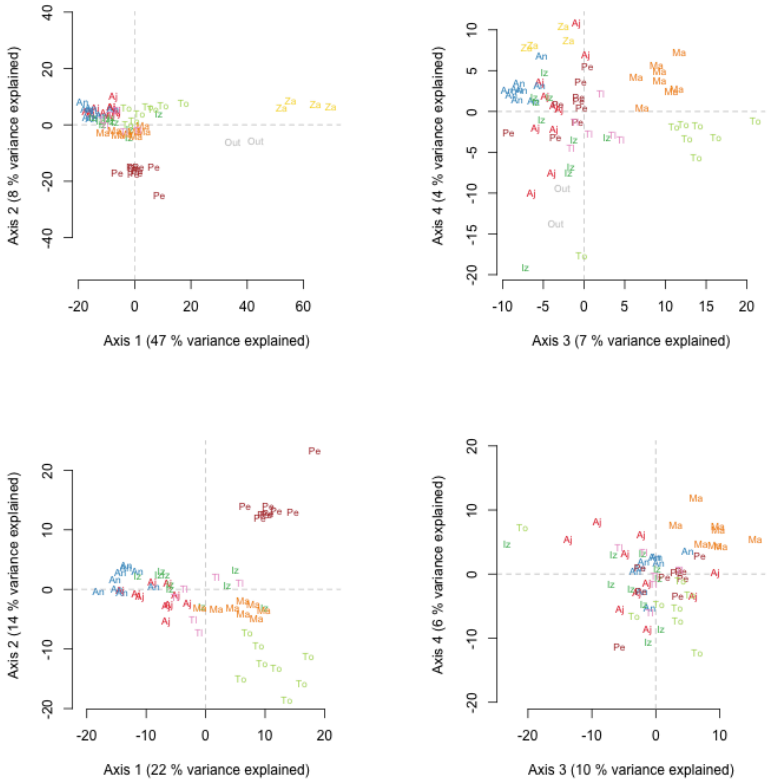
Optimal



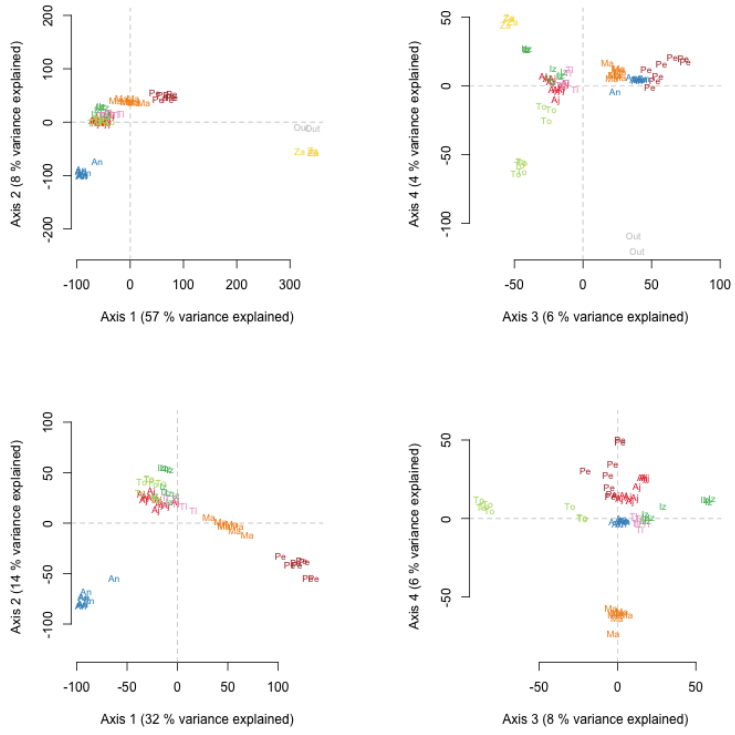
Near optimal



High coverage



Default



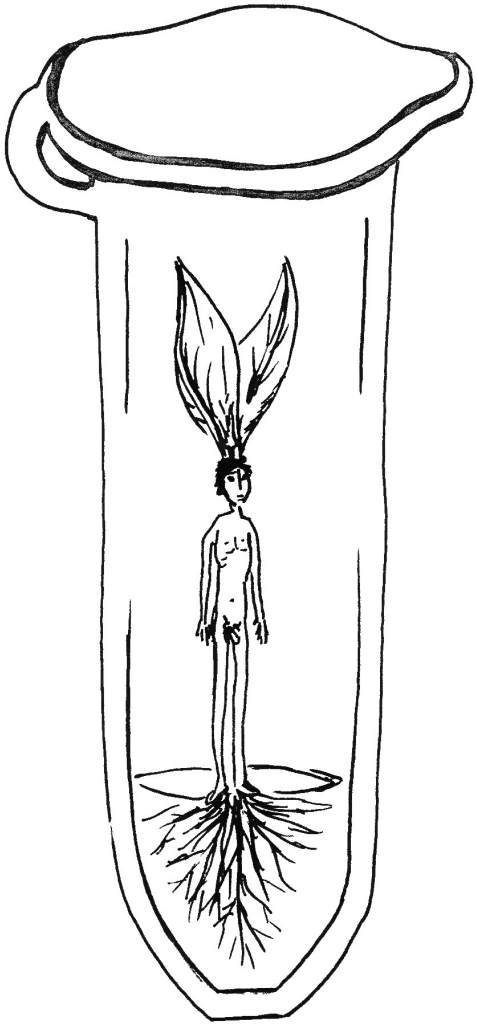
S5. PCoA for each of the four *Stacks* parameter profiles tested (optimal, near optimal, high coverage and default). Upper panels correspond to the PCoA

performed with all samples, and the bottom to the analyses removing the El Zamorano population and *B. trifolia* from the distance matrix.

APPENDIX II

Supporting information for Chapter 5

La verdad nunca es tan espectacular como cuando lo ignoras todo
-Alonso Zamora



In vitro growing, ticatla 2013

Supporting Information 1

INDEX

I. Summary of double digest RAD labwork, sequencing output and the bioinformatic pipeline

Experimental design

Two hundred specimens from seven *Juniperus* species, 30 replicated samples and 10 negative controls were used to construct double digest RAD (ddRAD) libraries with the reagents and conditions explained below. The *J. monticola* dataset analysed here consists of 130 individually tagged specimens of *J. monticola* (10 samples per mountain of 13 localities), four of *J. flaccida,* one of *J. deppeana,* one of *J. zanonii* and 20 replicated samples. Samples of *J. blancoi, J. virginiana* and *J. scopolorum* were also included and sequenced, but would not used for the present study.

Individual DNA extracts were randomly divided into ten groups, each of them corresponding to a pool of individuals for a total of 10 double indexed libraries (Ju01-Ju10, Table 1). Each group comprised 20 *Juniperus* sp. samples, three replicates (one of them replicated in a different group) and one negative control for a total of 24 barcoded (sequence-tagged) individuals. Replicates had the same DNA source but were treated and barcoded independently. Replicates were chosen randomly but included at least one replicate per outgroup and population, except *J. deppeana* which was not replicated. Within each group of 24 barcoded samples all positions on the PCR plates were randomly selected (Table 1). The digestion, ligation and PCR steps were performed in a total of three plates (Table 1). Samples of the same group were then pooled and the size selection for all groups was performed on the same gel. Samples were randomly allocated a well within the corresponding plate. Two libraries, Ju01 and Ju10, were sequenced in a separate lane on an Illumina HiSeq2000 with a single read run,

100bp long at the Lausanne Genomic Technologies Facility, Switzerland. The remaining libraries were pooled in pairs and each pool was sequenced in a single lane using the same service provider.

Table 1. Samples and barcodes used in the preparation of ten ddRAD libraries for Illumina sequencing.

SampleSEQ.ID*	Library	Plate	well	Barcode	ID.adaptor	Index(ILLPCR2)	Index_sequence
JbYc02_ir	Ju01	A	A1	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar05	ACAGTG
JbBk22	Ju01	A	B1	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar05	ACAGTG
JbMh12	Ju01	A	C1	AAGATA	P1_SbfI_03.1	ILLPCR2_bar05	ACAGTG
JmPpJ16	Ju01	A	D1	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar05	ACAGTG
JmTaJ01	Ju01	A	E1	TTACGG	P1_SbfI_05.1	ILLPCR2_bar05	ACAGTG
JbPr13	Ju01	A	F1	AACGAA	P1_SbfI_06.1	ILLPCR2_bar05	ACAGTG
NegCtrL01	Ju01	A	G1	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar05	ACAGTG
JbSl09_r	Ju01	A	H1	CCGACC	P1_SbfI_08.1	ILLPCR2_bar05	ACAGTG
JbZh14	Ju01	A	A2	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar05	ACAGTG
JbHu05	Ju01	A	B2	CATCAA	P1_SbfI_10.1	ILLPCR2_bar05	ACAGTG
JmTlJ16	Ju01	A	C2	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar05	ACAGTG
JmMaJ03	Ju01	A	D2	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar05	ACAGTG
JbHu11	Ju01	A	E2	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar05	ACAGTG
JmPeJ05	Ju01	A	F2	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar05	ACAGTG
JmToJ09	Ju01	A	G2	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar05	ACAGTG
JmPpJ16_r	Ju01	A	H2	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar05	ACAGTG
JbBk20	Ju01	A	A3	GATTAC	P1_SbfI_17.1	ILLPCR2_bar05	ACAGTG
JbZh05	Ju01	A	B3	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar05	ACAGTG
JbHu04	Ju01	A	C3	CAACCG	P1_SbfI_19.1	ILLPCR2_bar05	ACAGTG
JmCoJ01	Ju01	A	D3	GTATGA	P1_SbfI_20.1	ILLPCR2_bar05	ACAGTG
JbHu02	Ju01	A	E3	TGGATT	P1_SbfI_21.1	ILLPCR2_bar05	ACAGTG
JbSl09	Ju01	A	F3	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar05	ACAGTG
JmCiJ14	Ju01	A	G3	AACTCG	P1_SbfI_23.1	ILLPCR2_bar05	ACAGTG
JmBlJ14	Ju01	A	H3	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar05	ACAGTG
JmPeJ02	Ju02	A	A4	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar12	CTTGTA
JmPeJ19	Ju02	A	B4	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar12	CTTGTA
JmCiJ10	Ju02	A	C4	AAGATA	P1_SbfI_03.1	ILLPCR2_bar12	CTTGTA
JmMaJ03_ir	Ju02	A	D4	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar12	CTTGTA
JmMaJ20	Ju02	A	E4	TTACGG	P1_SbfI_05.1	ILLPCR2_bar12	CTTGTA
JbSn12	Ju02	A	F4	AACGAA	P1_SbfI_06.1	ILLPCR2_bar12	CTTGTA
JbSn07	Ju02	A	G4	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar12	CTTGTA
JmIzJ14	Ju02	A	H4	CCGACC	P1_SbfI_08.1	ILLPCR2_bar12	CTTGTA
JbYc10	Ju02	A	A5	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar12	CTTGTA
JmMaJ04	Ju02	A	B5	CATCAA	P1_SbfI_10.1	ILLPCR2_bar12	CTTGTA
JmCiJ16	Ju02	A	C5	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar12	CTTGTA
JbSl07_r	Ju02	A	D5	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar12	CTTGTA

NegCtrL02	Ju02	A	E5	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar12	CTTGTA
OutJzPt02	Ju02	A	F5	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar12	CTTGTA
JmTaJ05	Ju02	A	G5	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar12	CTTGTA
JmNeJ17	Ju02	A	H5	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar12	CTTGTA
OutJdAl201	Ju02	A	A6	GATTAC	P1_SbfI_17.1	ILLPCR2_bar12	CTTGTA
JmTlJ01	Ju02	A	B6	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar12	CTTGTA
OutJzPt02_r	Ju02	A	C6	CAACCG	P1_SbfI_19.1	ILLPCR2_bar12	CTTGTA
JbSl07	Ju02	A	D6	GTATGA	P1_SbfI_20.1	ILLPCR2_bar12	CTTGTA
JbHu12	Ju02	A	E6	TGGATT	P1_SbfI_21.1	ILLPCR2_bar12	CTTGTA
JbMh01	Ju02	A	F6	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar12	CTTGTA
JmToJ02	Ju02	A	G6	AACTCG	P1_SbfI_23.1	ILLPCR2_bar12	CTTGTA
JmPpJ10	Ju02	A	H6	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar12	CTTGTA
JbZh13	Ju03	A	A7	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar06	GCCAAT
JmNeJ18	Ju03	A	B7	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar06	GCCAAT
JbMh18	Ju03	A	C7	AAGATA	P1_SbfI_03.1	ILLPCR2_bar06	GCCAAT
JmIzJ05	Ju03	A	D7	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar06	GCCAAT
JbBk09	Ju03	A	E7	TTACGG	P1_SbfI_05.1	ILLPCR2_bar06	GCCAAT
JmPeJ20	Ju03	A	F7	AACGAA	P1_SbfI_06.1	ILLPCR2_bar06	GCCAAT
JmChJ04	Ju03	A	G7	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar06	GCCAAT
JbPr04	Ju03	A	H7	CCGACC	P1_SbfI_08.1	ILLPCR2_bar06	GCCAAT
JbBk01	Ju03	A	A8	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar06	GCCAAT
JmPeJ04	Ju03	A	B8	CATCAA	P1_SbfI_10.1	ILLPCR2_bar06	GCCAAT
JmMaJ04_ir	Ju03	A	C8	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar06	GCCAAT
JbBk04	Ju03	A	D8	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar06	GCCAAT
JmPpJ04	Ju03	A	E8	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar06	GCCAAT
JbHu15	Ju03	A	F8	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar06	GCCAAT
JbMh18_r	Ju03	A	G8	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar06	GCCAAT
JbSn19	Ju03	A	H8	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar06	GCCAAT
JbYc14	Ju03	A	A9	GATTAC	P1_SbfI_17.1	ILLPCR2_bar06	GCCAAT
JbSl12	Ju03	A	B9	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar06	GCCAAT
JmPpJ04_r	Ju03	A	C9	CAACCG	P1_SbfI_19.1	ILLPCR2_bar06	GCCAAT
NegCtrL03	Ju03	A	D9	GTATGA	P1_SbfI_20.1	ILLPCR2_bar06	GCCAAT
JmNeJ05	Ju03	A	E9	TGGATT	P1_SbfI_21.1	ILLPCR2_bar06	GCCAAT
JbHu01	Ju03	A	F9	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar06	GCCAAT
JmChJ01	Ju03	A	G9	AACTCG	P1_SbfI_23.1	ILLPCR2_bar06	GCCAAT
JmMaJ19	Ju03	A	H9	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar06	GCCAAT
JmIzJ01	Ju04	A	A10	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar12	CTTGTA
JmIzJ04	Ju04	A	B10	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar12	CTTGTA
JmTlJ18	Ju04	A	C10	AAGATA	P1_SbfI_03.1	ILLPCR2_bar12	CTTGTA
JmAjJ24	Ju04	A	D10	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar12	CTTGTA
JmIzJ13	Ju04	A	E10	TTACGG	P1_SbfI_05.1	ILLPCR2_bar12	CTTGTA
JmCoJ09	Ju04	A	F10	AACGAA	P1_SbfI_06.1	ILLPCR2_bar12	CTTGTA
JmNeJ18_ir	Ju04	A	G10	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar12	CTTGTA
JmTlJ03	Ju04	A	H10	CCGACC	P1_SbfI_08.1	ILLPCR2_bar12	CTTGTA
JbYc16	Ju04	A	A11	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar12	CTTGTA
JmAjJ04	Ju04	A	B11	CATCAA	P1_SbfI_10.1	ILLPCR2_bar12	CTTGTA

JbSl11	Ju04	A	C11	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar12	CTTGTA
JbHu03	Ju04	A	D11	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar12	CTTGTA
JbMh02	Ju04	A	E11	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar12	CTTGTA
JmCoJ06	Ju04	A	F11	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar12	CTTGTA
JmAjJ01	Ju04	A	G11	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar12	CTTGTA
JmCiJ09	Ju04	A	H11	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar12	CTTGTA
JbZh19	Ju04	A	A12	GATTAC	P1_SbfI_17.1	ILLPCR2_bar12	CTTGTA
JmTlJ03_r	Ju04	A	B12	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar12	CTTGTA
JbZh19_r	Ju04	A	C12	CAACCG	P1_SbfI_19.1	ILLPCR2_bar12	CTTGTA
JmTaJ03	Ju04	A	D12	GTATGA	P1_SbfI_20.1	ILLPCR2_bar12	CTTGTA
JmMaJ05	Ju04	A	E12	TGGATT	P1_SbfI_21.1	ILLPCR2_bar12	CTTGTA
JmCoJ25	Ju04	A	F12	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar12	CTTGTA
JbZh02	Ju04	A	G12	AACTCG	P1_SbfI_23.1	ILLPCR2_bar12	CTTGTA
NegCtrL04	Ju04	A	H12	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar12	CTTGTA
JmCoJ05	Ju05	B	A1	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar06	GCCAAT
JbYc06	Ju05	B	B1	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar06	GCCAAT
JmIzJ15	Ju05	B	C1	AAGATA	P1_SbfI_03.1	ILLPCR2_bar06	GCCAAT
NegCtrL05	Ju05	B	D1	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar06	GCCAAT
JmAjJ22	Ju05	B	E1	TTACGG	P1_SbfI_05.1	ILLPCR2_bar06	GCCAAT
JmCiJ12	Ju05	B	F1	AACGAA	P1_SbfI_06.1	ILLPCR2_bar06	GCCAAT
JbYc15	Ju05	B	G1	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar06	GCCAAT
JbMh19	Ju05	B	H1	CCGACC	P1_SbfI_08.1	ILLPCR2_bar06	GCCAAT
JbSn01	Ju05	B	A2	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar06	GCCAAT
JmTaJ04	Ju05	B	B2	CATCAA	P1_SbfI_10.1	ILLPCR2_bar06	GCCAAT
JmCiJ13	Ju05	B	C2	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar06	GCCAAT
JmCiJ11	Ju05	B	D2	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar06	GCCAAT
JmNeJ20_r	Ju05	B	E2	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar06	GCCAAT
JmTlJ05	Ju05	B	F2	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar06	GCCAAT
JmIzJ13_ir	Ju05	B	G2	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar06	GCCAAT
JmToJ11	Ju05	B	H2	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar06	GCCAAT
JmCoJ27	Ju05	B	A3	GATTAC	P1_SbfI_17.1	ILLPCR2_bar06	GCCAAT
JmToJ13	Ju05	B	B3	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar06	GCCAAT
JmNeJ20	Ju05	B	C3	CAACCG	P1_SbfI_19.1	ILLPCR2_bar06	GCCAAT
JmPpJ02	Ju05	B	D3	GTATGA	P1_SbfI_20.1	ILLPCR2_bar06	GCCAAT
JmChJ12	Ju05	B	E3	TGGATT	P1_SbfI_21.1	ILLPCR2_bar06	GCCAAT
JmIzJ16	Ju05	B	F3	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar06	GCCAAT
JmCoJ05_r	Ju05	B	G3	AACTCG	P1_SbfI_23.1	ILLPCR2_bar06	GCCAAT
JbZh04	Ju05	B	H3	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar06	GCCAAT
OutJfQr501	Ju06	B	A4	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar12	CTTGTA
JmPeJ22	Ju06	B	B4	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar12	CTTGTA
JbYc08	Ju06	B	C4	AAGATA	P1_SbfI_03.1	ILLPCR2_bar12	CTTGTA
JmPeJ21	Ju06	B	D4	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar12	CTTGTA
JmCoJ03	Ju06	B	E4	TTACGG	P1_SbfI_05.1	ILLPCR2_bar12	CTTGTA
JmChJ12_ir	Ju06	B	F4	AACGAA	P1_SbfI_06.1	ILLPCR2_bar12	CTTGTA
NegCtrL06	Ju06	B	G4	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar12	CTTGTA
JbHu14_r	Ju06	B	H4	CCGACC	P1_SbfI_08.1	ILLPCR2_bar12	CTTGTA

JmCiJ17_r	Ju06	B	A5	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar12	CTTGTA
JmChJ03	Ju06	B	B5	CATCAA	P1_SbfI_10.1	ILLPCR2_bar12	CTTGTA
JmBlJ05	Ju06	B	C5	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar12	CTTGTA
JmPpJ03	Ju06	B	D5	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar12	CTTGTA
JbMh04	Ju06	B	E5	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar12	CTTGTA
JmCoJ26	Ju06	B	F5	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar12	CTTGTA
JmBlJ02	Ju06	B	G5	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar12	CTTGTA
JbBk19	Ju06	B	H5	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar12	CTTGTA
JbPr01	Ju06	B	A6	GATTAC	P1_SbfI_17.1	ILLPCR2_bar12	CTTGTA
JmTaJ19	Ju06	B	B6	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar12	CTTGTA
JmNeJ21	Ju06	B	C6	CAACCG	P1_SbfI_19.1	ILLPCR2_bar12	CTTGTA
JbSl10	Ju06	B	D6	GTATGA	P1_SbfI_20.1	ILLPCR2_bar12	CTTGTA
JmPeJ07	Ju06	B	E6	TGGATT	P1_SbfI_21.1	ILLPCR2_bar12	CTTGTA
JbHu14	Ju06	B	F6	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar12	CTTGTA
JmCiJ17	Ju06	B	G6	AACTCG	P1_SbfI_23.1	ILLPCR2_bar12	CTTGTA
JmNeJ03	Ju06	B	H6	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar12	CTTGTA
OutJfIx301	Ju07	B	A7	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar06	GCCAAT
JbMh09	Ju07	B	B7	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar06	GCCAAT
JmBlJ03	Ju07	B	C7	AAGATA	P1_SbfI_03.1	ILLPCR2_bar06	GCCAAT
OutJfQr501_ir	Ju07	B	D7	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar06	GCCAAT
JmToJ07	Ju07	B	E7	TTACGG	P1_SbfI_05.1	ILLPCR2_bar06	GCCAAT
JbPr11	Ju07	B	F7	AACGAA	P1_SbfI_06.1	ILLPCR2_bar06	GCCAAT
JmBlJ17	Ju07	B	G7	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar06	GCCAAT
JbPr19	Ju07	B	H7	CCGACC	P1_SbfI_08.1	ILLPCR2_bar06	GCCAAT
JbMh06	Ju07	B	A8	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar06	GCCAAT
JmTaJ02	Ju07	B	B8	CATCAA	P1_SbfI_10.1	ILLPCR2_bar06	GCCAAT
JmNeJ02	Ju07	B	C8	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar06	GCCAAT
JmTlJ20	Ju07	B	D8	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar06	GCCAAT
JbZh16	Ju07	B	E8	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar06	GCCAAT
JmNeJ02_r	Ju07	B	F8	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar06	GCCAAT
NegCtrL07	Ju07	B	G8	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar06	GCCAAT
JbSl01	Ju07	B	H8	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar06	GCCAAT
JmPpJ11	Ju07	B	A9	GATTAC	P1_SbfI_17.1	ILLPCR2_bar06	GCCAAT
JbMh17	Ju07	B	B9	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar06	GCCAAT
JmToJ07_r	Ju07	B	C9	CAACCG	P1_SbfI_19.1	ILLPCR2_bar06	GCCAAT
JmTaJ20	Ju07	B	D9	GTATGA	P1_SbfI_20.1	ILLPCR2_bar06	GCCAAT
JmIzJ17	Ju07	B	E9	TGGATT	P1_SbfI_21.1	ILLPCR2_bar06	GCCAAT
JbZh12	Ju07	B	F9	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar06	GCCAAT
JmPeJ06	Ju07	B	G9	AACTCG	P1_SbfI_23.1	ILLPCR2_bar06	GCCAAT
JmToJ10	Ju07	B	H9	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar06	GCCAAT
JmBlJ04	Ju08	B	A10	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar12	CTTGTA
JmMaJ22	Ju08	B	B10	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar12	CTTGTA
JmNeJ01	Ju08	B	C10	AAGATA	P1_SbfI_03.1	ILLPCR2_bar12	CTTGTA
JmCoJ17	Ju08	B	D10	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar12	CTTGTA
JmPpJ05	Ju08	B	E10	TTACGG	P1_SbfI_05.1	ILLPCR2_bar12	CTTGTA
JbPr19_ir	Ju08	B	F10	AACGAA	P1_SbfI_06.1	ILLPCR2_bar12	CTTGTA

JmCiJ08	Ju08	B	G10	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar12	CTTGTA
JmToJ01	Ju08	B	H10	CCGACC	P1_SbfI_08.1	ILLPCR2_bar12	CTTGTA
JmCiJ15	Ju08	B	A11	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar12	CTTGTA
JmIzJ03	Ju08	B	B11	CATCAA	P1_SbfI_10.1	ILLPCR2_bar12	CTTGTA
OutJsUS10933_r	Ju08	B	C11	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar12	CTTGTA
JmAjJ06	Ju08	B	D11	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar12	CTTGTA
JmBlJ04_r	Ju08	B	E11	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar12	CTTGTA
NegCtrL08	Ju08	B	F11	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar12	CTTGTA
JmBlJ18	Ju08	B	G11	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar12	CTTGTA
JmAjJ10	Ju08	B	H11	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar12	CTTGTA
JmMaJ06	Ju08	B	A12	GATTAC	P1_SbfI_17.1	ILLPCR2_bar12	CTTGTA
OutJfFc402	Ju08	B	B12	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar12	CTTGTA
JmPeJ03	Ju08	B	C12	CAACCG	P1_SbfI_19.1	ILLPCR2_bar12	CTTGTA
OutJsUS10933	Ju08	B	D12	GTATGA	P1_SbfI_20.1	ILLPCR2_bar12	CTTGTA
JmIzJ02	Ju08	B	E12	TGGATT	P1_SbfI_21.1	ILLPCR2_bar12	CTTGTA
OutJfSin01	Ju08	B	F12	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar12	CTTGTA
JbHu13	Ju08	B	G12	AACTCG	P1_SbfI_23.1	ILLPCR2_bar12	CTTGTA
JbSl20	Ju08	B	H12	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar12	CTTGTA
JmBlJ16	Ju09	C	A1	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar06	GCCAAT
JmToJ06	Ju09	C	B1	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar06	GCCAAT
OutJfSin01_ir	Ju09	C	C1	AAGATA	P1_SbfI_03.1	ILLPCR2_bar06	GCCAAT
JmAjJ23	Ju09	C	D1	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar06	GCCAAT
JbYc07	Ju09	C	E1	TTACGG	P1_SbfI_05.1	ILLPCR2_bar06	GCCAAT
OutJvUS10220	Ju09	C	F1	AACGAA	P1_SbfI_06.1	ILLPCR2_bar06	GCCAAT
NegCtrL09	Ju09	C	G1	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar06	GCCAAT
JbSn03_r	Ju09	C	H1	CCGACC	P1_SbfI_08.1	ILLPCR2_bar06	GCCAAT
JmAjJ23_r	Ju09	C	A2	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar06	GCCAAT
JmBlJ15	Ju09	C	B2	CATCAA	P1_SbfI_10.1	ILLPCR2_bar06	GCCAAT
JmAjJ21	Ju09	C	C2	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar06	GCCAAT
JmTlJ04	Ju09	C	D2	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar06	GCCAAT
JmPpJ17	Ju09	C	E2	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar06	GCCAAT
JmMaJ21	Ju09	C	F2	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar06	GCCAAT
JmAjJ03	Ju09	C	G2	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar06	GCCAAT
JmMaJ02	Ju09	C	H2	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar06	GCCAAT
JmToJ08	Ju09	C	A3	GATTAC	P1_SbfI_17.1	ILLPCR2_bar06	GCCAAT
JbSn03	Ju09	C	B3	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar06	GCCAAT
JmToJ12	Ju09	C	C3	CAACCG	P1_SbfI_19.1	ILLPCR2_bar06	GCCAAT
JmTlJ02	Ju09	C	D3	GTATGA	P1_SbfI_20.1	ILLPCR2_bar06	GCCAAT
JmAjJ05	Ju09	C	E3	TGGATT	P1_SbfI_21.1	ILLPCR2_bar06	GCCAAT
JmMaJ23	Ju09	C	F3	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar06	GCCAAT
JmTlJ17	Ju09	C	G3	AACTCG	P1_SbfI_23.1	ILLPCR2_bar06	GCCAAT
JbMh03	Ju09	C	H3	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar06	GCCAAT
JbYc02	Ju10	C	A4	GGTCTT	P1_SbfI_01.1	ILLPCR2_bar05	ACAGTG
JbPr02	Ju10	C	B4	CTGGTT	P1_SbfI_02.1	ILLPCR2_bar05	ACAGTG
JmChJ11	Ju10	C	C4	AAGATA	P1_SbfI_03.1	ILLPCR2_bar05	ACAGTG
JmChJ14	Ju10	C	D4	ACTTCC	P1_SbfI_04.1	ILLPCR2_bar05	ACAGTG

JmCoJ04	Ju10	C	E4	TTACGG	P1_SbfI_05.1	ILLPCR2_bar05	ACAGTG
JmTaJ17	Ju10	C	F4	AACGAA	P1_SbfI_06.1	ILLPCR2_bar05	ACAGTG
JmBlJ16_ir	Ju10	C	G4	ATTCAT	P1_SbfI_07.1	ILLPCR2_bar05	ACAGTG
JmNeJ19	Ju10	C	H4	CCGACC	P1_SbfI_08.1	ILLPCR2_bar05	ACAGTG
JmTaJ18	Ju10	C	A5	ATCGTC	P1_SbfI_09.1	ILLPCR2_bar05	ACAGTG
JmChJ02	Ju10	C	B5	CATCAA	P1_SbfI_10.1	ILLPCR2_bar05	ACAGTG
JmTaJ16	Ju10	C	C5	GCCTGG	P1_SbfI_11.1	ILLPCR2_bar05	ACAGTG
JmNeJ04	Ju10	C	D5	TGCTTG	P1_SbfI_12.1	ILLPCR2_bar05	ACAGTG
JmBlJ01	Ju10	C	E5	TCGCAT	P1_SbfI_13.1	ILLPCR2_bar05	ACAGTG
JmChJ20	Ju10	C	F5	GGTAGA	P1_SbfI_14.1	ILLPCR2_bar05	ACAGTG
JbSn11	Ju10	C	G5	GGAGCG	P1_SbfI_15.1	ILLPCR2_bar05	ACAGTG
JmBlJ01_r	Ju10	C	H5	TTGAAC	P1_SbfI_16.1	ILLPCR2_bar05	ACAGTG
JbPr02_r	Ju10	C	A6	GATTAC	P1_SbfI_17.1	ILLPCR2_bar05	ACAGTG
JmChJ13	Ju10	C	B6	CGAGGC	P1_SbfI_18.1	ILLPCR2_bar05	ACAGTG
JmTlJ19	Ju10	C	C6	CAACCG	P1_SbfI_19.1	ILLPCR2_bar05	ACAGTG
JmPpJ01	Ju10	C	D6	GTATGA	P1_SbfI_20.1	ILLPCR2_bar05	ACAGTG
JmChJ05	Ju10	C	E6	TGGATT	P1_SbfI_21.1	ILLPCR2_bar05	ACAGTG
JbSl18	Ju10	C	F6	CCAGCT	P1_SbfI_22.1	ILLPCR2_bar05	ACAGTG
JmPpJ09	Ju10	C	G6	AACTCG	P1_SbfI_23.1	ILLPCR2_bar05	ACAGTG
NegCtrL10	Ju10	C	H6	ACCAGA	P1_SbfI_24.1	ILLPCR2_bar05	ACAGTG

* Sample IDs starting with "Jm" and "Jb" correspond to samples of *J. monticola* or *J. blancoi,* respectively. Next two letters of the code correspond to population IDs (as in Fig. 1 of main text). Outgroup species are labeled with the code "Out" and negative controls with "NegCtr". Replicated samples are labeled with "_r" or "_ir" at the end of the sample ID.

Library preparation

For library preparation we followed a modified version of the Parchman et al., (2012) and Peterson et al., (2012) double digest RAD protocols. For adapter, PCR primer sequences and full protocol see section II of this Supplementary Material. In summary, the library preparations consisted of the following steps: (1) Phenol-chloroform wash and ethanol precipitation of DNA extractions. DNA concentrations after the wash were standardized to 30-45 ng/μL with the exception of some samples where concentration was <10 ng/μL. (2) Digestion of each DNA sample with SbfI (HF) and MseI at 37°C for ten hours, followed by inactivation of restriction enzymes at 65°C for 20 minutes. (3) Adapter ligation was performed in the same well from the digestion reaction using T4 DNA ligase at 16^{o}C for six hours. A general (non-sample specific) MseI adaptor was added to all samples in the ligation master mix, followed by the addition of a sample-

specific SbfI adaptor for each DNA sample. For sample-specific SbfI adaptors a unique 6bp long barcode was used. In each set of 24, all of the barcodes are separated from each other by at least 3 substitutions. (4) Digestion-ligation products were diluted with 100 µL of water, purified using AMPure XP in a 0.8 ratio and eluted in Tris pH 8.5 buffer. (5) Amplification of adapter-barcode-ligated fragments using Illumina PCR primers. To ameliorate stochastic differences in PCR production of fragments across reactions, the following reaction procedure was performed individually for each restriction-ligation product and combined at a later stage (see step 8). Amplification reactions were performed with Phusion Taq, Phusion PCR buffer, dNTP, $MgCl_2$, DMSO and a PCR primer mix of ILLPCR1 and ILLPCR2-bar05, ILLPCR2-bar06 and ILLPCR2-bar012 (depending on experimental design, Table 1) under the following conditions: 98 °C for 30 seconds; 20 cycles of: 98 °C for 20 seconds, 60° C for 30 seconds, 72° C for 40 seconds; final extension at 72° C for 10 minutes. (6) Addition of primers and dNTPs for a final thermal cycle to reduce the concentration of single-stranded or heteroduplex PCR products. For this step, a reaction mix containing the Phusion PCR Buffer, dNTPs and the same PCR primer mix of the previous step (but excluding Phusion Taq and $MgCl_2$) was added to each of the previous reactions and cycled at 98° C for 3 minutes, 60° C for 2 minutes and 72° C for 12 minutes. (7) Electrophoresis of 3 µL of the reaction from step 6 in a 1.5% agarose gel, run at 100 V for 1 hr to confirm reaction success. (8) Pooling of reactions within each library (Ju01-Ju10) into a single 1.5 ml microcentrifuge tube which was then evaporated to half the volume. (9) Selection of a size range between 500-600 bp by manual excision from a 1.5% agarose gel run at 100 V for 2 hours. Purification of the gel extracts was performed with the MiniElute Qiagen gel extraction kit using one column per gel lane. The 10 libraries were run in the same gel, adding 80 µL per well in 3 wells per library, and separating each library with empty wells and DNA ladder. The final elutions of columns belonging to the same library were pooled together for a final ethanol precipitation. (10) Measurement of library concentration using Qubit fluorometer and submission to the Fragment Analyzer Automated CE System to evaluate the desired concentration and range of the fragments selected (Figure 1).

We used enzymes from New England Biolabs: SbfI-HF (R3642S), MseI (R0525S), T4 DNA Ligase (M0202S), Phusion Taq (M0530S) and their corresponding buffers.

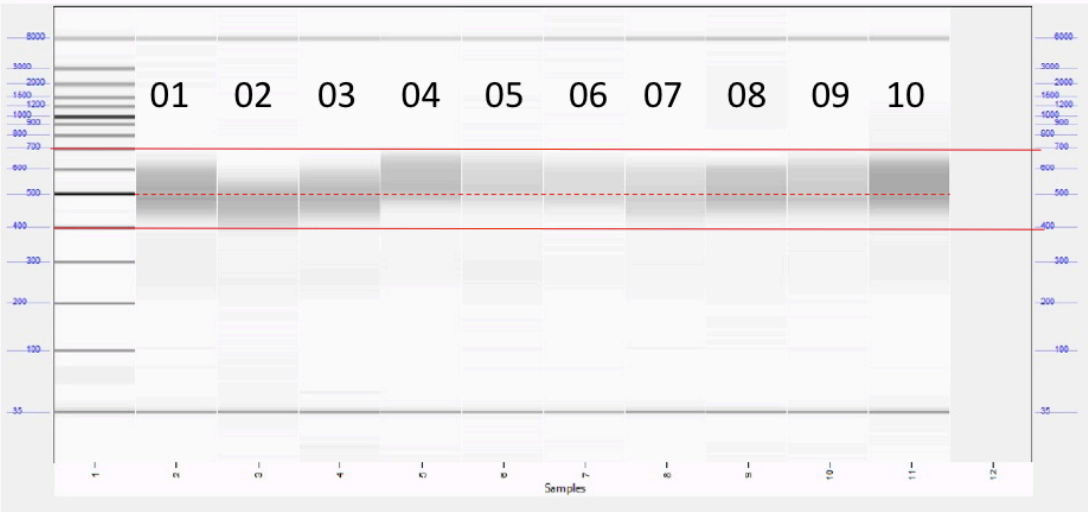


Fig 1. Fragment analyser run of Ju01-Ju10 libraries. Selected fragments ranged from 400 to 700 bp in each library with a peak at 500 bp. Library Ju02 shows a slight bias towards the lower side of the range, and Ju04-08 towards the upper side.

Final DNA concentrations after purification and ethanol cleaning were 23.8, 17.3, 13.8, 21.3, 11.0, 11.4, 11.8, 13.3, 10.3 and 23.2 ng/µl for libraries Ju01-Ju10, respectively.

Libraries Ju01 and Ju10 were sequenced in separate lanes on an Illumina HiSeq2000 with a single read run, 100bp long at the Lausanne Genomic Technologies Facility, Switzerland. The rest of the libraries were pooled in pairs (Ju08-Ju09, Ju06-Ju07, Ju04-Ju05 and Ju02-Ju03), each of which was sequenced in a single lane with the same specifications and service provider. Libraries from each pool-pair had Illumina indexes 06 and 12 (as recommended for pools of two Illumina indexes).

Demultiplexing

Juniperus raw reads were demultiplexed and quality filtered using *Stacks* v. 1.17 by (1) truncating final read length to 87 (because there was a quality drop after this position in lane 10); (2) removing any reads with an uncalled base; (3)

discarding reads with low quality scores (score limit 22 to 28, depending on the sequencing lane); (4) discard reads that had been marked by Illumina's chastity filter as failing; (5) filtering adapter sequences and (6) rescuing tags (maximum distance of 1 between barcodes). Sequencing yield and final number of reads per library are shown in Table 2.

Table 2. Sequencing yield, number of reads lost during quality filtering and total retained reads per library

	Library				
	Ju01	**Ju02**	**Ju03**	**Ju04**	**Ju05**
Total reads	**275,831,143**	**69,307,699**	**64,036,139**	**59,716,341**	**36,493,856**
Failed Illumina-filtered reads	7,050,929	1,575,333	1,767,812	1,031,160	700,303
Reads containing adapter sequence	471,931	139,136	162,690	122,123	149,252
Ambiguous barcode drops	140,313,314	17,386,128	13,009,783	9,165,502	3,113,741
Low quality read drops	10,167,375	3,593,548	3,788,019	8,052,859	5,525,816
Ambiguous RAD-Tag drops	59,336,530	23,011,590	22,786,391	23,443,528	14,491,246
Retained reads	**58,491,064**	**23,601,964**	**22,521,444**	**17,901,169**	**12,513,498**

	Library				
	Ju06	**Ju07**	**Ju08**	**Ju09**	**Ju10**
Total reads	**94,918,753**	**68,101,885**	**65,418,638**	**63,740,841**	**90,916,796**
Failed Illumina filtered reads	3,834,006	2,620,101	1,539,905	1,609,925	1,651,452
Reads containing adapter sequence	336,467	260,309	146,174	128,981	129,635
Ambiguous barcode drops	12,172,197	9,848,273	12,461,987	11,407,125	15,655,037
Low quality read drops	5,487,362	4,560,778	3,928,894	4,016,697	11,391,511
Ambiguous RAD-Tag drops	38,660,467	23,511,843	21,074,560	21,410,173	38,375,024
Retained reads	**34,428,254**	**27,300,581**	**26,267,118**	**25,167,940**	**23,714,137**

De novo *assembly*

Using the replicates set of samples, a range of *de novo* assembly parameters were tested as in Mastretta-Yanes *et al.* (2014a) to optimise for the recovery of a large number of loci while reducing the SNP and RAD allele error rates. Specifically, the following key parameters were tested with the values specified

in parentheses: the minimum number of raw reads required to form a stack (*–m* 2 to 15), the maximum number of mismatches allowed between stacks when processing an individual (*-M* 2 to 10), the allowed number of mismatches between loci when building the catalog (*-n* 0 to 5) and the maximum number of stacks per locus (*--max_locus_stacks* 2 to 6). Only one parameter was varied at a time while keeping the other parameters fixed to *m*=3, *M*=2, *n*=0 and *max_locus_stacks*=3. The value of *–N* was always defined as *M*+2.

After examining the yield on number of RAD-loci and SNP-loci (Fig. 2), the effect on missing data (Figs. 3-5) and error rates (Fig. 6-8), the chosen parameters for optimised *de novo* assembly were: m=10, M=2, N=4, n=3, max_locus_stacks=4 and default SNP calling model. Notice that the –m parameter was set to 10 because this recovered more loci than m=12 or 15 which provided the smallest allele and error rates. The dataset of assembled samples included in total 166 samples, out of which 10 were negative controls, 6 corresponded to *J. flaccida*, 1 to *J. deppeana* and 2 to *J. zanonii* and were used as outgroups. The rest (148) belong to *J. monticola* sampled from the TMVB.

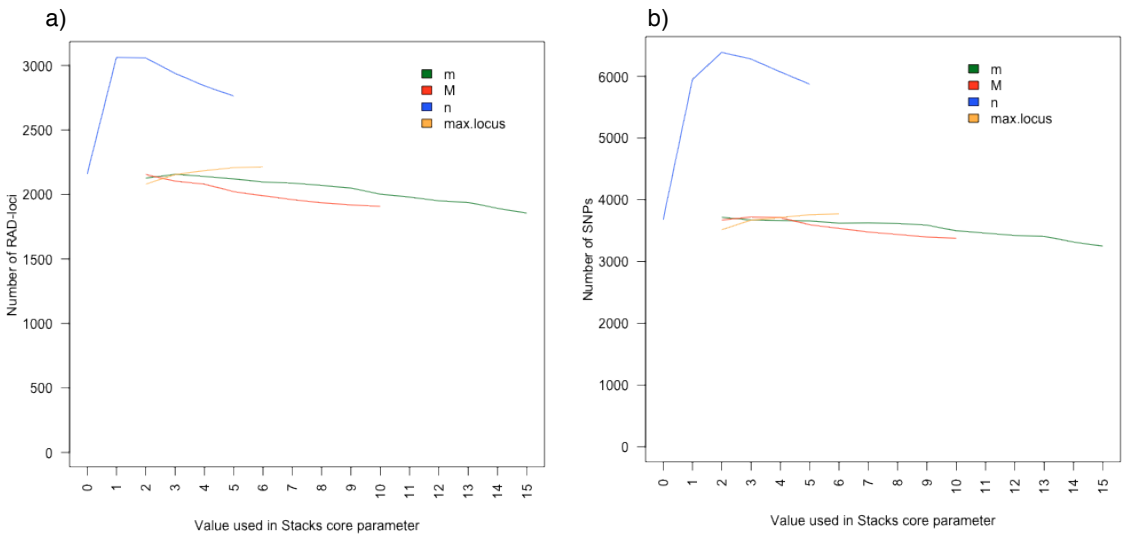


Fig. 2. Total number of (a) RAD-loci and (b) SNP-loci obtained using different *Stacks* core parameter settings. Only one parameter varied in each run with the remaining set to m = 3, M = 2, n = 0,max_locus_stacks (max.locus) = 3 and N = M + 2.

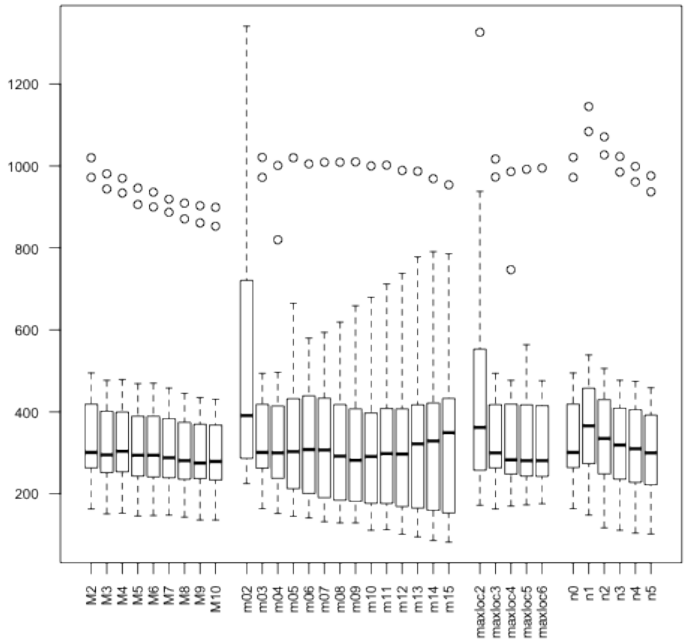


Fig. 3. Effect of different values for *Stacks* core parameters on total number of missing loci. Only one parameter varied in each run with the remaining set as explained in Fig. 1.

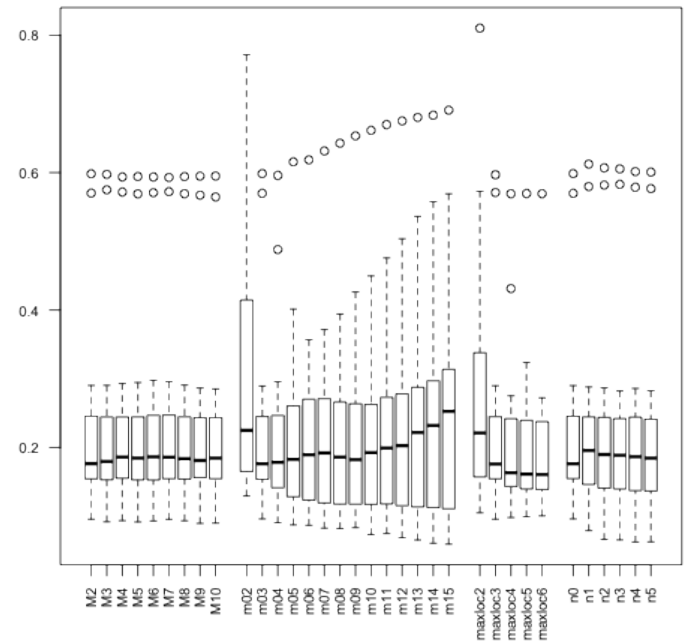


Fig. 4. Effect of different values for *Stacks* core parameters on the proportion of missing loci relative to the total. Only one parameter varied in each run with the remaining set as explained in Fig. 1.

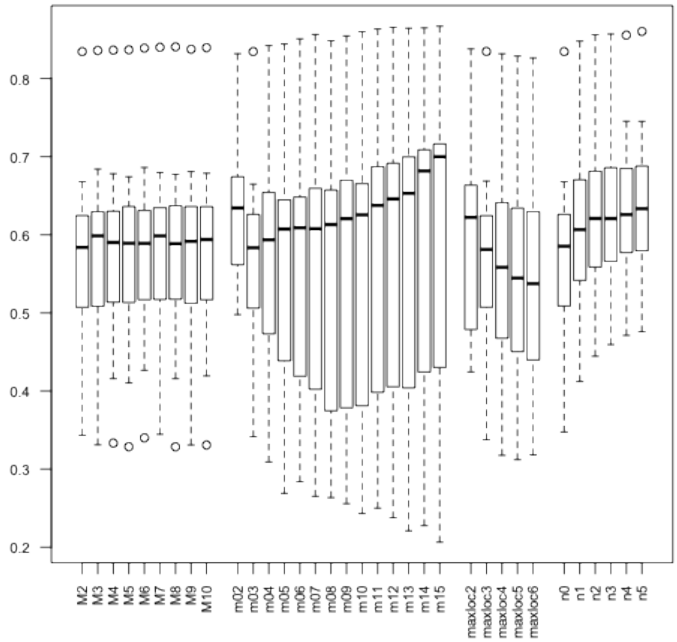


Fig. 5. Effect of different values for Stacks core parameters on the proportion of missing loci different within a replicate pair. Only one parameter varied in each run with the remaining set as explained in Fig. 1.

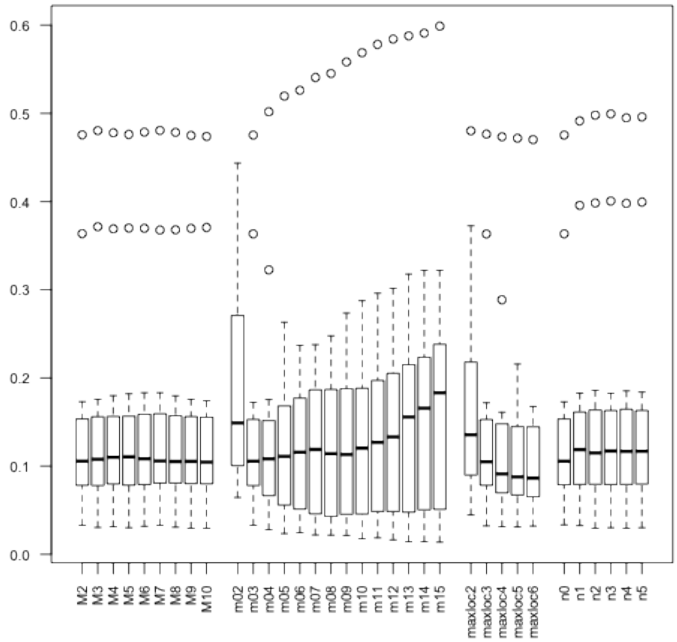


Fig. 6. Effect of different values for *Stacks* core parameters on RAD-locus error rate. Only one parameter varied in each run with the remaining set as explained in Fig. 1.

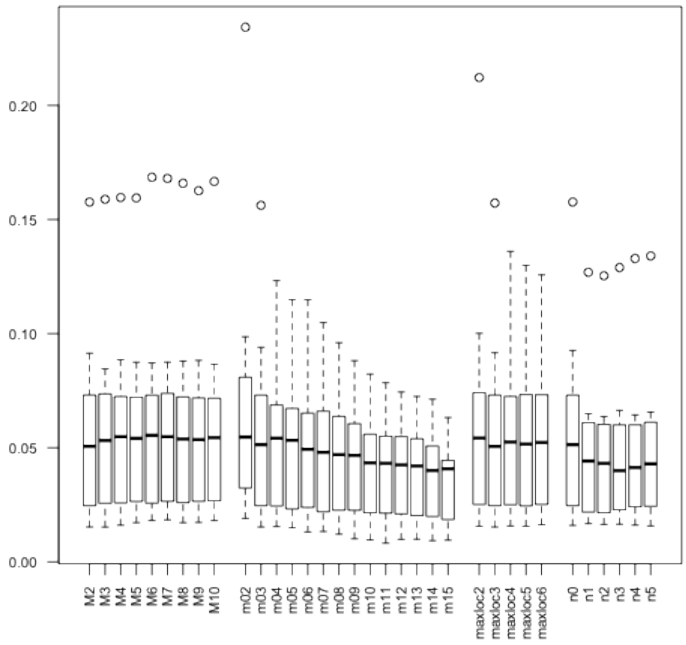


Fig. 7. Effect of different values for *Stacks* core parameters on RAD-allele error rate. Only one parameter varied in each run with the remaining set as explained in Fig. 1.

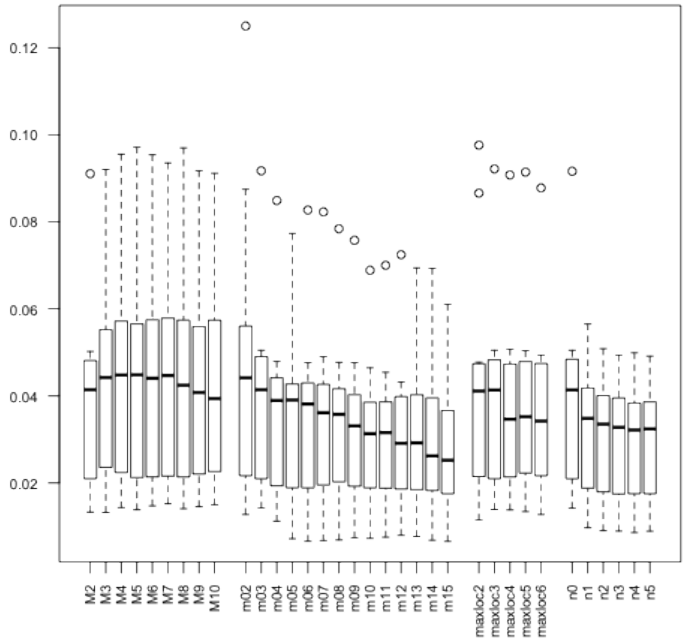


Fig. 8. Effect of different values for *Stacks* core parameters on SNP error rate. Only one parameter varied in each run with the remaining set as explained in Fig. 1.

RAD-seq data yield, error rates and loci filtering

Most samples of all species recovered a substantial (>4,000) number of stacks and negative controls showed a marginal number of sequences after *de novo* assembly with the optimised settings (Fig. 9). Samples and loci were

subsequently filtered to keep only those samples having more than 35% of the mean number of loci per sample and only those loci present in at least 80% the samples.

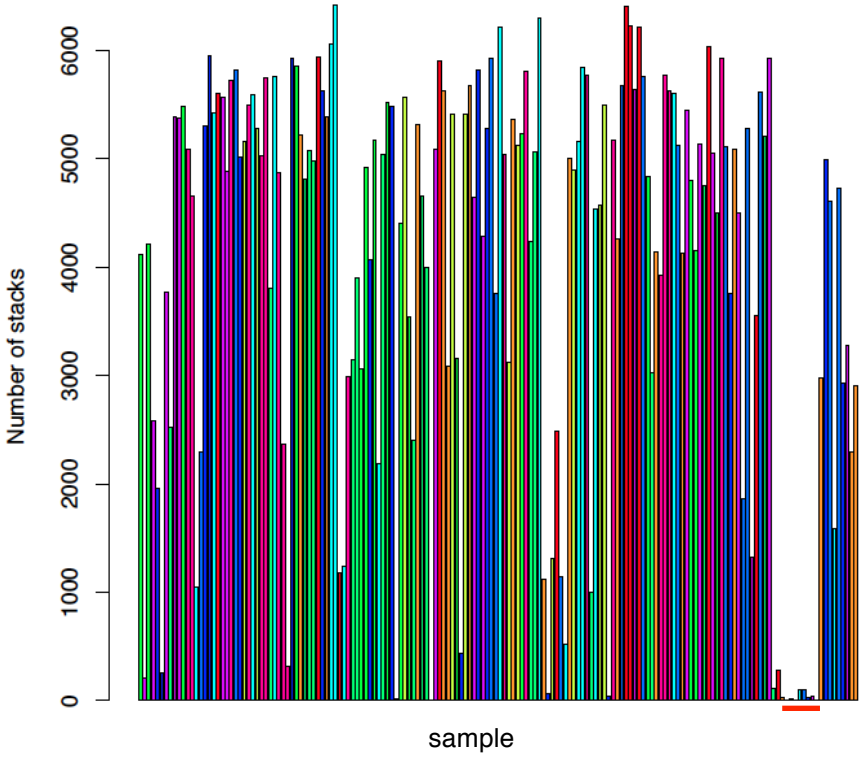


Fig. 9. Number of stacks per sequence-tagged sample. Colours correspond to each ddRAD library (Ju01-Ju10). Negative controls (underlined with red, right end) showed a negligible amount of reads and were discarded in the downstream analyses by the sample selection step (based on proportion shared number of loci among samples).

Fifteen (JmToJ13, JmPpJ05, JmPpJ02, JmPeJ07, JmPeJ06, JmPeJ03, JmPeJ02, JmMaJ06, JmIzJ17, JmIzJ03, JmCoJ01, JmChJ20, JmBlJ02, JmAjJ10, JmAjJ03) out of the 156 samples and replicates did not pass the threshold for numbers of shared loci and were discarded. Final number of samples was 141, with a mean coverage of 73.96 (SD 40.08). There were significantly more coverage in samples from lane Ju01 (Fig. 10), but it was randomly distributed among individuals of different populations (Fig. 11).

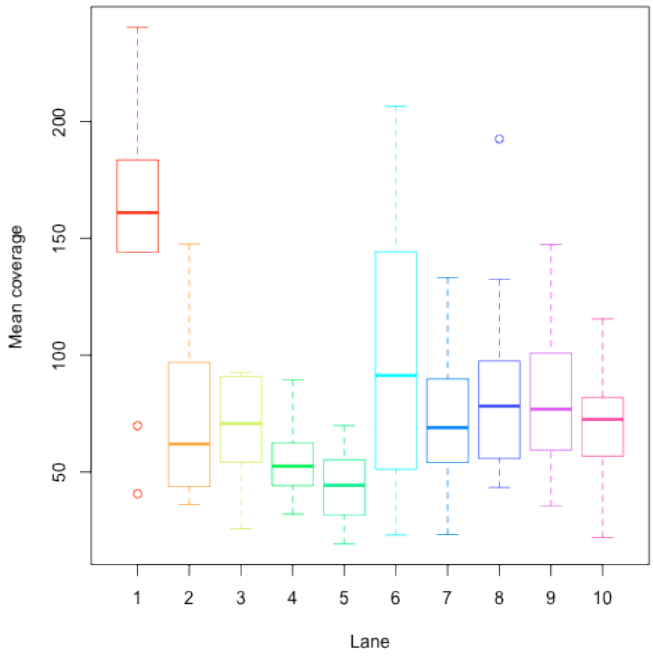


Fig. 10. Mean coverage per retained sample per ddRAd library after running *Stacks de novo* assembly with the optimised parameters.

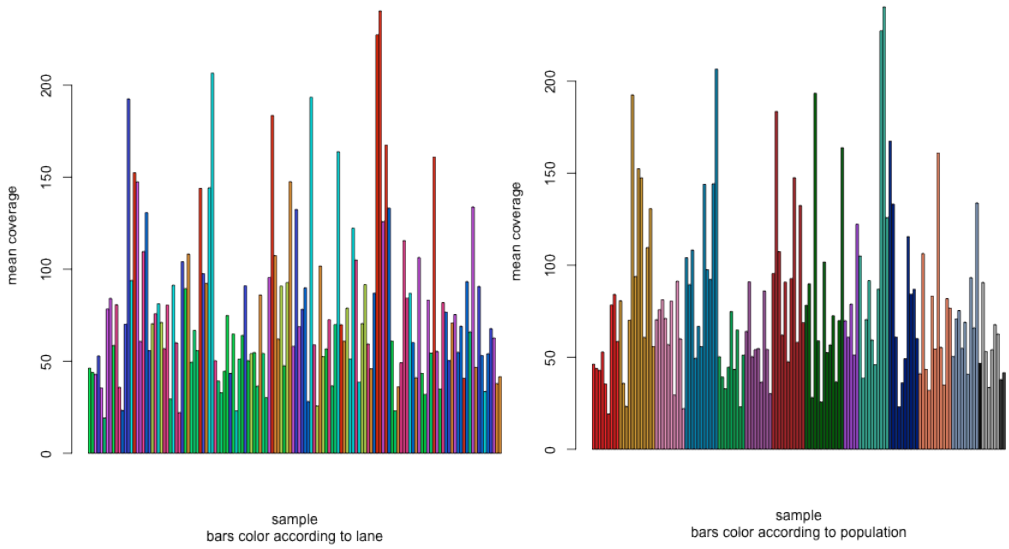


Fig. 11. Mean coverage per retained sample after running *Stacks de novo* assembly with the optimised parameters. Left: coloured by ddRAD libraries (as in Fig 11). Right: coloured by geographic origin of samples.

In total, 6,120 RAD-loci containing 25,823 SNPs were recovered. Error rates for this dataset of loci and samples are 21% (SD 10), 1.9% (SD 2.1) and 2.2% (SD 1.5) for RAD-loci, alleles and SNP error rates, respectively, with 33%

missing data. RAD-loci of this dataset were subsequently examined to identify potential paralogous loci and loci not sufficiently represented among individuals of each sampling location.

To identify potential paralogs and loci not sufficiently represented among individuals of each sampling location the *populations* program of *Stacks* was run to estimate allele frequencies and sampling size per locus per population. The following loci were identified from this output: (1) *putative shared paralogous loci*, defined as loci where the frequency of the major allele equalled p=0.5 in more than one population or species (as implemented in Mastretta-Yanes *et al.* (2014b) and showing deviations from Hardy-Weinberg Equilibrium (HWE, H_{obs} >0.9, negative F_{IS} or F_{IS}=1); (2) *putative paralogous loci private to a single population of J. monticola,* defined as loci where p=0.5 in any single sampling location, present in at least 4 individuals of that population and showing deviations from HWE, and; (3) *RAD-loci not sufficiently represented among individuals within populations,* defined as those that were present in only one individual in any given population. In total, 2,004 RAD-loci met one or more of the previous conditions, out of which 934 were putative shared paralogous loci, 458 putative private paralogous loci and 1,263 were not sufficiently represented among individuals within populations. These 2,004 RAD-loci were blacklisted and filtered from subsequent analyses.

After filtering the blacklisted loci, 3,249 RAD-loci, containing 11,407 SNPs with a mean coverage of 84.60 (SD 50.06) were recovered. Only the first SNP of each RAD-locus was used for population genomics analyses. A total 3,181 SNPs were recovered when the outgroups were included, with a RAD-locus error rate of 21% (SD 15), an allele error rate of 1.8% (SD 2.3), a SNP error rate of 1.5% (SD 1.4) and 18% missing data. In the *J. montiocola* ingroup dataset 2,925 SNPs were recovered, with a RAD-locus error rate of 21% (SD 15), an allele error rate of 1.8% (SD 2.3), a SNP error rate of 1.4% (SD 0.08) and 16% of missing data.

References

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014a) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly

optimization for population genetic inference. *Molecular Ecology Resources*, doi:10.1111/1755-0998.12291

Mastretta-Yanes A, Zamudio S, Jorgensen TH *et al.* (2014b) Gene duplication, population genomics and species-level differentiation within a tropical mountain shrub. *Genome Biology and Evolution*, evu205.

Parchman TL, Gompert Z, Mudge J *et al.* (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest radseq: an inexpensive method for de novo snp discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

II. Modified double digest RAD (ddRAD) sequencing protocol

April 2013

Modifications added by Mastretta-Yanes, A. (University of East Anglia) and Brelsford, A. (Universite de Lausanne), based on the published protocols Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7, e37135. and Parchman, T.L., Gompert, Z., Mudge, J., Schilkey, F.D., Benkman, C.W., and Buerkle, C.A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. Molecular Ecology 21, 2991–3005.

This protocol is more similar to Parchman et al. method because undertakes the size selection after performing the PCR independently in each sample. Peterson et al. protocol undertakes the size selection in an equimolar pool of purified ligation products and then the PCR step is performed.

Summary of modifications specific to this protocol:
Adapters match SbfI restriction enzyme, which is a rare cutter.
Added dual-index barcoding to allow multiplexing >96 samples per library
Modified restriction and ligation mixes to maintain consistent buffer concentration across both steps
Addition of primers and dNTPs for a final thermal cycle, in order to reduce production of single-stranded or heteroduplex PCR products

The present protocol dual indexing barcoding allows to pool 288 samples per library for the price of 61 oligos (24x2 for P1 adapters + 2 for P2 adapter + 1 PCR1 primer + 12 ILLPCR2 primers). It can be easily adapted to a pool of 1,152 samples if using 96 P1 barcoded adapters instead of 24.

Glossary

Adapter: fully or partially double-stranded product of annealing two oligos. Adapters are ligated to genomic DNA at restriction enzyme cut sites in order to add barcodes and common PCR priming sequences.

Barcode: short DNA sequence downstream of the sequencing primer annealing region of an adapter. Used to resolve products of different ligation reactions (usually separate individuals) after sequencing pooled libraries.

Fragment: section of genomic DNA resulting from restriction enzyme cleavage.

Index: short DNA sequence introduced during PCR amplification of the final library that uniquely identifies products of that PCR reaction. Used combinatorially with Adapter P1 barcodes to resolve multiplexed sample pools.

Library: a collection of sequencing-competent fragments.

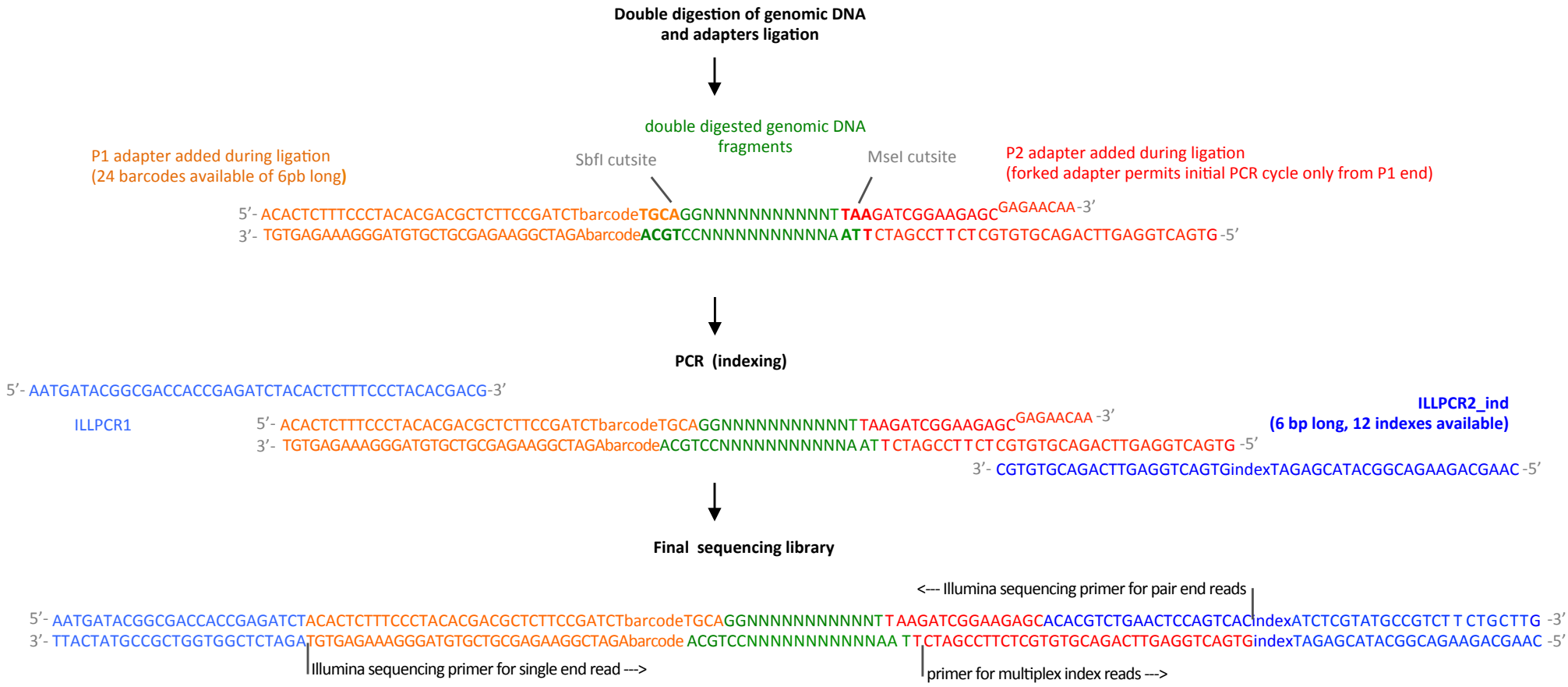


Figure 1. Diagram of library preparation and final structure of sequencing library.

Note on experimental design

Include replicates of some of the samples and a negative control. Randomize the position of the samples in the plate. The negative control should be treated as a normal sample though the whole protocol.

Note on starting DNA material

DNA should ideally be at a minimum concentration of 20 ng/μL and a maximum concentration of 150 ng/μL, but lower concentrations (up to 5 ng/μl) may still work. It is advisable to homogenize sample's concentration before digestion if the variation is orders of magnitude larger.

DNA can be extracted using either a phenol chloroform protocol or a Qiagen extraction kit. Some extractions can carry a salt excess or inhibitors for enzyme activity (e.g. some terpenoids in plant DNA extractions). If such is the case, it is advisable to perform a phenol chloroform DNA cleaning following these steps:

1. To 100 μl of eluted DNA, add 0.5 μl of 20% SDS and 100 μl phenol-chloroform (Sigma Aldrich P2069-100ML)
2. Mix well (vortex gently)
3. Centrifuge at room temperature for 5 min at 14,000 rpm.
4. Pipette the aqueous phase (upper phase, aprox. 80 μl, it is better to leave some DNA than to pipette the organic phase) to a new labeled tube.
5. Discard original tube
6. Add 1/10 volume Na acetate 3M pH 4.8 or 5.2 (i.e. 8 μl for 80 μl DNA solution in this example)
7. Add 2 volumes ethanol 100% (storage -20°C) (i.e. 176 μl in this example)
   Total volume: 264ul, possible with 264 ng. If concentration is below this (1ng/μl), you must add a carrier: glycogen or linear acrylamide.
8. Vortex gently
9. Put on dry ice for 30 min. or over night at -20°C
10. Centrifuge at 4°C for 30 min at 14,000 rpm.
11. Discard the supernatant.
12. Wash with 500ul ethanol 70% (storage 4°C)
13. Centrifuge at 4°C for 5 min at 14,000 rpm.
14. Discard the supernatant
15. Quick spin
16. Pipette out the last drop of ethanol
17. Speed Vac for 3 or 5-7 min at room temperature.
18. Resuspend in 25ul of Tris 10 mM pH 7.5 or 8.0

0. Preparation of adaptors and primers working solutions

P1 adapters:

The P1 adapters consist of 24 barcodes of 6bp long (Table 1) at the end of core sequence ACACTCTTTCCCTACACGACGCTCTTCCGATCT and an overhang (TGCA) at the end of the P1_n.1 oligo that matches the cutsite of SbfI or PstI restriction enzimes (change this overhang to adapt

to other enzimes). The barcodes were designed using the Python script at https://bioinf.eva.mpg.de/multiplex/. In this set of 24, all of the barcodes are separated from each other by at least 3 substitutions.

To prepare the adapter mix for P1 add 98 µl of water to as many PCR wells as P1 barcoded adapters are desired (24 in this case). Then add primer pairs (P1_n.1 with P1_n.2) by mixing 1 µL of each oligo in a pair (100 µM stock). If organized as in Figure 2, from the oligos 100 µM stock plate mix column 1 with column 4, column 2 with column 5, column 3 with column 6 f to generate the plate of annealed 1 µM P1 adapters. Heat to 95ºC for 5 minutes and bring to 20ºC with a ramp of 0.1 ºC/s to slowly cool down. Once they are ready it is possible to freeze it for later use. Keep the set of adaptors organized in plate format that is convenient for later use in setting up reactions.

Plate oligos P1 100µM stock

	1	2	3	4	5	6
A	P1_01.1	P1_09.1	P1_17.1	P1_01.2	P1_09.2	P1_17.2
B	P1_02.1	P1_10.1	P1_18.1	P1_02.2	P1_10.2	P1_18.2
C	P1_03.1	P1_11.1	P1_19.1	P1_03.2	P1_11.2	P1_19.2
D	P1_04.1	P1_12.1	P1_20.1	P1_04.2	P1_12.2	P1_20.2
E	P1_05.1	P1_13.1	P1_21.1	P1_05.2	P1_13.2	P1_21.2
F	P1_06.1	P1_14.1	P1_22.1	P1_06.2	P1_14.2	P1_22.2
G	P1_07.1	P1_15.1	P1_23.1	P1_07.2	P1_15.2	P1_23.2
H	P1_08.1	P1_16.1	P1_24.1	P1_08.2	P1_16.2	P1_24.2

Plate 1µM annealed P1 adapters

	1	2	3
A	P1_01	P1_9	P1_17
B	P1_02	P1_10	P1_18
C	P1_03	P1_11	P1_19
D	P1_04	P1_12	P1_20
E	P1_05	P1_13	P1_21
F	P1_06	P1_14	P1_22
G	P1_07	P1_15	P1_23
H	P1_08	P1_16	P1_24

Figure 2. Oligos and annealed P1 adapters

P2 adapter

The P2 adapters presented here (Table 2) are compatible with MseI. To prepare the working solution mix 100 µL of the P2.1_MseI and P2.2_MseI oligos (100 µM stock) with 800 µL of water to make 1000 µL of 10 pmole/µL (10 µM) stock. Heat to 95ºC for 5 minutes and bring to 20ºC with a ramp of 0.1 ºC/s to slowly cool down. Freeze for later use.

PCR primers

Mix 50 µL of the ILLPCR1 and ILLPCR2_ind oligos (Table 2) with 900 µL of water to make a working solution (5 µM of each oligo). The dual-indexing barcode is incorporated in the

ILLPCR2_ind oligo, so **this step must be repeated for each dual-indexing barcode** (mixing each uniquely barcoded version of ILLPCR2 with ILLPCR1, which will be the same oligo in all working solutions).

Note: If using only 2 indexed primers (i.e. to pool 24x2=48 samples) Illumina recommends to use the ILLPCR2_ind06 and ILLPCR2_ind12. If three primers, use 4, 6, 12. If six primers: 2,4,5,6,7,12.

I.a. Double restriction digest

1. Prepare master mix I (see below, 3 μL prepared per sample), mix and centrifuge. We have found that making 1.2x per sample is sufficient to avoid running out due to high viscosity and/or pipetting error. Work on ice all times.

MASTER MIX I: DIGESTION	
SbfI-MseI	**Vol (μl) 1x**
10X T4 Buffer	0.9
1 M NaCl	0.45
1 mg/mL BSA	0.45
H2O	0.85
MseI (10,000 U/ml)	0.1
SbfI (HF) (20,000 U/ml)	0.25
Total mix volume per sample	3

2. Place 6 μL of sample DNA in each well of a plate.

3. Add 3 μL of the combined master mix I to each well. The total reaction volume should be 9 μL.
4. Cover and seal the plate, centrifuge and incubate at 37°C for 10 hours* on a thermal cycler with a heated lid. Heat kill the enzime with 20 mins at 65ºC. Keep at 4ºC afterwards.

* The digestion time can be reduced to 3 hrs, but if the genome size is large it is advisable to perform the reaction during a long time to ensure complete digestion.

I.b. Fragment analyzer profiles of digestion (optional)

You can confirm the successful of the digestion with a Fragment Analyzer (FA) or Bioanalyzer profile (service usually provided by the sequencing facility). Use replicates of few representative samples perform a single digestion with both enzimes independently (use master mix 1 replacing one enzyme with water) and a double digestion as stated before. This can also be used to estimate the number of fragments that would be expected at different fragment size ranges. See Peterson et al protocol for this.

II. Adaptor Ligation

1. Thaw P1 and P2 adaptors. These adaptors should already be annealed (step 0).

2. Prepare master mix II (see below, 1.6 μL prepared per sample), mix well. As above, it is best to prepare an extra 20% (1.2x/sample).

MASTER MIX II: LIGATION

EcoRI-MseI	**Vol (μl) 1x**
10x T4 Buffer	0.16
1M NaCl	0.13
1 mg/mL BSA	0.13
Water	0.0125
P2 (MseI) adapter 10 uM	1
T4 DNA Ligase (400,000 U/ml)	0.1675
Total mix volume per sample	1.6

3. Add 1.6 μL to each well of the restriction digested DNA.

4. Add 1 μL of the P1 (SbfI) adaptor to each well (a unique barcoded adaptor for each DNA sample).

5. The total reaction volume should now be 11.6 μL. Cover and seal the plate, vortex softly, centrifuge and incubate at 16° C for 6 hours on a thermocycler.

6. Dilute the Restriction-Ligation reaction with 100 μL of Tris 10 mM (or 0.1x TE for long-term storage). Store at 4° C for a month, or -20° C for longer.

III. Purification (optional)

Clean the ligation product with AMPure XP beads following the protocol below, using a magnetic plate or a magnetic tube rack to separate beads from the solution. The Ampure XP original protocol recommends using a volume of beads equal to 1.8X the volume of the solution being cleaned, however ratiod of 1X or 1.5X are successful.

This step is optional, but it reduces the presence of adapter dimers and increases the success of the PCR in samples that otherwise may fail. Also, the AMPure reagent contains polyethylene glycol (PEG), and because higher molecular mass DNA precipitates at lower PEG concentrations than lower molecular mass DNA, the AMPure reagent can be used to discard small fragments. We found that a ratio of 1X keeps fragments >200 bp and a ratio of 0.8X fragments >300. Therefore, the AMPure XP purification can be use to discard adapter dimmers or to perform the size selection instead of doing so with a gel extraction (the range however will be wide, this is only recommended if working with a small genome size).

To perform the purification with the diluted ligation product from the previous step follow the protocol below. The original Agencourt AMPure XP protocol recommends a starting sample volume of 40 μl, but it can be done with 20 μl if pipetting with special care.

On lab bench:

1. Take Agencourt AMPure XP bottle out from the fridge 30 minutes before starting
2. Shake and vortex the Agencourt AMPure XP bottle to fully resuspend magnetic particles.
3. Samples to purify should be ready in a PCR plate (if using magnetic bead) or 1.5ml eppendorf tubes (if using tube rack).
4. Add Sample Vol µl X desired ratio (e.g. 1.5X, 1X or 0.8X) of Agencourt AMPure XP to each sample. Pipette mix 10 times.

Example: to perform a cleaning of 1.5X ratio for a sample volume of 40 µl add an AMPure XP volume of 40 x 1.5 = 60 µl of AMPure XP beads solution.

5. Incubate at room temperature for 5 minutes
6. Place the reaction plate/tubes onto the magnetic plate/rack

On magnetic plate/rack

7. Let it stand for 5 minutes to separate beads from solution.
8. Aspirate the supernatant from the reaction plate and discard (do not disturb the beads)
9. Dispense 200 µl of 70% ethanol (use a fresh preparation) and incubate at room temperature for at least 30 seconds. Aspirate out the ethanol and discard. Repeat for a total of two washes.
10. Wait until the ethanol gets completely dry (5-10 minutes) after the 2nd wash and remove from the magnet

On lab bench

11. Add a volume of elution buffer (Tris 10 mM) equal (or smaller, to concentrate) to the starting sample volume, pipette mix 10 times or until the magnetic particles are fully resuspended (brown color).
12. Incubate at room temperature for 5 minutes.
13. Place the reaction plate/tubes onto the magnetic plate/rack

On magnetic plate/rack

14. Let it stand for 5 minutes to separate beads from solution
15. Transfer the solution to a new plate and label it. This is the purified product. Be careful do not carry over the magnetic particles when aspirating (it is advisable to aspirate 2 µl less than the elution volume).

IV. PCR Amplification

This PCR step uses the Illumina PCR primers to amplify fragments that have our adapters + barcodes ligated onto the ends. To ameliorate stochastic differences in PCR production of fragments in reactions, we run two separate 10 µL reactions per restriction-ligation product (i.e. perform next two steps twice with the same samples), and later combine them. If your sequencing batch includes fewer than 32 individuals, run each PCR at double volume (20 µL) to produce sufficient library quantity.

1. Prepare master mix III (see below, 8 µL per sample, but remember to prepare 2 PCR reactions per sample), vortex and centrifuge. **If you are running the dual-indexing protocol, be sure to prepare separate master mixes for samples to be indexed with different Illumina barcodes- these will each require a different primer mix (see step 0).** Remember, if only 2

index primers will be used use the ILLPCR2_ind06 and ILLPCR2_ind12, if three primers, use 4, 6, 12. If six primers use 2,4,5,6,7,12.

MASTER MIX III: PCR

	Vol (µl) 1x
Water	4.875
Phusion Buffer	2
dNTP (25mM)	0.08
MgCl2 (50 mM)	0.2
PCR Primer Mix	0.67
Phusion Taq	0.1
DMSO	0.075
Total mix volume per sample	8

2. Add 8 µL of the combined master mix III to each well of a plate.
3. Add 2 µL of the diluted ligation product from step II or of the purification product if step III was done.
4. Thermal cycler profile for this PCR: 98° C for 30s; 20 cycles of: 98° C for 20s, 60° C for 30s, 72° C for 40s; final extension at 72° C for 10 min.
5. Prepare master mix IV (see below, 1 µL per sample), **remember to account for dual-indexing primers; they need to be prepared in separate mixes.** It is not necessary to add more polymerase or $MgCl_2$ as there is still enough from the previous PCR. This step reduce production of single-stranded or heteroduplex PCR products.

MASTER MIX IV: PCR final cycle

	Vol (µl) 1x
Water	0.385
Buffer (Phusion)	0.2
PCR primer mix	0.335
dNTP (25 mM)	0.08
Total mix volume per sample	1

6. Add 1 µL to each PCR product (keep cold), run thermocycler profile as follows: 98° C for 3 min, 60° C for 2 min, 72° C for 12 min.

Note: it is advisable to run all the reactions in the same thermocycler.

V. Confirm reaction success of each sample (optional)

Pool equal samples of the two PCR reactions into the same plate ("stack the plates) and run each PCR product on a 1.5% agarose gel for 20-30 minutes. You should see a smear of PCR product

from 150-300 bp (depending on the AMPure ratio used during the purification) to between 500 and 1000 bp, often with a bright band of primer dimer at 130 bp. Samples that failed to amplify, or amplified only the adapter dimer, can be excluded from the pool (except negative controls, those must be pooled).

Note: If there is an evident difference in the yield of some samples compare to others it is possible that the samples with the much brighter smears will take over the sequencing reaction (specially if sequencing a low number of samples). If such seems to be the case, it is advisable to perform a purification as in step III and then measure the concentrations (Qubit) and pooling in equimolar ratios.

VI. Size selection

In this protocol we used an agarose gel extraction to undertake the size selection, but it can also be done using a Blue Pippin (Sage Science) or using different Agencourt AMPure XP ratios (you will need to perform a series of experiments and to send test profiles to the Fragment Analyzer).

In this step it is possible to pool together samples with different index (done with ILLPCR2 primer in step IV), as this guaranties that the size selection will be homogeneous among indexes, however it is possible to perform the size selection independently (but preferably in the same gel, just in separate wells) and to pool together the libraries with different indexes after the purification. Pool in equimolar ratios if following this approach.

To perform the standard agarose gel extraction follow the steps below.

Agarose gel size selection

1. Pool PCR product from both replicates and all samples from into one tube (see note before regarding pooling samples with different indexes). Measure DNA concentration using the Qubit. Depending on the genome size, enzymes and number of samples you should expect a concentration between 8 and 40 ng/µl.
2. Use a SpeedVac (keeping the temperature low) to evaporate the pool of PCR to increase concentration and reduce the number of wells needed tin the gel. Usually a final volume of half the original works fine (do not go below this due to salts overconcentration), but if the original concentration is high the final sample may represent and overload for the gel. The final amount of DNA in the gel should not be larger than 240 ng/mm for a tick gel or 120 ng/mm for a standard. As a guideline, 200 µl of PCR pool at 65 ng/µl + 40 µl LB can be run loading 50-80 µl in 3-4 wells of 18 mm width. More volume will require more wells.
3. Fill a gel rig with new, clean TBE buffer and prepare a 1.5 or 2% agarose gel. Run the pooled PCR product at 100 volts for 2 hours. Include a good ladder on multiple gel lanes so that a clear line can be visualized across the gel, leave an empty lane between the ladder and the library sample. Ethidium bromide in the gel will not interfere after gel purification. A good approach is to tape together several gel combs to allow for larger wells (e.g. tape 5 1.5mm combs to generate a single one of 18 mm width), and to load 50-80 µL of the pool into each well.
4. Cut the desired region out of the gel using the large end of sterile 1000 µl pipette tips or with a sterile razor. We have used the region from 400-500 bp because it will exclude most fragments that consist mostly of adaptor sequence. To minimize gel exposure to UV it is possible to perform the extraction with the UV off by first using it only to mark with a 10 µl

pipette tip the bands of interest in the ladders. Then use a dark straight paper or ruler below the gel bead to create a guide using the ladders marks a reference.

5. Store the excised gel fragments in clean 1.5 or 2 ml colorless eppendorf tubes (ensure tube size will be enough for QG buffer and isopropanol volume added in next steps). Proceed to extraction purification or store at 4ºC until then.

Extraction purification

The following steps use the QIAquick Minielute Gel Extraction Kit with modifications in the incubation and centrifuge conditions.

6. Weigh the gel slice (tare an empty tube first, then weight the one from step 5. Add 3 volumes of Buffer QG to 1 volume of gel (100 mg gel ~ 100 µl). The maximum amount of gel slice per spin column is 400 mg. For >2% agarose gels, add 6 volumes Buffer QG.
7. Incubate at 22°C for 30 min or until the gel slice has completely dissolved. This enriches GC bonds. Vortex gently the tube every 2–3 min during incubation to help dissolve the gel.
8. After the gel slice has dissolved completely, check that the color of the mixture is yellow (similar to Buffer QG without dissolved agarose). If the color of the mixture is orange or violet, add 10 µl 3 M sodium acetate, pH 5.0, and mix. The color of the mixture will turn to yellow. Note: if your gel slice contained LB the mixture color may change due to the LB pigment and not because of a pH change, so it is not necessary to add sodium acetate.
9. Add 1 gel volume of isopropanol to the sample and mix by inverting.
10. Place a MinElute spin column in a provided 2 ml collection tube.
11. Apply sample to the MinElute column and centrifuge for 1 min at 10,000 rpm.
12. Discard flow-through and place the MinElute column back into the same collection tube. For sample volumes of more than 800 µl, simply load and spin again.
13. Add 500 µl Buffer QG to the MinElute column and centrifuge* for 1 min at 10,000 rpm.
14. Discard flow-through and place the MinElute column back into the same collection tube.
15. Add 750 µl Buffer PE to MinElute column. Let the column stand 2–5 min after addition of Buffer PE.
16. Centrifuge* for 1 min at 10,000 rpm.
17. Discard flow-through and place the MinElute column back into the same collection tube.
18. Centrifuge the column in a 2 ml collection tube (provided) for 1 min. Residual ethanol from Buffer PE will not be completely removed unless the flow-through is discarded before this additional centrifugation.

If more than one column was used to purify a gel extract from the same library, perform the following steps independently with each column in the same same eppendorf tube.

19. Place the MinElute column into a clean 1.5 ml eppendorf tube. To elute DNA, add 10 µl Buffer EB (10 mM Tris·Cl, pH 8.5) to the center of the MinElute membrane. (Ensure that the EB is dispensed directly onto the membrane for complete elution of bound DNA.)
20. Let the column stand for 1 min, and then centrifuge the column for 1 min.

* To increase the amount of DNA recovered it is advisable to increase the speed gradually. If the centrifuge does not has this option it can be done by first centrifuging at around 2,000 rpm for few seconds, then stopping it, centrifuging at around 5,000 rpm for few seconds, stopping again and finally centrifuging at the desired revolutions and time (10,000 rpm for 1 min in this case).

VI. Preparing final template for Illumina sequencing

1. Measure DNA concentration with the Qubit.
2. Perform an ethanol precipitation to increase concentration and remove excess salts. Note: the concentration of DNA in the precipitation solution (i.e. library solution + NaAc + 100% ethanol) should be a minimum of 1 ng/µl, otherwise it would not precipitate and will be lost.
   A. Add 1/10 volume 3 M sodium acetate 3M pH 4.8 or 5.2 (e.g. 2ul for 20 µl DNA solution)
   B. Add 2 volumes of 100% ethanol (molecular biology grade) stored at -20°C, chill in dry ice for 30 min or overnight in a -20ºC freezer.
   C. Centrifuge at max speed for 15 minutes, remove supernatant carefully.
   D. Add 200 µL 70% Ethanol (diluted from absolute, not technical)
   E. Centrifuge 10 minutes, remove supernatant
   F. Dry DNA Pellet
   G. Resuspend using 20-40 µL of Tris 10 mM or TE
3. Measure concentration again. A total concentration of >25 ng/µL is ideal for Illumina sequencing, but we can go as low as 2 ng/µL.
4. Make an aliquot of the library and submit it to Fragment Analyzer or Bioanalyzer. You should expect to see a curve with a peak in the middle of the range of the size selection. A peak around 130 bp indicates that there was primer dimer carry over. It is possible to perform a 0.9X or 1X ampure purification to discard the primer dimers, but if the peak is small relatively to the library, it is possible to sequence the as it is, as they will represent a small percentage of the total reads.
5. If the Fragment Analyzer profile and concentration are the desired the library is now ready for sequencing. The library can be submitted for sequencing in a Illumina HiSeq2000 (or similar) system in a single or pair-end run. The index sequencing is done separately from the insert sequencing, and the index sequence is not effected by the insert length, so it is not necessary to run the pair end to get the indexes sequence. If you used this protocol with more than one index, then you will be asked by the sequencing facility to provide their ID and sequence (Table 2) so that they can demultiplex the reads by index. Then your pipeline will have to include a second demultiplexing step to separate the reads by individual. Happy sequencing.

Table 1. Oligos sequence for P1 adapters

barcode #	barcode sequence	reverse complement	ID	P1_SbfI_n.1 sequence (5'-3')		ID	P1_SbfI_n.2 sequence
1	GGTCTT	AAGACC	P1_SbfI_01.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTCTTTGCA		P1_SbfI_01.2	AAGACCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
2	CTGGTT	AACCAG	P1_SbfI_02.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGGTTTGCA		P1_SbfI_02.2	AACCAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
3	AAGATA	TATCTT	P1_SbfI_03.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGATATGCA		P1_SbfI_03.2	TATCTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
4	ACTTCC	GGAAGT	P1_SbfI_04.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTCCTGCA		P1_SbfI_04.2	GGAAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
5	TTACGG	CCGTAA	P1_SbfI_05.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTACGGTGCA		P1_SbfI_05.2	CCGTAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
6	AACGAA	TTCGTT	P1_SbfI_06.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACGAATGCA		P1_SbfI_06.2	TTCGTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
7	ATTCAT	ATGAAT	P1_SbfI_07.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATTCATTGCA		P1_SbfI_07.2	ATGAATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
8	CCGACC	GGTCGG	P1_SbfI_08.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGACCTGCA		P1_SbfI_08.2	GGTCGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
9	ATCGTC	GACGAT	P1_SbfI_09.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCGTCTGCA		P1_SbfI_09.2	GACGATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
10	CATCAA	TTGATG	P1_SbfI_10.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCATCAATGCA		P1_SbfI_10.2	TTGATGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
11	GCCTGG	CCAGGC	P1_SbfI_11.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTGGTGCA		P1_SbfI_11.2	CCAGGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
12	TGCTTG	CAAGCA	P1_SbfI_12.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCTTGTGCA		P1_SbfI_12.2	CAAGCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
13	TCGCAT	ATGCGA	P1_SbfI_13.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGCATTGCA		P1_SbfI_13.2	ATGCGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
14	GGTAGA	TCTACC	P1_SbfI_14.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTAGATGCA		P1_SbfI_14.2	TCTACCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
15	GGAGCG	CGCTCC	P1_SbfI_15.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAGCGTGCA		P1_SbfI_15.2	CGCTCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
16	TTGAAC	GTTCAA	P1_SbfI_16.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGAACTGCA		P1_SbfI_16.2	GTTCAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
17	GATTAC	GTAATC	P1_SbfI_17.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATTACTGCA		P1_SbfI_17.2	GTAATCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
18	CGAGGC	GCCTCG	P1_SbfI_18.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGAGGCTGCA		P1_SbfI_18.2	GCCTCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
19	CAACCG	CGGTTG	P1_SbfI_19.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAACCGTGCA		P1_SbfI_19.2	CGGTTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
20	GTATGA	TCATAC	P1_SbfI_20.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTATGATGCA		P1_SbfI_20.2	TCATACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
21	TGGATT	AATCCA	P1_SbfI_21.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGATTTGCA		P1_SbfI_21.2	AATCCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

22	CCAGCT	AGCTGG	P1_SbfI_22.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCAGCTTGCA		P1_SbfI_22.2	AGCTGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
23	AACTCG	CGAGTT	P1_SbfI_23.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACTCGTGCA		P1_SbfI_23.2	CGAGTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
24	ACCAGA	TCTGGT	P1_SbfI_24.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCAGATGCA		P1_SbfI_24.2	TCTGGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

These oligos do not include a protective base, but it is possible to add a C after the barcode and before the restriction enzime overhang. Order sequences as unmodified oligos with HPSF purification.

Table 2. Oligos sequence (5'-3') for P2 adapter and PCR primers.

P2.1_MseI	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	
p2.2_MseI	/5Phos/TAAGATCGGAAGAGCGAGAACAA	
ILLPCR1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG	**Index sequence ****
ILLPCR2_ind01	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGC	ATCACG
ILLPCR2_ind02	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGC	CGATGT
ILLPCR2_ind03	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGC	TTAGGC
ILLPCR2_ind04	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGC	TGACCA
ILLPCR2_ind05	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGC	ACAGTG
ILLPCR2_ind06	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGC	GCCAAT
ILLPCR2_ind07	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGC	CAGATC
ILLPCR2_ind08	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGC	ACTTGA
ILLPCR2_ind09	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGC	GATCAG
ILLPCR2_ind10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGC	TAGCTT
ILLPCR2_ind11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGC	GGCTAC
ILLPCR2_ind12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGC	CTTGTA

Modifications key: /5Phos/ = 5' phosphate. Note: it is optional to add phosphorothioate bonds (*) to the first two bases of the PCR2 primers to add resistance to degradation by exonucleases. Order HPSF purification for the unmodified oligos and HPLC for the modified. ** As needed for demultiplexing, the reverse complement of each sequence is inside the ILLPCR2 primer.

Reagents and estimated cost

Table 3. Summary of reagents and services prices to perform the ddRAD protocol in 288 samples as to April 2013.

Item	Company	Catalog	Size/volume	Unit cost (GBP)	Units needed for library of 288 samples	Cost for library of 288 samples (GBP)
Oligos for PCR primers*	Eurofims	custom order	unmodified oligo HPSF purification	£12.72	12	£152.64
Oligos for P1 adapters*	Eurofims	custom order	Unmodified Plate Oligos HPSF (48 oligos for 24 adapters)	£302.88	1	£302.88
Oligos for P2 adapters*	Eurofims	custom order	Unmodified oligo HPSF purification (P2.1) and /Pho modified oligo HPLC (P2.2)	£30.00	1	£30.00
SbfI (HF)	NEB	R3642S	500 units at 20,000 units/ml	£41.63	1	£41.63
MseI	NEB	R0525S	500 units at 10,000 units/ml	£39.15	1	£39.15
T4 DNA Ligase	NEB	M0202S	20,000 units at 400,000 cohesive end units/ml	£39.15	1	£39.15
Phusion Taq	NEB	M0530S	100 units	£64.00	2	£128.00
QIAquick Gel Extraction Kit (50)	Quiagen	28704	50 reactions	£79.30	1	£79.30
Qubit® dsDNA HS Assay Kit**	Life-Technologies	Q32854	500 assays, 0.2–100 ng	£136.00	1	£136.00
Ampure XP***	Agencourt	A63880	5 ml	£272.00	1	£272.00

Phenol-Chloroform (optional)	Sigma-Aldrich	P2069-400ML	100 µl	£195.00	1	£195.00
200 µl filter tips	StarLab	S1120-8810	10 x 96-Tip Sterile Racks	£91.12	2	£182.24
10 µl filter tips extended length	StarLab	S1120-3810	10 x 96-Tip Sterile Racks	£91.12	2	£182.24
10000 µl filter tips	StarLab	S1126-7810	10 x 96-Tip Sterile Racks	£97.92	1	£97.92
Subtotal without sequencing						**£2,209.68**
Lane HiSeq2000	GTF-Lausanne		Single pair	£1,360.00	Depends on genome size, enzymes and number of samples pooled.	
Total if sequencing 288 samples in 1 lane (maximum pool with these number of adapters)						**£3,569.68**
Total if sequencing 288 samples in 12 lanes (no pooling, will guaranty high coverage, likely more than needed)						**£18,529.68**

Notes: price of ethanol, isopropanol, Tris and agarose are not included because they are common molecular biology reagents normally bought in much larger quantities than what is needed for this protocol. * Oligos have to be bought only once and then can be used for many libraries, price can be shared among different projects/labs. **Requires having the Qubit flourometer (£1,518.44). DNA concentration can be done instead with Picogreen, but notwith for Nanodrop. ***Requires having a magnetic rack or plate (£200-400)

So long story short:

Biodiversity rocks, Mexican mountains are awesome and genomes are wibbly wobbly... time-y wimey... stuff.