Identification and analysis of a transcriptome of Douglas-fir (*Pseudotsuga menziesii*) and population structure inference using different next-generation sequencing techniques

Dissertation zur Erlangung des Doktorgrades
der Naturwissenschaften (Dr. rer. nat.)

Fakultät Naturwissenschaften
Universität Hohenheim

Institut für Pflanzenzüchtung, Saatgutforschung und Populationsgenetik (350)
und
Institut für Physiologie und Biotechnologie der Pflanzen (260)

vorgelegt von

Thomas Müller

aus Bayreuth

Stuttgart, 2015

Dekan	Prof. Dr. Heinz Breer
1. berichtende Person:	Prof. Dr. Karl J Schmid
2. berichtende Person:	Prof. Dr. Andreas Schaller
3. Prüferin:	Prof. Dr. Waltraud Schulze
Eingereicht am:	06.02.2015
Mündliche Prüfung am:	15.06.2015

Die vorliegende Arbeit wurde am 17.04.2015 von der Fakultät Naturwissenschaften der Universität Hohenheim als „Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften“ angenommen.

Contents

List of Figures

List of Tables

I. Zusammenfassung

Vorhersagen und Simulationen gehen davon aus, dass sich die klimatischen Bedingungen in Zentraleuropa in den kommenden Jahrzehnten entscheidend verändern werden. Es werden längere Trockenperioden und weniger Niederschläge im Sommer erwartet. Im Gegensatz zu Tieren können Pflanzen ihren Standort nicht wechseln, so dass sie sich an die neuen Gegebenheiten adaptieren oder über ihre Nachkommen neue ökologische Nischen besiedeln müssen. Aufgrund der langen Generationszeit bei Bäumen ist es hier besonders wichtig zu wissen, ob und wie sie mit den erwarteten klimatischen Bedingungen umgehen können. Förster machen sich bereits heute über die Zusammensetzung künftiger Wälder Gedanken, da Baumarten und Populationen ausgewählt werden müssen, die mit dem veränderten Klima keine oder nur wenig Probleme haben.

Die Douglasie (*Pseudotsuga menziesii*) ist hierbei eine vielversprechende Baumart, da sie sich in ihrem natürlichen Verbreitungsgebiet in Nordamerika an unterschiedliche Habitate und Klimazonen adaptiert hat. Sie kann in zwei Varietäten unterteilt werden, die Küsten- und die Inlandsdouglasie, die sich genotypisch und phänotypisch, z.B. in der Trockentoleranz, unterscheiden. Bei Anbauversuchen in Deutschland zeigten die Bäume, hauptsächlich Küstendouglasien, sehr gute Wuchsleistungen. Daher wurde ein Forschungsprojekt, "DougAdapt", konzipiert, das genotypische und phänotypische Unterschiede zwischen verschiedenen Provenienzen der Küsten- und Inlandsdouglasien analysieren und den Einfluss des Genotyps auf den Phänotyp untersuchen sollte. In dem Projekt wurden Bäume aus Feldversuchen in Deutschland und aus Gewächshausexperimenten beprobt.

Um die genetische Diversität der Provenienzen zu untersuchen, war es zunächst nötig, Referenzsequenzen zu generieren, da es, bis auf eine begrenzte Anzahl von Genen, keine Sequenzinformation für die Douglasie gab. Selbst mit kostengünstigen modernen Sequenziertechnologien ist es sehr teuer das $\sim$19 Gigabasen große Genom der Douglasie vollständig zu erfassen und zu entschlüsseln. Eine Alternative stellt die Transkriptomsequenzierung dar, bei der nur Gene, d.h. die für Proteine codierenden Bereiche des Genoms, sequenziert werden. Die in dieser Arbeit erstmals für Douglasie durchgeführte Transkriptomsequenzierung resultierte in einer großen Anzahl Referenzsequenzen, die, wie Vergleiche mit bekannten Transkriptomen anderer Pflanzenarten zeigten, das Transkriptom umfänglich repräsentieren. Durch die Verwendung von Setzlingen, die zuvor unter kontrollierten Bedingungen im Rahmen eines Trocken-

stressexperimentes aufwuchsen, war es des Weiteren möglich, Kandidatengene zu identifizieren, die vermutlich bei der Reaktion der Bäume auf Trockenstress von Bedeutung sind. Darüber hinaus konnten mehr als 27,000 bis dahin unbekannte Punktmutationen (single nucleotide polymorphisms, SNPs) in Douglasien detektiert werden. SNPs können großen Einfluss auf den Phänotyp eines Individuums haben und werden beispielsweise als Marker oder zur Analyse der genetischen Diversität verwendet.

Die Analyse der genetische Diversität in Douglasienprovenienzen und die Suche nach Genen, die wahrscheinlich an der lokalen Anpassung der Bäume beteiligt sind, wurde mit Hilfe eines *sequence capture* Experiments durchgeführt. Dabei werden nur vorab definierte Bereiche eines Genoms sequenziert, in diesem Fall die potentiellen Trockenstresskandidatengene sowie ungefähr 57,000 weitere potentielle Gensequenzen. Wir konnten nachweisen, dass *sequence capture* basierend auf Transkriptomsequenzen in Arten mit einem großem und weitestgehend unbekannten Genom anwendbar ist. Obwohl die polymorphen Kandidatengene für Trockenstress einen höheren Grad an genetischer Differenzierung aufwiesen als die restlichen Gene, waren sie nicht unter den gefundenen Kandidatengenen die vermutlich unter positiver Selektion sind. Letztere wiederum spielen wahrscheinlich eine Rolle bei der lokalen Anpassung der Bäume. Trotz eines starken Genflusses zwischen Inlands- und Küstenprovenienzen zeigten die SNP-Daten eine genetische Differenzierung zwischen beiden Varietäten, aber nur eine sehr geringe Differenzierung innerhalb der Küstenpopulationen.

Eine weitere Studie untersuchte die Anwendbarkeit von *Genotyping-by-sequencing* (GBS) in Douglasien und verglich die Ergebnisse zweier GBS Versuche mit dem *sequence capture* Experiment. In GBS wird ein Genom durch ein oder mehrere Restriktionsenzyme verdaut. Anschließend werden nur Fragmente einer bestimmten Länge sequenziert, was den Anteil des Genoms, der sequenziert wird, und dadurch auch die Kosten erheblich reduziert. Der Vorteil gegenüber *sequence capture* liegt darin, dass mit weniger Aufwand und Kosten mehr Individuen parallel beprobt werden können. Wir konnten zeigen, dass ein Verdau mit zwei Restriktionsenzymen mehr SNPs mit weniger fehlenden Daten ergibt, als ein Verdau mit einem Restriktionsenzym. Im Vergleich zum *sequence capture* wurden in beiden GBS deutlich weniger SNPs detektiert. Dennoch war es mit den SNP-Daten aus beiden GBS Ansätzen möglich, südliche Inlands-, nördliche Inlands- und Küstendouglasienprovenienzen zu unterscheiden. GBS, insbesondere mit zwei Restriktionsenzymen, stellt einen vielversprechenden Ansatz dar, um eine

große Anzahl Douglasien kostengünstig zu genotypisieren und um SNPs zu erhalten, die für verschiedene Zwecke, z.B. genomweite Assoziationsstudien, verwendet werden können.

In dieser Arbeit wurden eine große Anzahl Sequenzdaten und SNPs des Douglasiengenoms analysiert. Mit zusätzlichen phänotypischen Informationen werden diese Daten bei der Analyse nützlicher Eigenschaften wie Trockentoleranz von großer Bedeutung sein. Die hier gewonnenen Informationen über das Douglasiengenom und die genetische Diversität zwischen unterschiedlichen Provenienzen können außerdem in Züchtungsprogrammen und Assoziationsstudien verwendet werden, die wiederum bei der Auswahl optimaler Provenienzen für bestimmte Standorte hilfreich sein können.

II. Abstract

Simulations and predictions assume severe changes in the climatic conditions in Central Europe in the coming decades. Longer periods of drought and less precipitation during summer are expected. In contrast to animals, plants cannot change their habitat and have to adapt to the new conditions or their offspring has to colonize new ecological niches. Due to the long generation times in trees it is important to know if and how trees can cope with the expected climatic conditions. Forest managers already give thought to the composition of future forests, because they have to choose species and populations which have no or only few problems with the changed climate.

Douglas-fir (*Pseudotsuga menziesii*) is a promising tree species for this purpose, because it is adapted to different habitats and climate zones in its natural distribution range in North America. The two main varieties, coastal and interior Douglas-fir, differ genotypically and phenotypically, e.g. in drought tolerance. Douglas-fir trees, mainly of the coastal variety, showed good growth performances in field trials in Germany. Hence, a research project called "DougAdapt" was designed to analyze and to link genotypic and phenotypic differences in several coastal and interior Douglas-fir provenances. In this project, trees from field trials and from greenhouse experiments were sampled.

To analyze the genetic diversity of the provenances we first generated reference sequences, because with the exception of some genes, there was no reference sequence information available for Douglas-fir. Even with modern and cost-efficient next-generation sequencing technologies it would be very expensive to decipher the $\sim$ 19 gigabases of the Douglas-fir genome completely. An alternative to whole genome sequencing is transcriptome sequencing, in which only the coding regions of a genome are sequenced. The transcriptome sequencing, which was performed for the first time in Douglas-fir, resulted in a large number of putative unique transcripts (PUTs). Comparisons with published transcriptomes of other plant species showed that the PUTs represented the transcriptome of Douglas-fir comprehensively. As the sampled seedlings were part of a drought stress experiment and grew under controlled conditions, we were able to identify drought related candidate PUTs, which may be part of the trees' response to drought. Furthermore, more than 27,000 previously unknown single nucleotide polymorphisms (SNPs) in Douglas-fir could be identified. SNPs can influence the phenotype of individuals, and they can be used for instance as markers or to analyze genetic diversity.

The analysis of genetic diversity of Douglas-fir provenances and the search for genes which may be part of the local adaptation of the trees were performed with a sequence capture experiment. In sequence capture only predefined regions of a genome are sequenced, in this case the drought related candidate PUTs as well as approximately 57,000 further PUTs. We showed that sequence capture based on PUTs as target regions is applicable in species with large and mostly unknown genomes. The polymorphic drought related candidate PUTs showed higher genetic differentiation than the remaining genes. Nevertheless, none of them was among the candidate PUTs for positive selection, which in turn are probably part of the local adaptation of the trees. Despite a high level of gene flow between coastal and interior provenances, the SNP data showed genetic differentiation between both varieties but only very low differentiation between the coastal provenances.

We also investigated if genotyping-by-sequencing (GBS) is a suitable method to detect polymorphisms in Douglas-fir and compared the results of two GBS experiments with the sequence capture. The genome is digested with one or several restriction enzymes in GBS. Afterwards, only fragments with a specific length are sequenced, which considerably reduces the part of the genome that is sequenced as well as the costs. The advantage compared to sequence capture is the possibility to sample more individuals at the same time with less effort and costs. We showed that a digestion with two restriction enzymes results in more SNPs with less missing data, compared to using only one restriction enzyme. Both GBS methods returned considerably less SNPs than the sequence capture. Nevertheless, it was possible to distinguish between southern interior, northern interior, and coastal provenances using SNP data of the GBS experiments. GBS, especially with two restriction enzymes, seems to be a promising approach to genotype a large number of Douglas-fir trees and to obtain SNPs at low costs, which can be used in several tasks like genome-wide association studies.

A large amount of sequence data and SNPs were analyzed in this thesis. Together with phenotypic information, these data will be crucial for the analysis of useful traits in Douglas-fir, like drought tolerance. Furthermore, the results concerning the Douglas-fir genome and the genetic diversity of different provenances will be beneficial in breeding programs and association studies, which in turn can be helpful to choose the optimal provenances for a given location.

1. General introduction

> A forest of these trees is a spectacle too much for one man to see.
>
> *(David Douglas)*

The Intergovernmental Panel on Climate Change (IPCC) expects increasing summer temperatures and decreasing precipitation in Central Europe in the future (IPCC, 2007). Simulations and further studies support these predictions (Brohan et al, 2006, Fink et al, 2004, Meehl and Tebaldi, 2004). Because the expected changes will have a strong influence on the forest landscape and due to long generation times of trees, forest managers need to know which trees are able to cope with future conditions. Douglas-fir, which is adapted to many different ecozones in its natural habitat in North America, seems to be a promising species for this purpose.

This thesis analyzes the transcriptome of Douglas-fir and the genetic variations in several provenances of coastal and interior Douglas-fir.

1.1. About Douglas-fir

Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) is an evergreen, monoecious conifer and one of the most important and valuable timber trees world wide. During the Tertiary period the species *Pseudotsuga* was present in Europe, but became extinct during the Ice Age (Konnert et al, 2008). Due to the north-south orientation of the mountains in North America, the species was able to migrate south and to survive there. After the Ice Age, Douglas-fir re-populated the north from possibly two centers, one at the Columbia Valley and one in the Rocky Mountains (Halliday and Brown, 1943, Tsukada, 1982).

The scientific name, *Pseudotsuga menziesii*, refers to the Scottish botanist Archibald Menzies (*15 March 1754 – †15 February 1842), who discovered the tree in 1792 on Vancouver Island (Aas, 2008, Menzies et al, 1923). He brought dried parts of the plants back to England, where

they were described systematically by Aylmer Bourke Lambert in 1803 (Hermann, 1981, Kownatzki et al, 2011). The common name, Douglas-fir, refers to David Douglas (*25 June 1799 – †12 July 1834), another Scottish botanist and rival of Menzies, who also collected seeds and specimens of the trees in the 1820s and sent them to England. In contrast to Menzies' attempts to send seeds to Europe, the samples of Douglas arrived and were planted in England. Therefore, Douglas is considered the person who introduced Douglas-fir to Europe (together with more than 200 further plant species; Aas, 2008, Douglas, 1914).

1.1.1. Origin of Douglas-fir

Douglas-fir occurs in two main varieties in its natural distribution range in North America (Figure 1.1). The coastal or green Douglas-fir (*P. menziesii* var. *menziesii*) is distributed over 2,200 km along the American west coast from British Columbia, Canada, to central California, USA. The interior, blue, or Rocky Mountain Douglas-fir (*P. menziesii* var. *glauca*) extends over a range of 4,500 km along the Rocky Mountains from Alberta, Canada, to Colorado, USA, down to Mexico. The distribution of Douglas-fir is rather continuous, but there are also many isolated populations, especially in the southern part of the habitat of interior trees. Generally, coastal Douglas-fir grows from sea-level up to 1,520 m, but this variety can also be found at altitudes of 2,300 m in the Sierra Nevada. Interior Douglas-fir grows at higher elevations than the coastal variety. In the northern part of its distribution range it grows from 550 to 2,440 m and in the Rocky Mountains usually from 1,830 to 2,900 m, but it was also reported from stands at Mount Graham at an elevation of 3,260 m (Aas, 2008, Hermann and Lavender, 1990). No clear distinction can be drawn between coastal and interior varieties in Canada, probably due to interbreeding between the two in the *menziesii-glauca* transition zone (Eckert et al, 2009a, Kohnle et al, 2012).

Coastal and interior Douglas-fir differ in several phenotypical and morphological characteristics, which can be used to distinguish the varieties (Table 1.1). Furthermore, differences for example in growth rate, terpene content of needle oil, nuclear characteristics, or bark thickness were reported (El-Lakany and Sziklai, 1970, Hermann and Lavender, 1990, Kohnle et al, 2012, Rudloff, 1972). With allozyme and RAPD markers it was possible to further subdivide the interior variety into a northern and a southern subgroup (Aagaard et al, 1998, Li and Adams, 1989). Possibly as a consequence of interbreeding, the coastal populations are genetically more similar to northern than to southern interior populations.

Figure 1.1: Natural habitat of coastal (light-green) and interior (dark-green) Douglas-fir in North America. Map modified from a map provided by USGS with data from Little (1971).

1.1.2. Adaptation and population genomics

A process that leads to adaptation is positive selection (Darwin, 1859). Positive selection occurs if random mutations cause genomic variations which are beneficial for individuals carrying the new allele. As a consequence, the frequency of the advantageous allele increases in the population, which may eventually lead to fixation. Therefore, populations may adapt to different natural conditions due to an interplay of genomic variations caused by random mutations and positive selection.

To identify candidate genes for positive selection F_{ST} outlier tests, as implemented in Lositan or BayeScan, can be used (Chapter 3; Antao et al, 2008, Beaumont and Nichols, 1996, Foll and Gaggiotti, 2008). F_{ST} is a measure of population differentiation, which is based on allele frequencies (Weir and Cockerham, 1984, Wright, 1951). Another possibility to detect loci under selection is to use Tajima's D to estimate whether observed DNA sequences evolve under a neutral model (Chapter 3; Tajima, 1989). Departures of Tajima's D values from the neutral model can be caused by selection, but also by non-selective forces like demographic processes.

Table 1.1.: Phenotypical and morphological differences of *P. menziesii* var. *menziesii* and *P. menziesii* var. *glauca* (modified from Aas (2008))

	var. *menziesii*	var. *glauca*
Height	up to 80 m (max. 120 m)	up to 40 m (max. 48 m)
Diameter at breast height	up to 2.2 m (max. 4.9 m)	up to 0.9 m (max. 2.2 m)
Needles	yellow- to dark-green	gray- to blue-green
Cones	6 – 11 cm	4 – 8 cm
Cold hardiness	lower	higher
Shade tolerance	lower	higher
Susceptibility to *Rhabdocline pseudotsugae*	lower	higher

For Douglas-fir and other conifers, only a small number of candidate genes for positive selection were reported (Eckert et al, 2009d, Krutovsky and Neale, 2005, Palmé et al, 2008). However, the large natural habitat reflects Douglas-fir's ability to adapt to different climatic conditions. Depending on elevation or geographic location, the populations differ for example in cold hardiness (Darychuk et al, 2012, Rehfeldt, 1979), or drought tolerance (Darychuk et al, 2012, Jansen et al, 2013, Martinez-Meier et al, 2009, Pharis and Ferrell, 1966). Due to the adaptation to different ecozones, a variety of genetically diverse populations evolved (Campbell, 1979, Kleinschmit and Bastien, 1992, Rehfeldt, 1979, 1989). Nevertheless, no or only a weak population structure was found within coastal Douglas-fir populations using allozymes, RAPDs, or microsatellite markers (Aagaard et al, 1998, Krutovsky et al, 2009, Li and Adams, 1989, Viard et al, 2001).

1.1.3. Experimental sites in Germany

Douglas-fir was first planted in Germany in 1831 in the arboretum of Flottbek close to Hamburg by John Richmond Booth (Kownatzki et al, 2011). The first plantations were coastal Douglas-firs, which grew very well and resulted in healthy stands (Kleinschmit and Bastien, 1992, Schober, 1972). It was soon realized that the origin of the trees was correlated with their performance in Europe (Schwappach, 1907). In general, coastal Douglas-firs grew very well, while the later introduced interior trees performed less well (Kleinschmit et al, 1979, Kohnle et al, 2012). Ten field trails with trees of several provenances were established around 1960 in south-west Germany in the course of an international provenance trial (Kenk and Thren,

1984). Phenotypic data and genomic material of trees from three of the ten field trials (Wiesloch/Philippsburg, Schluchsee, Sindelfingen) were analyzed in the "DougAdapt" project, of which this thesis was part of (see Chapter 1.1.5).

1.1.4. Economic value

Douglas-fir is of large economic value for timber production in the Pacific Northwest, where it is one of the most important timber species, due to its rapid growth and favorable wood quality (Lowell et al, 2014). The area of Douglas-fir production in Europe is also increasing (Reyer et al, 2010), with coastal provenances being planted more frequently, due to superior growth performances compared to interior provenances (Ducić et al, 2008, Hermann and Lavender, 1999, Kleinschmit et al, 1979). In Germany an area of approximately 217,000 ha is currently covered with Douglas-fir (https://www.bundeswaldinventur.de/). Wood quality is not the only reason for the interest in the species in Europe. Resistance against many European pathogens (Ducić et al, 2008, Hermann and Lavender, 1999) and the expected better adaptation to future climatic conditions compared to for example Norway spruce (Hanewinkel et al, 2013) are also important. Therefore, the identification and characterization of differentially adapted provenances is important for forest managers to be able to select trees that are capable to cope with the anticipated future climate. For this purpose, the "DougAdapt" research project was initialized.

1.1.5. DougAdapt project

Because of the large natural range with various climatic conditions it is expectable that different Douglas-fir provenances are differentially adapted to drought. In general, it is considered that interior Douglas-fir provenances are more drought tolerant than coastal provenances (Pharis and Ferrell, 1966), but coastal provenances showed better growth performances in Central Europe (see Chapter 1.1.3; Kleinschmit et al, 1979, Kohnle et al, 2012). Due to the long generation times in trees and the expected climatic changes in Central Europe, it is important for forest managers to understand how the different provenances respond to drought stress and which provenances are suited for plantation. Therefore, the project "DougAdapt: Adaptation of forest trees to climatic change – Diversity of drought responses in Douglas-fir provenances" aimed to link genotypic with phenotypic variation and to identify molecular mechanisms involved in the response to drought in several Douglas-fir provenances. In this project, trees subjected to

drought stress experiments under controlled conditions in green houses as well as trees planted approximately 50 years ago in the course of the international provenance trial in Germany were analyzed (Kenk and Thren, 1984). Several groups collaborated for this project and published several research papers focusing on different topics, like nitrogen content, gene expression profiling, or tree ring isotopic composition (Du et al, 2014, Hess et al, 2013, Jansen et al, 2013).

The main goals of this thesis within the "DougAdapt" project were to establish reference sequences for Douglas-fir, to detect drought-stress candidate genes, to identify differentially adapted genes, and to analyze the allelic variation within the provenances. For these tasks we conducted several experiments and applied state-of-the-art methods like next-generation sequencing (NGS).

1.2. Next-generation sequencing

The genetic information about development and functioning of organisms is encoded in their genome, which consists of DNA (deoxyribonucleic acid). Knowledge of the DNA sequence enables scientists to perform a multitude of different tasks, like searching for genes, comparing DNA of various species, or searching for differences in the genomes of several individuals of the same species (e.g., Altschul et al, 1990, Burge and Karlin, 1997). Changes in the genomic sequence, for instance deletions, insertions, or point mutations, can have a huge impact on the fitness and phenotype of an organism (Eyre-Walker and Keightley, 2007). Point mutations, also called single nucleotide polymorphisms (SNPs), play an important role in research today. They can be used for many tasks like revealing footprints of selection of a species (e.g., Cao et al, 2011, Oleksyk et al, 2010) or producing genotyping arrays that test if individuals have specific DNA polymorphisms (e.g., Ganal et al, 2011, Yang et al, 2009). SNPs in a gene are called synonymous SNPs if they do not change the amino acid sequence, and non-synonymous SNPs if they alter the amino acid sequence. Polymorphisms in non-coding regions are of interest, because they can have an influence on how genes are transcribed. SNPs in coding and non-coding regions without an obvious impact can still be used as markers for many tasks as they may be linked to an effect which they do not cause (Williams and Oleksiak, 2011).

The process of deciphering DNA fragments is called sequencing. In 1977 Frederick Sanger and his colleagues developed a method to sequence DNA using chain-terminating di-deoxy-

nucleotidetriphosphates (ddNTPs) during DNA replication (Sanger et al, 1977). Improvements, like the use of fluorescently tagged ddNTP, were the precursor for high-throughput DNA sequencing (Smith et al, 1986). Sanger sequencing is especially suited if rather long and/or high quality DNA reads (i.e. fragments) are required, but even with the improvements it cannot compete with next-generation sequencing methods in terms of throughput and costs (Table 1.2). With second-generation (methods requiring amplification of DNA prior to sequencing) and third-generation (methods not requiring DNA amplification) sequencing technologies, which came up in the early 2000s, it is possible to sequence millions to billions of base pairs (bp) in hours to days (Table 1.2, Glenn, 2011). A short overview of the sequencing technologies used in this thesis will be presented in this chapter.

Table 1.2.: Comparison of sequencing methods (values from http://www.molecularecologist.com/next-gen-fieldguide-2013/, see also Glenn (2011)). Applied Biosystems 3730xl represents Sanger sequencing, while 454 FLX Titanium and Illumina HiSeq 1000 use next-generation sequencing. The latter two were used in this thesis. PCR - polymerase chain reaction, Mbp - megabase pair, PE - paired end.

Instrument	Applied Biosystems 3730xl (capillary)	454 FLX Titanium	Illumina HiSeq 1000
Amplification	PCR, cloning	Emulsion PCR	Bridge PCR
Run time (max. read length)	2 hrs.	10 hrs	8.5 days
Millions of reads / run	9.6×10^{-5}	1	≤ 1500
Bases / read	650	400	100+100 (PE)
Yield Mbp / run	0.06	400	$\leq 300,000$
Reagent cost / run	\$144	\$6,200	\$10,220
Reagent cost / Mbp	\$2,308	\$12	\$0.04

1.2.1. 454 technology

454 pyrosequencing was the first commercially available second-generation sequencing technique, and was used in Chapter 2 (Margulies et al, 2005). The technique generally produces reads with a length of 400 to 650 bp, which are advantageous in *de novo* applications if no reference genome is available or if gaps in a reference should be closed. To perform 454 sequencing, specific primers are ligated to fragmented and denaturated DNA strands (Figure 1.2a). Each fragment is then bound to a bead, and an emulsion PCR is performed with the result that each bead carries millions of clonal copies of the same DNA fragment (Dressman et al, 2003). DNA-

bound beads are then loaded on a picotiter plate, where one bead is placed into each well. After putting the plate in the sequencing device, the four possible nucleotides are added successively. If nucleotides complementary to the DNA strand in a bead are incorporated, a pyrophosphate is released. The release causes a light signal (with the help of ATP sulfurylase and luciferase) that is recorded by a camera (Mardis, 2008, Shendure and Ji, 2008). Because the wells of a plate are recorded in parallel, this method yields high-throughput data sets. Compared to other second-generation sequencing methods, this method produces rather long reads, but the costs in time and price per bp are rather high (Table 1.2, Glenn, 2011, Mardis, 2008).

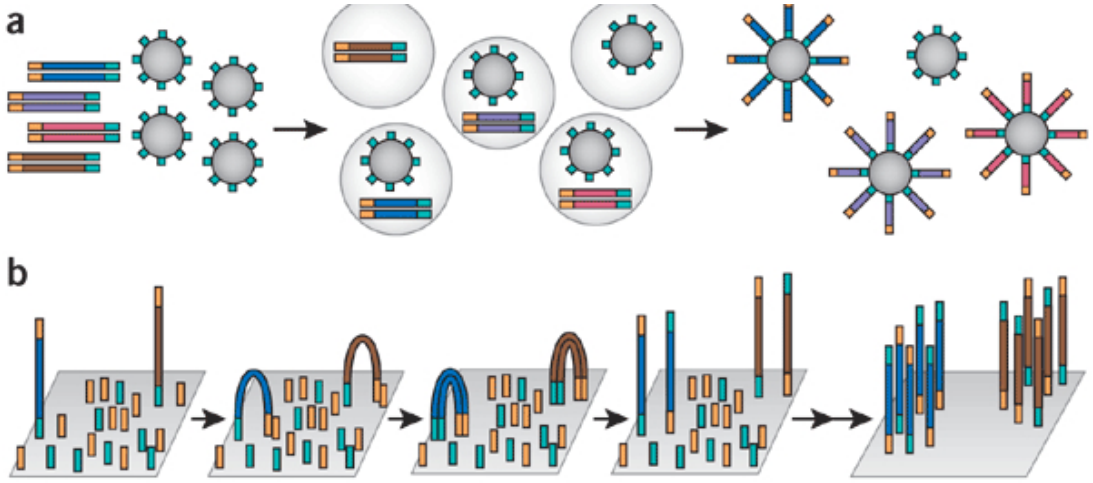


Figure 1.2.: a) Emulsion PCR used in 454 sequencing (Dressman et al, 2003). Adapter (gold and turquoise) flanked DNA fragments are attached to the surface of small beads, which carry an oligonucleotide complementary to one of the adapters in a water-oil emulsion. Due to low template concentration most compartments with beads contain either one or zero DNA molecule. At the end of emulsion PCR, the beads carry a large number of copies of the same DNA fragment. b) Bridge PCR used in Illumina technology. The surface of the flowcell is densely covered with both primers of adapter flanked DNA molecules. The molecules bind to the surface and form a bridge to a nearby primer. After second strand synthesis, the DNA is denaturated, and the process starts again. In the end the flowcell contains a large number of clusters, each of them consisting of $\sim$1,000 single strand copies of the same DNA fragment. Figure from Shendure and Ji (2008).

1.2.2. Illumina technology

Another second-generation sequencing method is Solexa or Illumina sequencing, which was applied in Chapter 3 and Chapter 4. It is a sequencing-by-synthesis method which usually produces reads of 100 bp to 150 bp, but can also be used to generate reads with lengths up to 600 bp (Glenn, 2011). *In vitro* generated adapter flanked DNA fragments are amplified on a solid

surface (the flowcell) via bridge PCR (Figure 1.2b, Adessi et al, 2000, Fedurco et al, 2006). The solid surface is coated with oligonucleotides complementary to both adapters ligated to the DNA fragments. The adapter of a fragment binds to the complementary oligo, and then builds a bridge to the second oligo on the surface. Primers and enzymes are added to synthesize the second strand, which results in a double stranded DNA bridge that is denaturated again, leaving two copies of the same fragment. Bridge amplification is repeated several times, resulting in a flowcell covered with several millions of dense clusters, whereby each cluster contains ∼1,000 single strand copies of the same fragment (Shendure and Ji, 2008). During sequencing fluorescently labeled dNTPs are used, where the label also serves as a terminator for polymerization, ensuring that only one nucleotide is added in each sequencing cycle. After each cycle the incorporated nucleotide is imaged and the dye is cleaved enzymatically allowing the addition of another labeled dNTP in the next sequencing cycle. The average raw error rate is higher with Illumina sequencing compared to other sequencing platforms, but because of generally high coverages (each nucleotide of the DNA is sequenced in several reads) sequencing errors can be identified. In comparison to other second-generation sequencing methods, the costs per gigabase are low with Illumina sequencing (Table 1.2, Glenn, 2011).

1.3. Sequencing approaches

While for other plant taxa such as *Arabidopsis thaliana* reference genomes are available (The Arabidopsis Genome Initiative, 2000), genome analysis of conifer species lags behind because of their large genome sizes. Douglas-fir genome size is ∼ 19 Gbp (Ahuja and Neale, 2005), which is about 100 times the genome size of *A. thaliana* (Weigel and Mott, 2009) or 37 times the genome size of poplar (Bradshaw and Stettler, 1993, Tuskan et al, 2006). In this chapter, the sequencing approaches performed during this study to obtain Douglas-fir sequence data are briefly summarized.

1.3.1. RNASeq

At the start of this thesis in 2010, whole genome sequencing of a species with a large genome like Douglas-fir was very expensive. Until now only few whole genome assemblies are available, for instance for Norway spruce (Nystedt et al, 2013) and for loblolly pine (Neale et al, 2014, Zimin

et al, 2014). A Douglas-fir whole genome sequencing project is underway, but not finished yet (Neale et al, 2013). A cost-efficient alternative to sequence a whole genome is RNASeq, in which only the transcriptome, i.e., the regions coding for proteins, are sequenced (Wang et al, 2009). The coding regions of the DNA are transcribed to mRNAs (messenger ribonucleic acid), which are then translated into protein. In RNASeq, reverse transcriptase is used to synthesize cDNA (complementary DNA) from mRNA. The resulting cDNA is sequenced and assembled into unigenes or putative unique transcripts (PUTs). Due to long reads 454 sequencing technology is suited very well for RNASeq. Advantages of this approach are for example the suitability for non-model species and the possibility to find alternatively spliced transcripts (Wang et al, 2009). A drawback of this method is that it is not possible to identify all proteins, but only those which are expressed at the time of measurement in the cells or tissues under consideration.

1.3.2. Sequence capture

Sequence capture is a method which can be used to re-sequence complex genomes and to detect SNPs in individuals in a cost-efficient way (Grover et al, 2012). Target regions, for example coding regions, are defined based on available sequence information, and oligonucleotides complementary to those regions are synthesized. Fragmented DNA libraries of individuals are generated and mixed with the oligonucleotides. After ligation only DNA-oligo-complexes are captured that represent the predefined target regions. Sequencing costs are reduced because only the regions of interest are captured and sequenced. Two major methods are available, using either DNA or RNA oligonucleotides as probes. In this thesis, we used DNA-based probes, which were already applied in several plant species (Haun et al, 2011, Henry et al, 2014, Mascher et al, 2013). Illumina sequencing is well suited for this task, since the reference sequence is known, and defined oligonucleotides are overlapping. It is important to obtain a sufficient coverage per nucleotide to detect SNPs reliably (at least ten reads should cover a nucleotide), which is given using Illumina technology.

1.3.3. Genotyping-by-sequencing

Genotyping-by-sequencing (GBS) is a cheap, technically simple, and highly multiplex-able method suitable for a large variety of applications such as the identification of a large number of SNPs and population studies (Elshire et al, 2011). The key idea is to use one or two restriction

enzyme(s) to digest and fragment a genome (Poland et al, 2012b). A digestion with one restriction enzyme is called a single-digest (SD) and a digestion with two enzymes a double-digest (DD). Only fragments with a specific size are sequenced using Illumina technology resulting in a reduced representation of the genome after sequencing. Due to the reduction of the genome it is possible to sequence more individuals on the same lane if different barcodes are added to the fragments of each individual. Since restriction enzymes always cut at the same position, the digestion should result in the same fragments in different individuals of the same species, as long as there is no polymorphism in the restriction site. The suitability of this method with conifer species was shown for lodgepole pine (*Pinus contorta*) and for white spruce (*Picea glauca*, Chen et al, 2013). If reference sequence information is available the reads can be mapped against the reference and SNPs can be identified. However, if no reference information is available, a *de novo* analysis needs to be performed. The *de novo* analysis is more complex because reads from the same restriction sites have to be identified first, and then have to be aligned to each other (Catchen et al, 2011). Nevertheless, it is possible to detect polymorphisms, if no reference information is available.

1.4. Objectives

The goal of this thesis was to analyze the genetic variation of Douglas-fir provenances by generating an extensive resource of sequence and polymorphism information, which can be used in population genomics and genome-wide association studies.

Due to the lack of a reference genome or transcriptome, the first study aimed at establishing a set of PUTs representing the transcriptome and at searching for genes which are part of the trees drought response. Since cDNA of trees subjected to drought stress experiments was used (i.e. from control and from drought stressed trees), drought related candidate genes were identified.

Using the reference transcriptome of the first study, a targeted sequence capture experiment was performed. The main objectives of the second experiment were to identify SNPs in several individuals of five provenances, and to analyze the genetic differentiation within and between the provenances with a special focus on the drought candidate genes. Furthermore, the data was screened for differentially adapted genes by searching for patterns of directional selection.

Finally, in a third experiment the sequence capture and two GBS approaches were compared to assess their abilities to genotype a large number of individuals at low cost and to infer population structure.

2. A catalogue of putative unique transcripts from Douglas-fir (*Pseudotsuga menziesii*) based on 454 transcriptome sequencing of genetically diverse, drought stressed seedlings

Thomas Müller[1], Ingo Ensminger[2,3] and Karl J. Schmid[1]

[1]Department of Crop Biodiversity And Breeding Informatics, University of Hohenheim, Stuttgart, Germany, [2]Department of Biology, University of Toronto at Mississauga, Mississauga, ON, Canada, [3]Forest Research Institute of Baden-Württemberg (FVA), Freiburg i. Brsg., Germany

BMC Genomics 2012, 13:673 **doi:10.1186/1471-2164-13-673**

2.1. Abstract

Background Douglas-fir (*Pseudotsuga menziesii*) extends over a wide range of contrasting environmental conditions, reflecting substantial local adaptation. For this reason, it is an interesting model species to study plant adaptation and the effects of global climate change such as increased temperatures and significant periods of drought on individual trees and the forest landscape in general. However, genomic data and tools for studying genetic variation in natural populations to understand the genetic and physiological mechanisms of adaptation are currently missing for Douglas-fir. This study represents a first step towards characterizing the Douglas-fir transcriptome based on 454 sequencing of twelve cDNA libraries. The libraries were constructed from needle and wood tissue of coastal and interior provenances subjected to drought stress experiments.

Results The 454 sequencing of twelve normalized cDNA libraries resulted in 3.6 million reads from which a set of 170,859 putative unique transcripts (PUTs) was assembled. Func-

tional annotation by BLAST searches and Gene Ontology mapping showed that the composition of functional classes is very similar to other plant transcriptomes and demonstrated that a large fraction of the Douglas-fir transcriptome is tagged by the PUTs. Based on evolutionary conservation, we identified about 1,000 candidate genes related to drought stress. A total number of 187,653 single nucleotide polymorphisms (SNPs) were detected by three SNP detection tools. However, only 27,688 SNPs were identified by all three methods, indicating that SNP detection depends on the particular method used. The two alleles of about 60 % of the 27,688 SNPs are segregating simultaneously in both coastal and interior provenances, which indicates a high proportion of ancestral shared polymorphisms or a high level of gene flow between these two ecologically and phenotypically different varieties.

Conclusions We established a catalogue of PUTs and large SNP database for Douglas-fir. Both will serve as a useful resource for the further characterization of the genome and transcriptome of Douglas-fir and for the analysis of genetic variation using genotyping or re-sequencing methods.

2.2. Background

Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) is an ecologically highly variable species that occurs in two main varieties in North America. The natural range of the coastal or green Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*) extends over 2,200 km from the Pacific Northwestern fog belt and the adjacent summer-dry Coastal Range and Cascade mountains to the drier coastland of Southern California. The interior or blue Douglas-fir (*Pseudotsuga menziesii* var. *glauca*) is distributed over more than 4,500 km along the dry continental climates of the montane to the subalpine Rocky Mountains from Alberta to Colorado with isolated populations reaching into Mexico. Douglas-fir grows from sea level on Vancouver Island up to 3,000 m altitude in the southern Rocky Mountains (Hermann and Lavender, 1990). Within its natural range, Douglas-fir has evolved into a variety of genetically diverse populations adapted to contrasting ecozones (e.g., Campbell, 1979, Dean, 2007).

Douglas-fir populations differ in their response to frost (Darychuk et al, 2012, Rehfeldt, 1979), drought (Andrews et al, 2012, Darychuk et al, 2012, Martinez-Meier et al, 2009), and along environmental gradients (Jansen et al, 2013, Rehfeldt, 1979). Like most conifer species, it is

able to cope with limitations in soil water availability within its natural range (Carter and Klinka, 1990, Coops et al, 2007). There is a negative relationship between shoot water potential and the photosynthesis rate (Andrews et al, 2012), which decreased by about 70 % in water-stressed trees with a pre-dawn shoot water potential of about -1.7 MPa. In conifers such as Douglas-fir or pine, the recovery of photosynthesis upon rainfall and re-watering occurs within days together with the rapid recovery of predawn shoot water potential from stressed (around -1.5 MPa), or mildly stressed (around -1.0 MPa) to values higher than -0.5 MPa (Andrews et al, 2012, Watkinson et al, 2003). This high ecological, genetical and physiological diversity provides an excellent system to study the adaptation of conifer trees to contrasting environments.

Due to its rapid growth and favorable wood quality, Douglas-fir is an economically relevant species for timber production. In Europe, the area of Douglas-fir production is rising (Reyer et al, 2010). Forest practitioners appreciate the resistance of Douglas-fir against many European pathogens (Ducić et al, 2008, Hermann and Lavender, 1999). It is also expected that Douglas-fir is better adapted to future climate conditions in Central Europe than e.g. Norway spruce (Hanewinkel et al, 2013).

The Intergovernmental Panel on Climate Change (IPCC) expects increasing summer temperatures and decreasing precipitation in Central Europe in the coming years (IPCC, 2007). A trend towards warmer summers and more frequent summer droughts was reported in recent studies and simulations (Brohan et al, 2006, Fink et al, 2004, Meehl and Tebaldi, 2004). For this reason, it is important for forest managers to select suitable tree species or provenances that are adapted to the anticipated future climate. Currently, coastal Douglas-fir provenances are more frequently planted in Central Europe due to their superior growth performance compared to interior Douglas-firs (Ducić et al, 2008, Hermann and Lavender, 1999, Kleinschmit et al, 1979). The identification and characterization of differentially adapted provenances of coastal and interior Douglas-fir varieties has therefore high practical value.

Because of the large genome size of Douglas-firs (18.7 Gbp, about 100 times the genome size of *Arabidopsis thaliana* (Ahuja and Neale, 2005, Weigel and Mott, 2009) or 37 times the genome size of poplar (Bradshaw and Stettler, 1993, Tuskan et al, 2006)), transcriptome analysis is a cost-effective and suitable approach for the identification of candidate genes for adaptive traits and molecular markers that are linked to phenotypic variation. Transcriptomes of many species have been analyzed by next-generation sequencing technologies (Novaes et al, 2008,

Parchman et al, 2010, Pauchet et al, 2009), and numerous coding single-nucleotide polymorphisms (SNPs) were identified in conifer species such as *Pinus contorta*, *Picea glauca* and *Pinus taeda* (González-Martínez et al, 2006, Parchman et al, 2010, Pavy et al, 2006).

Douglas-firs, like other forest trees, have a high level of genetic diversity (Hamrick et al, 1992, Viard et al, 2001). For example, one study identified 933 SNPs in 121 candidate genes for cold-hardiness (1 SNP per 43 bp to 1 SNP per 112 bp) in coastal Douglas-firs (Eckert et al, 2009d). For this reason, transcriptome sequencing of different provenances will lead to candidate genes for differential adaptation and to many new genetic markers for the characterization of different populations.

The purpose of this study was to establish a catalogue of Douglas-fir putative unique transcripts (PUTs) enriched for drought stressed genes and to identify genetic polymorphisms as resource for further analysis such as re-sequencing projects, association studies, and gene expression profiling.

2.3. Results

Sequencing and assembly

The sequencing of twelve cDNA libraries resulted in 3,619,544 reads with an average length of 338 bp. After preprocessing, the number of reads decreased to 2,957,373. Read numbers were not equally distributed among libraries (Supplementary Figure A.1). The DINM, DINS and DIWC libraries consisted of less than 200,000 reads each and the DIWM library of less than 100,000 reads (see Table 2.1 for an explanation of the library abbreviations). The average length of the reads decreased to 315 bp after pre-processing (Supplementary Figure A.1). More than 99 % of reads in each library were used for the construction of the assembly after quality trimming, with the exception of the DIWM library (95 % used). A total of 2,793,051 (94.44 %) reads were assembled into 141,626 isotigs (of which 275 were contigs) of at least 100 bp length. Supplementary Figure A.2 contains the origin and the number of assembled reads. All isotigs were clustered in 116,311 isogroups. The mean isotig length was 623.22 bp (s.d. 437.67 bp, median: 474 bp), the mean coverage per base was 5.0 reads (s.d. 8.07), and the mean number of reads per isotig was 44.5 (s.d. 145.54). For 21,837 isotigs longer than 999 bp, the mean coverage increased to 13.66 (s.d. 11.77) reads per base. Furthermore, the mean number of reads

per isotig reached 181.27 (s.d. 274.75). Length of the isotig was positively correlated with the number of reads ($r = 0.4972$, $P < 0.0001$; Supplementary Figure A.3).

Table 2.1.: Explanation of the cDNA library abbreviations. D = Douglas-fir, C/I = coastal/interior, N/W = needle/wood tissue, C/M/S = no/mild/severe drought stress.

		Treatment		
Variety	**Tissue**	**Control**	**Mild stress**	**Severe stress**
Coastal	Needles	DCNC	DCNM	DCNS
	Wood	DCWC	DCWM	DCWS
Interior	Needles	DINC	DINM	DINS
	Wood	DIWC	DIWM	DIWS

Based on the results of the assembly, we constructed a set of PUTs as outlined in the Methods section. 42,159 of 71,392 reads with a length >99 bp initially labeled as singletons were mapped to isotigs and were considered as false positive singletons. Therefore, the final PUT set consisted of 170,859 sequences (141,626 isotigs and 29,233 singletons) with an average sequence length of 564.6 bp (s.d. 420.86 bp, median: 431 bp, Supplementary Figure A.4). As no reference sequence of Douglas-fir was available, we used the PUT set as reference for the following analysis, including functional annotation and SNP detection.

Functional annotation of the PUTs

For functional annotation, we compared all PUTs to the NCBI *nr* database using BLASTX with an e-value cutoff of 10^{-10}. At least one BLAST hit was obtained with 46,645 transcripts. If only the best hit of each transcript is considered, a total of 20,604 different sequences (unique hits) were hit in the *nr* database. The largest number of hits was against *Picea sitchensis* sequences, followed by *Vitis vinifera* (Figure 2.1). In the subsequent analysis, Blast2GO assigned at least one GO term to 39,624 transcripts. For the three main GO categories, 34,660 transcripts were assigned a GO term from the molecular function category, 28,714 from the biological process, and 24,166 from the cellular component category. To compare the distribution of GO terms of the Douglas-fir transcriptome with the distribution of GO terms of transcriptomes from other species, we also applied Blast2GO to the *Arabidopsis thaliana* and *Picea sitchensis* sequences downloaded from TAIR and NCBI, respectively. We chose these two species for comparison because *A. thaliana* is a well studied model species with a well studied transcriptome and *P.*

sitchensis is the species with most top BLASTX hits in our Douglas-fir PUT set. Figure 2.2 and Supplementary Figure A.5 show that the distributions of GO terms at GO level 2 to 5 for each of the three ontology classes are highly similar for all three species.

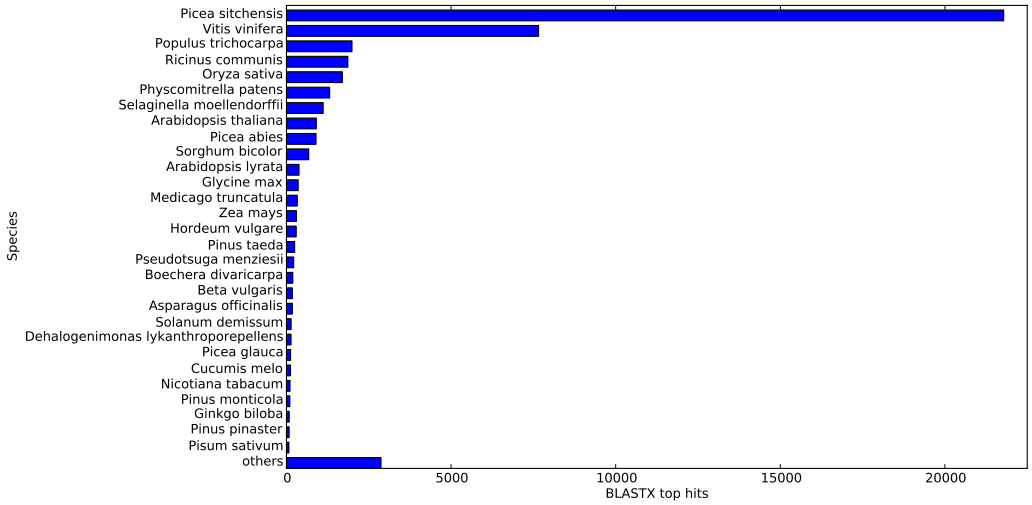


Figure 2.1.: Distribution of the top hits species of the BLASTX search of the PUTs from the assembly against NCBI's *nr* database.

Identification of treatment-specific PUTs

The isotigs (non-singleton transcripts of the PUT set) were divided and clustered according to the origin of their reads (Figure 2.3). About one third (34.38 %) of the isotigs contained reads from all three treatments and were therefore grouped in the *cms* group. The average length of isotigs in the *cms* group was 995.32 bp (Table 2.2). Each of the *cm*, *cs*, and *ms* groups contained 6-8 % of the isotigs with an average isotig length of 489 to 496 bp. The *c*, *m*, and *s* groups contained 14-15 % of the isotigs in each case. The average lengths of those isotigs were between 393 and 405 bp. The search for specific keywords in the BLASTX results revealed that 1,503 different isotigs coming from 998 isogroups had a BLASTX hit containing one of the keywords related to stress (Table 2.3, Supplementary Table A.1, Supplementary file 7). 134 of those isotigs coming from 132 isogroups were part of the *m*, *s*, or *ms* groups and will serve as top candidate genes in future studies. We expected that *cms* group sequences are more conserved than sequences assigned to the remaining groups because drought stress specific sequences may evolve faster or

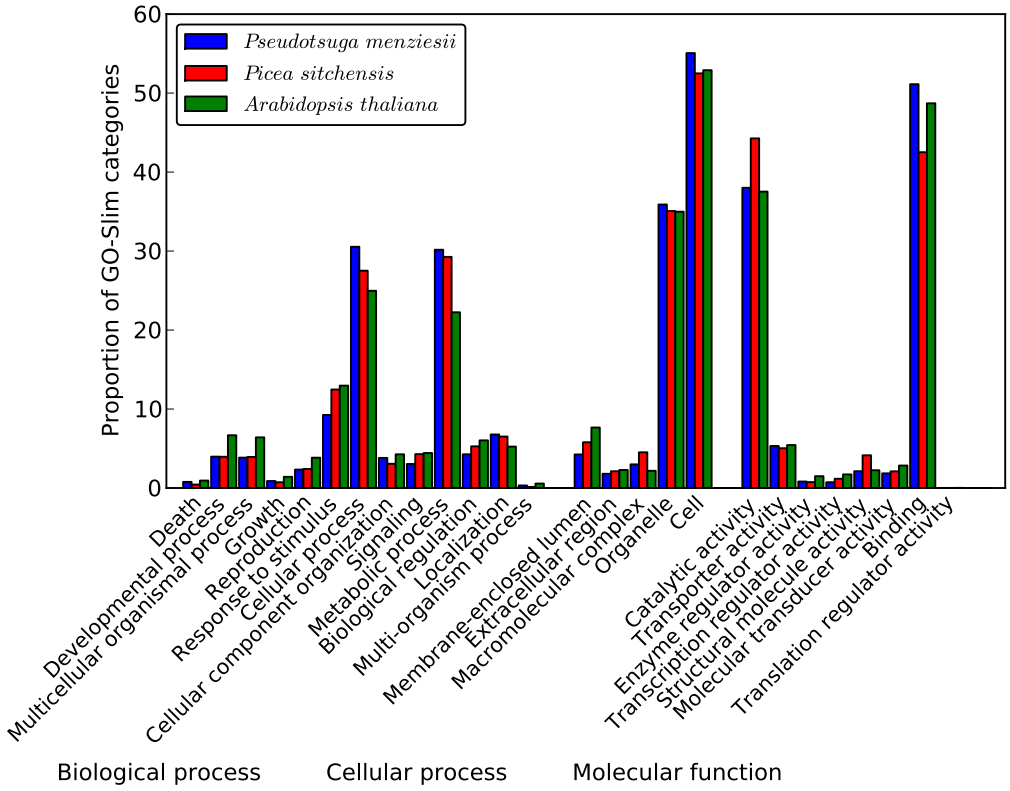


Figure 2.2.: Comparison of the distribution of the GO-Slim categories of the Douglas-fir PUT set versus *Picea sitchensis* and *Arabidopsis thaliana* at GO level 2. Transcriptome data of *P. sitchensis* and *A. thaliana* were obtained from NCBI and TAIR databases, respectively.

are of a more recent evolutionary origin than common or widely expressed genes.

To test this hypothesis, we determined the proportion of significant BLASTX hits within each group of isotigs against the *nr*, the *ara*, and the *picea* databases (Table 2.2). Most hits were observed in the *cms* group (e.g., 58.11 % against *nr*) and the least number of hits in the *m* group (14.13 % against *nr*). However, there is a highly significant correlation between the average length of isotigs and percent BLAST hits (e.g., hits against *ara*, $P < 0.0001$, Table 2.4), and also between the total sequence length of each isotig group with the proportion of BLAST hits (e.g., hits against *ara*, $P = 0.003$). Hence, the differences in the proportion of BLAST hits among classes of isotigs are not a result of differential evolutionary conservation, but of the amount of sequence data in each class.

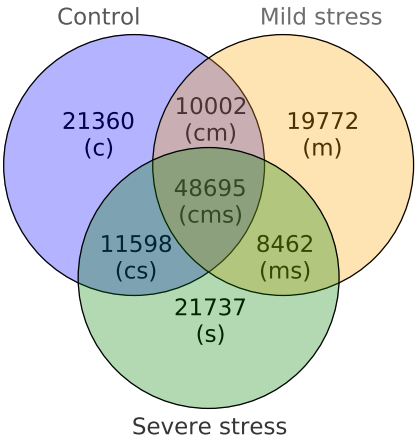


Figure 2.3: Venn diagram showing the number of non-singleton PUTs (i.e., isotigs) consisting of reads from (one or several) libraries of one or more treatment(s). E.g., 21,737 isotigs are composed of reads originating from one or several of the severe stress libraries (DCNS, DINS, DCWS, DIWS). *c* = control, *m* = mild stress, *s* = severe stress.

Table 2.2.: Percentages of isotigs with BLASTX hits. Percentages of isotigs (non-singleton PUTs) with a BLASTX hit against the *nr*, *ara*, and *picea* databases within the groups (see Figure 2.3). *c* = control, *m* = mild stress, *s* = severe stress, *cm* = control and mild stress, *cs* = control and severe stress, *ms* = mild and severe stress, *cms* = control, mild and severe stress.

			% of isotigs with BLASTX hits vs.		
	Avg. length of isotigs	**% of all isotigs**	***nr***	***ara***	***picea***
c	393.33	15.08	16.29	9.11	10.59
m	392.77	13.96	14.13	6.22	7.62
s	405.5	15.35	14.77	7.32	8.75
cm	488.89	7.06	24.59	15.8	16.12
cs	493.72	8.19	25.38	17.44	17.19
ms	496.32	5.97	19.14	11.05	11.18
cms	995.32	34.38	58.11	47.24	46.06

SNP identification

SNP detection was performed with three different programs, GSMapper, ssahaSNP, and bwa/SAMtools, to minimize the number of false positives. PUTs obtained from the assembly served as reference. The programs detected 57,691 (Newbler), 155,269 (ssahaSNP), and 85,346 (bwa/SAMtools) SNPs, resulting in a total number of 187,653 different SNPs. However, only 27,688 SNPs were detected by all three tools (Figure 2.4). These SNPs were selected for further analysis because we consider them as most reliable true positive polymorphisms. These SNPs were distributed over 10,517 different PUTs of 10,054 different isogroups. Most transcripts harbored only a single SNP and 2,499 transcripts contained more than three SNPs. A total of 23 SNPs were detected in the most polymorphic PUT. In the 7,684 transcripts with at least one SNP and a

Table 2.3.: Keyword search in BLASTX results. Number of isotigs (non-singleton PUTs) with a BLASTX hit containing a keyword for each group (see Figure 2.3). *c* = control, *m* = mild stress, *s* = severe stress, *cm* = control and mild stress, *cs* = control and severe stress, *ms* = mild and severe stress, *cms* = control, mild and severe stress.

	Isotig group						
Search term	***c***	***m***	***s***	***cm***	***cs***	***ms***	***cms***
"Drought"	4	3	8	5	13	3	103
"Water-deficit"	0	0	4	0	5	0	29
"Water-stress"	6	2	6	4	6	1	109
"Osmotic-stress"	4	1	6	2	7	2	58
"Heat-stress"	2	0	1	0	2	0	27
"Heat-shock"	24	17	31	21	23	15	466
"Dehydration"	20	7	17	14	18	1	205
"Abscisic acid"	7	1	8	5	10	2	142
"ABA-responsive"[1]	0	0	2	1	1	1	25
"ABA-induced"	1	0	2	2	1	0	27
"ABA receptor"	0	0	2	0	0	0	20
"Pyrabactin resistance 1"	0	0	2	0	0	0	10
"Snf1-related protein kinases"[2]	4	2	2	2	3	7	69
"DREB1"[3]	2	0	0	1	1	0	9
"DREB2"	2	0	0	2	2	0	14
"C-repeat binding"	0	0	0	1	1	0	4
"ERD"[4]	7	4	9	3	8	2	112
"CIPK"[5]	2	3	2	0	2	6	47
"CDPK"[6]	0	2	1	7	1	0	39
"CBL1" [7]	5	1	0	3	6	1	72
"PKS3"[8]	0	2	0	0	2	0	12
Different isotigs	66	33	69	58	71	32	1,174

[1]ABA = abscisic acid
[2]Snf = sucrose non-fermenting
[3]DREB = dehydration-responsive element-binding
[4]ERD = early responsive to dehydration
[5]CIPK = CBL-interacting protein kinase
[6]CDPK = calcium-dependent protein kinase
[7]CBL = calcineurin B-like protein
[8]PKS = phytochrome kinase substrate

significant match against the *nr* database, 5,378 SNPs were classified as synonymous and 4,129 as non-synonymous.

In addition, we estimated the polymorphism level of the transcriptome by dividing the number of SNPs with the total number of different nucleotides in PUTs (as the same contig can contribute

Table 2.4.: BLASTX and Blast2GO results divided by isotig length. Results of similarity searches with BLASTX and functional annotation using Blast2GO subdivided by transcript length in bp.

	All PUTs	< 501	501 – 1,000	1,001 – 1,999	$> 2{,}000$
Number of PUTs	170,859	106,296	42,760	19,589	2,214
Total sequence [Mbp]	96.5	35	26.4	29.7	5.4
Hits with *nr*	27.3%	13.5%	36.9%	75.5%	87.8%
Hits with *ara*	19.6%	7.9%	25.1%	64.4%	82.6%
Hits with *picea*	19.6%	8.5%	26.3%	58.7%	71.4%
Isotigs with assigned GO term	23.2%	11.9%	26.6%	63.9%	82%

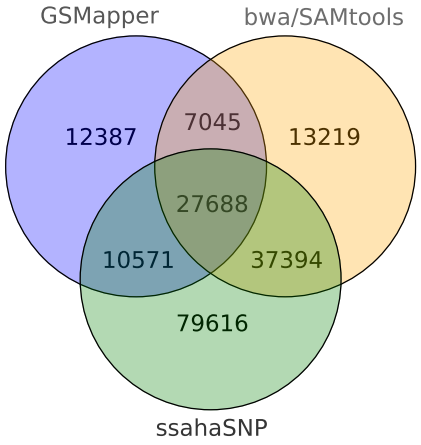


Figure 2.4: Number of SNPs identified by the SNP detection tools GSMapper, ssahaSNP, and bwa/SAMtools. 27,688 SNPs were detected by all three tools and are considered to be the most reliable SNPs.

to several isotigs, see Supplementary Figure A.6). If only the most reliable SNPs are considered, the estimated nucleotide diversity (0.04 %, corresponding to approximately 1 SNP per 2,530 nucleotides) is very low. Using all SNPs identified by Newbler, bwa/SAMtools and ssahaSNP separately, resulted in estimated polymorphism levels of 0.08 % (1 SNP per 1213 bp), 0.12 % (1 SNP per 820 bp), and 0.22 % (1 SNP per 451 bp), respectively.

To investigate differences in the level of genetic diversity between coastal and interior Douglas-firs, we divided the SNPs into groups depending on whether their alleles segregated in coastal or interior provenances, or in both (Table 2.5). The majority of SNPs are polymorphic in both coastal and interior provenances (Table 2.6), but coastal provenances have a higher number of provenance-specific alleles, as seen in the comparison of ci/c (7,158 SNPs) versus ci/i (2,547 SNPs) groups.

Table 2.5.: Groups of SNPs. Partitioning of SNPs into groups depending on the origin (coastal vs. interior) of sequence reads. c: only reads of coastal libraries; i: only reads of interior libraries, ci: reads of coastal as well as interior libraries; ci/ci: both possible nucleotides were confirmed by reads of coastal and interior libraries; c/i: one of the possible nucleotides at the SNP position was confirmed only by reads of coastal libraries, the other nucleotide only by reads of interior libraries; etc.

Origin of reads confirming the reference nucleotide	c	i	c	i	c	ci	i	ci	ci
Origin of reads confirming the variant nucleotide	c	i	i	c	ci	c	ci	i	ci
Group name	c/c	i/i	c/i		ci/c		ci/i		ci/ci

Table 2.6.: Summarized number of SNPs. Number of SNPs with a specific composition of reads. ci/ci: variant and reference nucleotide appeared in reads from coastal and interior libraries; ci/c, ci/i: variant or reference nucleotide appeared only in reads of the coastal or interior libraries, the other one appeared in reads of both kind of libraries; c/i: variant or reference nucleotide appeared only in reads of the coastal libraries, the other one only in reads of the interior libraries; c/c, i/i: variant and reference nucleotides appeared only in reads of coastal or interior libraries.

Origin of reads at SNP position	Number of SNPs
ci/ci	15,843
ci/c	7,158
ci/i	2,547
c/i	886
c/c	817
i/i	437

2.4. Discussion

Sequencing and assembly

Next-generation sequencing (NGS) has now a major impact on the genome-wide analysis of transcriptomes in non-model species (Novaes et al, 2008, Parchman et al, 2010, Pauchet et al, 2009). To achieve a comprehensive characterization of the protein-coding genome of Douglas-fir, we exposed young seedlings from different provenances to drought stress treatments and generated normalized cDNA libraries to enrich for rare transcripts or genes not constitutively expressed. All libraries were assembled into a single assembly to maximize the number of reads per transcript and to improve the quality of assembly and SNP annotation. There is a strong relationship between the number of reads and the length of a transcript, confirming the observation that longer transcripts consist of more reads than shorter transcripts (Meyer et al,

2009). The number and average length of the reads of four libraries (DINM, DINS, DIWC, and DIWM) were below expectation (Kumar and Blaxter, 2010, Parchman et al, 2010, Schuster et al, 2010), probably because of problems during the sequencing process. However, we did not exclude these libraries, because they contributed the same proportion of reads to the assembly as the other libraries ($>$ 95 % of the reads of each library).

94.44 % of all reads were assembled into isotigs during assembly, which is a high proportion compared to similar 454 transcriptome assemblies. For example, 88 % of reads were assembled in *Melitaea cinxia* (Vera et al, 2008) and *Eucalyptus grandis* (Novaes et al, 2008), 78 % in *Pandinus imperator* (Roeding et al, 2009), and 48 % in *Pinus contorta* (Parchman et al, 2010). One cause for the high proportion in our data is the stringent preprocessing of reads, which excluded most uninformative reads prior to the assembly. The number of PUTs in the assembly (170,859) exceeds the number of expected genes in conifer genomes, which ranges from 30.000 to 50.000 genes (Rigault et al, 2011). Nevertheless, the PUT set is smaller than the one obtained with *Pinus contorta* (303,450 transcripts; Parchman et al, 2010), but larger than in other 454 transcriptome sequencing projects (e.g., 44,469 transcripts in waterhemp (Riggins et al, 2010), 108,297 transcripts in a butterfly species (Vera et al, 2008)). It should be noted that it is difficult to compare numbers of transcripts in different projects, because they are influenced by the genome and transcriptome sizes of the sequenced organism, the assembly method used (Kumar and Blaxter, 2010), and the number of reads used for the assembly. Overall, the high number of transcripts compared to the expected number of genes is likely a result of incompletely assembled genes.

The average length of transcripts is 564.60 bp (median 431 bp, Supplementary Figure A.4), which is approximately half of the expected average gene length in eukaryotes (1346 bp; Xu et al, 2006).

Functional annotation of PUTs

We annotated the PUTs by using stringent BLASTX searches against the *nr* database from NCBI with a cutoff e-value of 10^{-10}. Assuming that each unique blast hit represents a different transcribed gene, we tagged 20,604 genes of the Douglas-fir genome. This number is similar to other projects in *Pinus contorta* with 17,321 tagged genes (Parchman et al, 2010) and is lower than the estimated total number of about 30,000 transcribed genes in white spruce *Picea glauca* (Rigault

et al, 2011). If we further assume that the number of unique blast hits equals the number of transcribed genes and that the transcriptome sizes of Douglas-fir and white spruce are comparable, it seems that the PUT set generated in this study tags about two thirds of the protein-coding genes of Douglas-fir. The missing third likely comprise (1) cDNAs that were excluded from assembly because of low quality; (2) genes that are expressed at different developmental stages, growth conditions, or tissues and were thus not represented in the twelve cDNA libraries despite the normalization process; and (3) non-conserved genes, which are either lineage-specific and not yet present in the *nr* database or rapidly evolving genes with e-values $> 10^{-10}$ in the BLASTX comparisons. Since about 75 % of the transcripts do not result in significant BLASTX hits, one may speculate that a large proportion represents non-conserved genes.

Gene Ontologies (GOs) provide a standardized set of terms to describe genes and gene products consistently in different species and databases (Ashburner et al, 2000). GO terms are widely used for annotation and for comparisons of gene products of different species (e.g., Parchman et al, 2010, Salem et al, 2010). The similarity of the GO annotation distributions of Douglas-fir PUTs to the well-characterized transcriptome of *A. thaliana* and the protein sequences of *P. sitchensis* (Figure 2.2) suggests that the PUT set broadly represents the Douglas-fir transcriptome and can be viewed as being representative for further applications and investigations.

Identification of drought stress related genes

Dividing PUTs consisting of multiple reads (i.e., the isotigs) by the origin of their reads is a simple, but useful method to identify potential treatment-specific sequences. About 50 % of isotigs consist of reads from the single treatment groups *c*, *m*, or *s*. On average, they are shorter than isotigs containing reads from at least two different treatments. The isotigs of the *m* and *s* groups, but also the *ms* group, were considered to be top candidates for drought stress tolerance or resistance. However, since most treatment-specific isotigs consist of only two or three reads that originated from a total of twelve cDNA libraries, we reasoned that the power of a statistical model to detect significant differences is low. Therefore, we compared the extent of evolutionary conservation between groups as judged by the proportion of significant BLAST hits. This analysis is based on the notion that widely expressed genes are under stronger selective constraint than treatment-specific genes (Mukhopadhyay et al, 2008, Zhang and Li, 2004). Under the assumption that constitutively expressed genes are expressed in all different treatments, we expected

that *cms* group isotigs are more conserved than isotigs from the *c*, *m*, and *s* groups. Since the libraries were normalized and cDNA levels do not represent true expression levels, we restricted our analysis to presence-absence patterns. The differences in the fractions of BLASTX hits in single treatment groups and the *cms* group suggested that genes expressed in all three treatments are more conserved. However, if groups are corrected for the total amount of sequence data, *cms* group isotigs are not more conserved than treatment-specific isotigs, because the main determinant for a BLAST hit is isotig length which is longer in *cms* isotigs (Table 2.4). This pattern was also observed in white spruce (Rigault et al, 2011).

In addition to testing the general hypothesis that treatment-specific genes are less conserved than widely-expressed genes, we also parsed BLASTX results for drought stress related keywords to find potential drought stress related PUTs. We expect that the 1,503 transcripts with a BLASTX hit containing one of the keywords are probably involved in the Douglas-firs response to drought (Table 2.3). More than 1,100 candidate PUTs are part of the *cms* group and only 134 candidates are part of the *m*, *s*, and *ms* groups. This reflects that the response to drought seems to be mainly facilitated through changes in gene expression levels via up- or down-regulation. The small set of 134 PUTs exclusively induced by drought stress appears to contribute to a specific drought response, but this needs to be further verified because their expression pattern may reflect a sampling artifact. Even though the function of those PUTs may not be conserved across large evolutionary distances, the identified PUTs serve as top candidates for further analysis of sequence and expression variation in comparisons of differentially adapted (e.g., coastal and interior) Douglas-fir provenances.

Analysis of genetic variation

The construction of the cDNA libraries representing genetically different provenances allowed the detection of SNPs for later analysis of patterns of genetic variation. The two most important results are the high proportion of shared polymorphisms and the strong influence of the SNP calling algorithm on the number of detected SNPs. By taking a conservative approach and considering only those SNPs that were called by all three programs, only 27,688 (highly reliable) SNPs were obtained, which is only about half of the number detected with gsMapper, which identified the lowest number of SNPs (57,691). Since the numbers of called SNPs differed highly between SNP detection tools, our results indicate that SNP calling from next-generation

sequencing data depend to a high degree on the software tools used. Therefore, results should be interpreted with caution, if relying on a single SNP detection approach only. To our knowledge there are no systematic studies yet that compared the accuracy of different SNP callers with next-generation sequencing data.

A comparison of the SNP density (SNPs per kb) of the most reliable SNPs with published data shows that the former is an underestimate of the true level of sequence variation in Douglas-fir. The SNP density is 1 SNP for every 2,530 bp, whereas other studies estimated an average SNP density from 1 SNP per 43 bp to 1 SNP per 112 bp using Sanger sequencing protocols (Eckert et al, 2009d). The reasons for the large difference to the reported SNP density are probably the stringency criteria used and the better quality of base-calling using Sanger sequencing. If we take only the SNPs identified by bwa/SAMtools or ssahaSNP in account, the SNP density increases to 1 SNP per 820 bp and 1 SNP per 451 bp, respectively.

Nevertheless, our sequence data make a significant contribution to the number of Douglas-fir SNPs available for further applications. Until now, only around 1,300 SNPs have been identified in Douglas-fir (Eckert et al, 2009b,d). If only the most reliable SNPs are considered, a key result is the large number of SNPs whose alleles are segregating in both the coastal and interior provenances (15,483 SNPs, ci/ci category in Table 2.6). In only 5 % of SNPs (886, c/i) the two alleles are specific to coastal and interior provenances, respectively. This high proportion of shared polymorphisms indicates either a high level of shared ancestral polymorphisms between the two main Douglas-fir varieties, or recent, possibly pollen-mediated gene flow. The comparison of SNPs, in which only one of the two alleles is shared between coastal and interior provenances suggest a higher level of genetic diversity in coastal provenances because three times as many SNPs are polymorphic for both alleles in the coastal (7,158 SNPs in the ci/c group) than in the interior accessions (2,547 SNPs in the ci/i group). This difference is also observed for SNPs which were called only in either the interior or coastal provenances because no reads were available from the other provenance, respectively (817 SNPs in the c/c *versus* 473 SNPs in the i/i group). Although these results are consistent with earlier studies on the genetic diversity of Douglas-fir varieties (Aagaard et al, 1998, Li and Adams, 1989), they are also certainly influenced by the different numbers of reads originating from coastal and interior cDNA libraries (1,757,542 vs. 1,076,192). Since there are 70 % more reads from the coastal provenances, the probability of finding a polymorphism in these provenances is increased and needs to be accounted for in

further conclusions.

Different numbers of reads can be accounted for by using methods for population genetic inference developed for next-generation sequencing that account for differences in read numbers from individuals or pools of individuals in estimating allele frequencies and population parameters (Hellmann et al, 2008, Kofler et al, 2011). However, such an approach does not work in the present study because allele frequencies depend on the sampled individuals in a library, the gene expression level and the effect of normalization on read numbers. Unbiased population genetic estimators like Tajima's π can be calculated from 454 data (Futschik and Schlötterer, 2010), but as the coverage at most SNP positions is much smaller than the total number of individuals, the results are not reliable. The development of genotyping and re-sequencing arrays using the present set of PUTs to estimate SNP allele frequencies and population genetic inference will allow accurate and unbiased estimates of nucleotide diversity.

2.5. Conclusions

In this study we established a catalogue of Douglas-fir putative unique transcripts (PUTs) enriched for drought stress induced genes. Although the real magnitude of the transcriptome is yet unknown, we estimate that the majority of the transcriptome has been tagged by the PUT set presented here. This is based on the results of the functional annotation and the comparison of the GO term distributions with those of *Arabidopsis thaliana* and *Picea sitchensis*. By analyzing sequence variation in the PUTs we detected thousands of new SNPs. Furthermore, we identified drought stress specific candidate sequences. Taken together these data represent a useful resource for the next steps in the characterization of the Douglas-fir genome and transcriptome and the association of genetic variation with phenotypic traits such as adaptation to different ecogeographic environments.

2.6. Methods

Plant material and library preparation

1.5 year old Douglas-fir seedlings were obtained from tree nurseries in British Columbia (Canada), Washington, Colorado, and New Mexico (USA) and grown in the greenhouse in a mixture

of soil:perlit:sand (50:25:25). All seedlings were fertilized with Osmocote Exact Hi End 5-6m (Scotts International BV, Heerlen, NL). Potted seedlings were watered every second day. Drought stress experiments started after two month of growth in the greenhouse, when visual inspection of the seedlings indicated a well developed root system. For the experiments, seedlings were randomly assigned to one of three different treatments: (1) control seedlings kept under well watered conditions, (2) mildly water stressed seedlings (predawn water potential between -0.7 and -1.0 MPa) and (3) severely water stressed seedlings (predawn water potential between -1.5 and -2.0 MPa). Water stress was imposed by withholding watering until a desired water potential had been reached (Watkinson et al, 2003). Water potential was assessed by repeated measurements of predawn needle water potential using a Scholander pressure chamber to assess the level of water stress (Ensminger et al, 2008). Within about 3-4 and 5-6 weeks, the target water potential was observed in the mildly and severely water stressed seedlings, respectively, and needles and sections of the stem (wood tissue) were harvested. Tissue from control seedlings was harvested in parallel in order to obtain samples from similarly aged plant material. Tissue samples were immediately frozen in liquid nitrogen and stored at -80°C for later extraction of RNA.

Frozen needles or sections of the stem were homogenized using mortars and pistils chilled with liquid nitrogen until a fine powder was obtained. Total RNA was extracted from 300 mg aliquots of the frozen powder using the CTAB method (Chang et al, 1993). Isolated RNA from individual seedlings was then quality checked using Qiaxcel (Qiagen, Hilden, Germany).

Aliquots of the RNA from several seedlings and several provenances were then combined for synthesis into a total of twelve pooled RNA samples. Six of these pooled samples represented a subset of coastal and six samples represented a subset of interior Douglas-fir. Each of these two subsets included two sets of pooled RNA samples from either needle tissue or from wood tissue. Finally, each of these tissue specific sets consisted of one pooled RNA sample from control, mildly water stressed or severely water stressed seedlings (Table 2.1, Supplementary Tables A.2 and A.3).

Normalized cDNA libraries were generated by Evrogen LAB (Moscow, Russia). Starting from 0.3 µg of total RNA double-stranded cDNA was synthesized using SMART Oligo II oligonucleotides and CDS primers (SMART Oligo II 5'–AAGCAGTGGTATCAACGCAGA GTACGCrGrGrG– 3', CDS primer 5'–AAGCAGTGGTATCAACGCAGAGTA-d(T)30– 3')

(Zhu et al, 2001). Amplified cDNA was then purified using the QIAquick PCR purification kit (Qiagen, CA, USA), concentrated by ethanol precipitation and then diluted to a final cDNA concentration of 50 ng/µl. SMART amplified cDNAs were then normalized (Zhulidov et al, 2004). Normalization included a cDNA denaturation/reassociation step followed by treatment with duplex-specific nuclease (DSN, Shagin et al, 2002) and subsequent amplification of the normalized fraction by PCR using SMART PCR primers (SMART PCR primer 5' –AAGCAG TGGTATCAACGCAGAGT– 3').

454 sequencing of the normalized cDNA libraries was carried out by Seq-IT (Kaiserslautern, Germany) using a Genome Analyzer FLX with 454 titanium chemistry (Roche, Basel, Switzerland). Prior to sequencing, each cDNA library was first fragmented. Fragments were tagged with multiplex identifiers (MIDs) to allow library identification of the reads obtained from parallel sequencing of the libraries on the Genome Analyzer FLX. In total three titanium runs were performed, with 1.5 runs analyzing the needle libraries, and 1.5 runs analyzing the wood libraries. The proprietary genome analyzer software was used for the first preprocessing of sequence reads including the assignment of quality scores to generate .sff-files for further processing and assembly of the data.

Preprocessing

The resulting .sff-files were extracted with the sff_extract tool (Blanca and Chevreux, 2012). All sequences with at least one 'N' were removed. The preprocessed files were used as input for SnoWhite (release 1.1.3; Dlugosch et al, 2013), a cleaning pipeline for cDNA sequences that uses SeqClean (SeqClean, 2012) and trims polyA/T. All sequences shorter than 50 bp or with a polyA/T repeat of at least 8bp at either end were discarded. The longer part of the sequence was retained if internal polyA/T tracts were detected. As the assembly program operates in flowgram signal space it is recommended to use .sff-files as input. Thus, the original .sff-files were altered according to the changes made during the preprocessing steps using custom Python scripts. Those altered .sff-files were loaded into the assembler.

Assembly and mapping

Sequences were assembled with Newbler v2.6 using default parameters supplemented by the `-cdna` and `-urt` options (454 Life Science, 2009, Kumar and Blaxter, 2010). Newbler con-

structs a set of contigs (contiguous sequences), representing assembled reads. Unassembled reads were marked as singletons, repeats, outlier (e.g. chimeric reads), or too short. Isotigs consist of contigs connected by a subset of reads (Supplementary Figure A.6). An isogroup is a group of different isotigs of the same multiple alignment. Isogroups represent genes, isotigs correspond to alternatively spliced transcripts, and contigs correspond to exons. This is a simplified view because contigs and isotigs can also contain sequences of untranslated regions. Independent contigs that were not part of an isotig were simply considered as isotigs to facilitate the analysis.

All twelve libraries were assembled together. Based on the assembly, we created a set of PUTs. We first searched for false positive singletons, i.e., reads that were marked as singletons although they matched nearly perfect to an existing isotig. For this purpose, all reads marked as singletons were mapped to the isotigs of the assembly using ssaha2 (Ning et al, 2001) with default parameter settings. Reads were mapped only if the pairwise sequence identity with a reference isotig was at least 98 % of the alignment length. Unmapped reads were considered as real singletons and checked for duplicates. The final PUT set consisted of the isotigs and the singletons of the assembly representing all different transcripts found in the data set. In particular, i.e., PUTs can be the only possible transcript of a gene, only a part of a longer transcript that can not be found within the data, alternatively spliced variants of a gene, but also the product of misassemblies. Sequences shorter than 100 bp were excluded to dismiss potentially uninformative sequences.

SNP detection

SNPs were identified with GSMapper (454 Life Science, 2009), ssahaSNP (Ning et al, 2001), and bwa/SAMtools (Li and Durbin, 2009, Li et al, 2009). Each program detected a different number of SNPs. Therefore, we combined the results of the three programs and considered the SNPs identified by all three tools as a set of potentially most reliable SNPs (Figure 2.5). We used the sequences of the PUTs derived from the assembly as reference for the SNP detection. To avoid sequencing errors from being considered as SNPs, we required for each tool that the reference nucleotide as well as the variant nucleotide were confirmed by at least three reads each. Hence, the minimum coverage per SNP position was six reads.

GSMapper v2.6 was run with default parameters for cDNA libraries. We constructed a .sff-

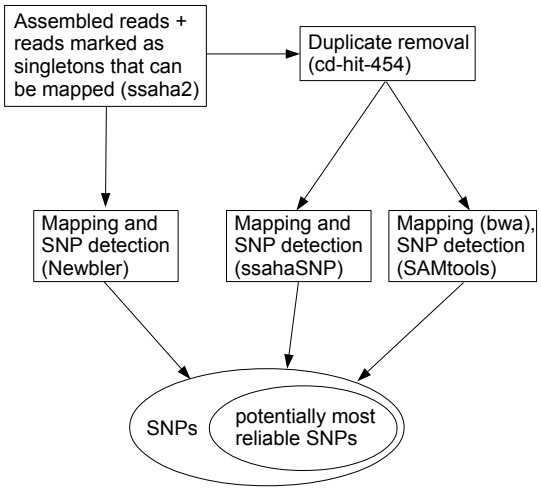


Figure 2.5.: All reads that were assembled and all reads that were marked as singletons but that can be mapped using ssaha2 (false positive singletons) served as input. Mapping of the reads and SNP detection was performed by three programs: Newbler, ssahaSNP, and bwa/SAMtools. For the latter two, the duplicates were removed using cd-hit-454. The workflow resulted in a set of SNPs, whereby those found by all tools are probable the most reliable SNPs.

file including all assembled reads of the assembly and all singleton reads that were mapped to the isotigs with ssaha2 (false positive singletons). All reads of that .sff-file were mapped against the reference sequences with GSMapper. The resulting file was parsed for SNPs using a custom script.

ssahaSNP v2.5.3 does not handle duplicate reads internally like GSMapper. Therefore, duplicate sequences were removed from the set of all assembled reads and all false positive singletons using cd-hit-454 v3.1.2 (Niu et al, 2010). A .fastq-file was produced using the corresponding fasta and quality files of the non-duplicate sequences. ssahaSNP was run with default parameters and mapped the reads of the .fastq-file against the PUTs. The results were further processed using the parse_SNP tool provided within the ssahaSNP package and custom scripts to extract SNPs that matched our criteria.

The third approach combined two tools, bwa v0.5.9 for mapping and SAMtools v0.1.16 for variant calling. The same .fastq-file that was used for ssahaSNP was used as input for bwa. As recommended for 454 reads, the `bwasw` option of bwa was used. The reads were mapped against the sequences of the PUTs. SAMtools was applied to convert the resulting .sam-file to a sorted

.bam-file and to call the variants in that .bam-file. The resulting SNPs were again parsed to report only those SNPs with at least three reads confirming the variant and at least three reads confirming the reference nucleotide. The final set of potentially most reliable SNPs was obtained by combining the results of the three approaches and extracting those SNPs that were detected by all three programs.

Synonymous and non-synonymous polymorphisms were detected using the results of the BLASTX search of the assembly against NCBI's non-redundant *nr* database (see below). All high-scoring segment pairs (HSP) of the top hit of each query were considered. Using the information of the BLAST results, we examined whether a SNP was in a coding or non-coding region of a gene. For SNPs in coding regions, we determined the amino acid at the corresponding position with the reference nucleotide as well as with the variant nucleotide to call synonymous or non-synonymous SNPs.

According to the criteria for SNP detection, a transcript was covered by at least six reads at each SNP position and at least three reads had to confirm each allele of a SNP. Each of the two alleles of a biallelic SNP can therefore include reads from coastal or interior varieties only, or from both varieties, resulting in nine combinations, which are summarized in Table 2.5. Since there is no reference genome sequence of Douglas-fir available, it was not possible to decide which of the two nucleotides was the reference or the variant nucleotide. Therefore, we pooled some combinations to compare the results independently of the classification of a nucleotide as reference or variant in our results.

BLAST searches and annotation

To investigate the evolutionary conservation of the transcripts, we constructed two databases: one containing *Picea sitchensis* protein sequences downloaded from the NCBI data repository (*picea* database NCBI data repository, 2012) and one containing *Arabidopsis thaliana* sequences downloaded from TAIR (*ara* database Swarbreck et al, 2008). The *picea* database consisted of 18,816 and the *ara* database of 35,381 sequences. PUTs were blasted against those two databases as well as against NCBI's non-redundant *nr* database using BLASTX v2.2.25+ with an e-value threshold of 10^{-10}.

Results of the BLASTX search of PUTs against *nr* database were used as input for Blast2GO v2.4.9 (Conesa et al, 2005). Blast2GO was utilized for the functional annotation with gene on-

tology (GO) terms. The first step in Blast2GO was the mapping, in which GO terms associated with the hits obtained during the BLASTX search were retrieved. In the annotation step, functional terms were assigned to the sequences based on the retrieved set of GO terms per sequence using Blast2GO's annotation score. Furthermore, we used a local version of InterProScan version 4.8 (Zdobnov and Apweiler, 2001) to search protein signatures in the InterPro database (Hunter et al, 2009). With the local version it was possible to analyze nucleotide sequences in all six possible open reading frames. Due to the long running time of some of the InterProScan applications, we used only a subset of them that included blastprodom, fprintscan, hmmpfam, hmmpanther, hmmtigr, hmmsmart, patternscan, and seg (Hunter et al, 2009). The results of the InterProScan were imported into Blast2GO to improve annotations. Annotations were further refined using Annex and GO-Slim, both of which were available within Blast2GO (Conesa and Götz, 2008, Myhre et al, 2006). Annex augments annotations by finding relationships between different GO terms and adding implicit annotations. GO-Slim represents a reduced set of GO terms that gives a useful summary of the all GO terms. Blast2GO provides organism specific GO-Slim mappings of which the plant specific mapping was chosen. For a better comparison of GO terms, functional annotations were generated for the protein sequences of *P. sitchensis* and *A. thaliana* used in the *picea* and *ara* databases. A BLASTP (v2.2.25+) search with an e-value of 10^{-5} against NCBI's non-redundant protein sequences was done before running Blast2GO. We did not annotate these two data sets with InterProScan, but with Annex and GO-Slim. The results of functional annotation of PUTs were compared to the results of the functional annotation of *P. sitchensis* and *A. thaliana*.

Identification of drought stress related genes

Two approaches were used to identify potential drought stress related genes. In the first approach, we divided the non-singleton PUTs, i.e., the isotigs, of the assembly by the origin of their reads into seven groups. The groups were named according to the libraries from which the reads were derived (*c*, *m*, *s*, *cm*, *cs*, *ms*, *cms*, where *c* stands for control, *m* for mild stress, and *s* for severe stress, *cm* for control and mild stress etc.). The isotigs in the *c*, *m*, and *s* groups were assumed to contain most likely treatment-specific sequences, as they contained isotigs composed of only reads of one treatment. Therefore, we expected to find drought stress related sequences mainly in the *m* and *s*, but also in the *ms* groups. For the second approach, the BLASTX results

were searched for specific keywords to identify candidate genes previously assigned to drought, water stress, or other stress related pathways (Table 2.3; Hamanishi and Campbell, 2011, Shinozaki and Yamaguchi-Shinozaki, 2007, Wang et al, 2004).

Data availability

The sequence reads were submitted to the ENA Sequence Read Archive (SRA) under study accession number ERP001358 (http://www.ebi.ac.uk/ena/data/view/ERP001358). PUTs, annotated SNPs, and Blast2GO results will be available from http://www.treeversity.org. BLASTX results and Python scripts used for the analysis are available upon request.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

IE designed, conducted and coordinated greenhouse and experimental work. IE and KS designed the sequencing experiment. TM analyzed the sequence data. TM, IE and KS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study is part of the collaborative project 'DougAdapt' with funding from the DFG to IE (DFG EN 829/4-1) and KS (DFG SCHM1354/3-1). The authors are grateful to Daniel Landwehr, Claudia Waack and Anna-Maria Weisser for assistance during greenhouse experiments. James Joyce, BC Timber sales, BC, Randy Moench, Colorado State University Nursery, CO and Bill Taylor, Webster Forest Nursery, WA kindly provided seedlings material for this study. We are grateful to Fabian Freund for advice about statistical analysis and to Henning Wildhagen and Sariel Hübner for their feedback and suggestions on an earlier version of the manuscript. We gratefully thank the bwGRiD project (http://www.bw-grid.de), member of the German D-Grid initiative, funded by the Ministry for Education and Research (Bundesministerium fuer Bildung und Forschung) and the Ministry for Science, Research and Arts Baden-Wuerttemberg (Ministerium fuer Wissenschaft, Forschung und Kunst Baden-Wuerttemberg) for

the use of the computational resources.

3. Targeted re-sequencing of five Douglas-fir provenances reveals population structure and putative target genes of positive selection

Thomas Müller[1], Fabian Freund[1], Henning Wildhagen[2,3] and Karl J. Schmid[1]

[1]Department of Crop Biodiversity And Breeding Informatics, University of Hohenheim, Stuttgart, Germany, [2]Forest Research Institute (FVA) Baden-Württemberg, Freiburg, Germany, [3]Department of Forest Botany and Tree Physiology, Georg-August-University, Göttingen, Germany

With kind permission from Springer Science and Business Media: Tree Genetics & Genomes 2015, 11:816, doi:10.1007/s11295-014-0816-z

3.1. Abstract

Douglas-fir (*Pseudotsuga menziesii*) occurs in a coastal and an interior variety that differ in drought tolerance and other adaptive traits. To characterize genome-wide levels of genetic diversity in coding regions and to identify genes involved in local adaptation, we used targeted sequence capture to re-sequence 72 trees representing one interior and four coastal provenances. A custom NimbleGen sequence capture array was designed from 57,110 putative unique transcripts (PUTs) to enrich genomic sequencing libraries for these regions. Sequence analysis revealed that almost 100% of target regions were captured and sequenced in at least one individual. We found 79,910 single nucleotide polymorphisms (SNPs) whose genotypes were called in all individuals. The data confirmed genetic differentiation between interior and coastal provenances and revealed little differentiation between coastal provenances. The nucleotide diversity of the total sample was estimated as $\pi = 0.0032$, which is at the lower end of values observed in conifers. Outlier tests of genetic differentiation identified 58 high-confidence candidate genes for directional selection with a broad functional diversity. *A priori* defined genes

involved in drought tolerance showed a significantly higher genetic differentiation between interior and coastal Douglas-fir suggesting a different evolution despite a low level of polymorphism. The observed data showed a reduced level of polymorphisms with low minor allele frequencies compared to standard demographic models with two populations and migration. Targeted sequence capture is an efficient method to characterize the genetic diversity of conifer trees with a complex genome.

3.2. Introduction

Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) is one of the most important timber trees in North America because of its rapid growth characteristic and high-quality wood. In its native distribution range in North America, this conifer exists in two main varieties. The coastal Douglas-fir (*P. menziesii* var. *menziesii*) occurs in the coastal range from British Columbia to central California, and the interior Douglas-fir (*P. menziesii* var. *glauca*) extends inland from British Columbia along the Rocky Mountains down to Mexico. The large area inhabited by Douglas-fir reflects its adaptation to diverse ecozones by genetic differentiation and natural selection (Campbell, 1979, Campbell and Sugano, 1979, Dean, 2007, Rehfeldt, 1989), which was confirmed in common garden experiments or provenance trials (Kleinschmit and Bastien, 1992). Studies in other conifer species found similar patterns of adaptive divergence suggesting the selective fixation of genetic variants (Neale and Ingvarsson, 2008, Neale and Savolainen, 2004).

The ongoing climate change presents a significant challenge for the conservation of autochthonous populations and the selection of suitable provenances for commercial forestry in North America and other temperate regions of the world, where Douglas-fir has been introduced as timber tree (Kleinschmit and Bastien, 1992). For example, the future climate of Central Europe will likely be characterized by more frequent and longer periods of summer drought and less precipitation (IPCC, 2007). Douglas-fir is of great interest for the European timber industry because it is more drought-tolerant than native European conifers such as Norway spruce (Hanewinkel et al, 2013). Furthermore, it shows strong growth potential and low susceptibility to diseases and pathogens in Europe (Ducić et al, 2008, Hermann and Lavender, 1999, Kohnle et al, 2012). Differentially adapted Douglas-fir provenances are already evaluated in a diversity of European environments using phenotypic traits (Kohnle et al, 2012) to identify suitable provenances.

The rapid development of genome sequencing technologies and the statistical analysis of genetic and phenotypic variation will greatly facilitate the identification of provenances that harbor useful genetic variation for required traits such as drought tolerance (Neale and Kremer, 2011). Compared to other plant taxa, genome analysis of conifer species lags behind because of their large genome sizes. Only few whole genome assemblies are currently available that include Norway spruce (Nystedt et al, 2013) and loblolly pine (Neale et al, 2014, Zimin et al, 2014). Douglas-fir has a genome size of 19 Gbp (Ahuja and Neale, 2005) and a whole genome sequencing project is underway (Neale et al, 2013). Whole genome sequencing is still prohibitively expensive for surveying genome-wide genetic variation in conifers, but sequencing of transcriptome or reduced representation libraries are suitable alternatives. For Douglas-fir, transcriptome sequencing data for gene identification, annotation, and surveys of sequence diversity are already available (Howe et al, 2013, Müller et al, 2012).

Sequence capture is a convenient and cost-efficient method to re-sequence complex genomes (Grover et al, 2012). Predefined target regions are captured from fragmented DNA libraries of individuals using custom oligonucleotides that are complementary to target regions such as exons. Only captured library fragments are sequenced, which reduces the extent of sequenced regions and therefore sequencing costs. The two major methods for sequence capture use either DNA or RNA oligonucleotides as probes. DNA-based oligonucleotides were successfully applied in plant species with complex genomes including soybean (Haun et al, 2011), barley (Mascher et al, 2013), and wheat (Henry et al, 2014).

The purpose of the present study was to establish sequence capture for investigating patterns of genetic diversity and adaptation in Douglas-fir. We re-sequenced 72 individual trees representing five differentially adapted Douglas-fir provenances included in European provenance trials. Since no reference genome was available, DNA-based capture probes were designed from 57,110 putative unique transcripts (PUTs) that likely represent a large proportion of the Douglas-fir protein coding genes (Müller et al, 2012). We included genes with a putative role in drought tolerance based on sequence similarity to homologs in other species. Tests of natural selection based on genetic differentiation identified high-confidence candidate genes for directional selection in provenances. However, we could not find any typical selection patterns comparing observed genetic diversity with standard neutral models. Further analysis suggested that such models were not suitable to describe the observed diversity. Our results show that sequence

capturing of target regions in complex genomes is a suitable cost-efficient approach to survey genetic variation in species with a large genome size and without a reference genome.

3.3. Materials and methods

Plant material

The Douglas-fir trees analyzed in this study were grown at three experimental sites in South-western Germany (Wiesloch/Philippsburg, Sindelfingen, and Schluchsee; Figure 3.1b), which were established in 1961 for an international Douglas-fir provenance trial (Kohnle et al, 2012). At each site, trees of the five provenances Salmon Arms (AR), Conrad Creek (CR), Cameron Lake (LA), Santiam River (RI), and Timber (TI) were sampled (Figure 3.1a; Online resource 2). The AR provenance is an interior and the other four provenances are coastal varieties. Needles of 25, 24, and 23 dominant trees with superior shapes of stem and crown were sampled on the experimental sites in Wiesloch/Philippsburg, Sindelfingen, and Schluchsee, respectively. Based on this design, we sampled four to five trees per provenance and site.

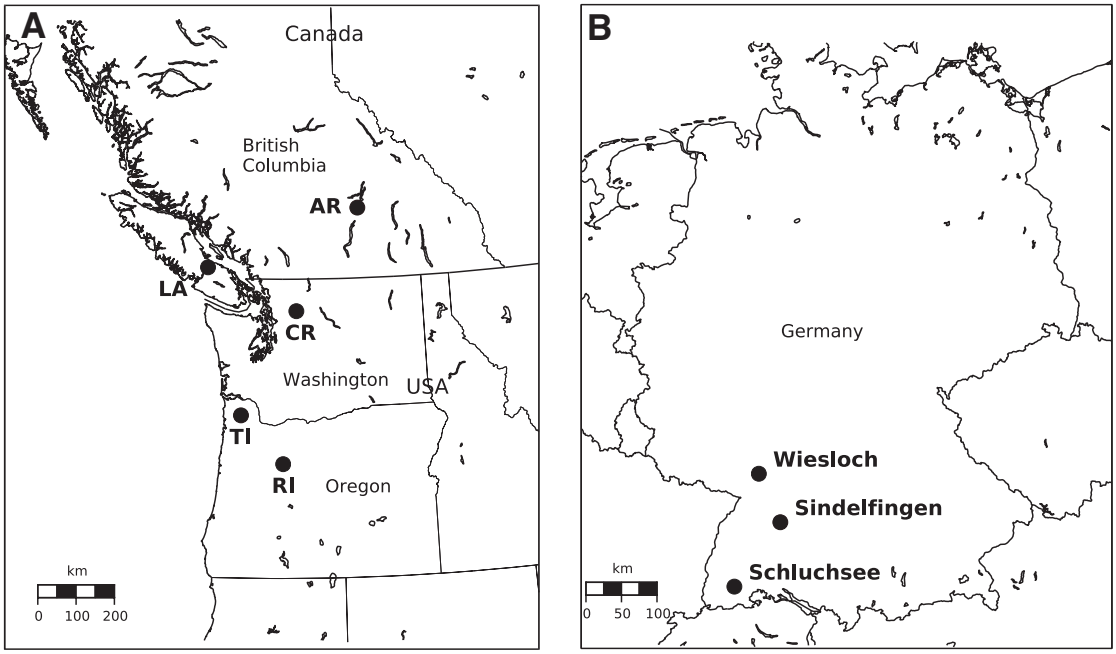


Figure 3.1.: a) Geographic origin of the provenances in North America (AR - Salmon Arms, CR - Conrad Creek, LA - Cameron Lake, RI - Santiam River, TI - Timber). b) Location of the field trials in Germany from which the needles for DNA extraction were harvested.

DNA extraction

Needle samples were collected in the field and immediately frozen. In the lab, samples were homogenized in a mixer mill (MM30, Qiagen, Hilden, Germany) and DNA was extracted with the DNeasy 96 plant kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol with the following modification. After grinding in the lysis solution, samples were incubated for 40 min at 65 °C in a water bath and after the addition of the precipitation buffer AP2, samples were incubated for 30 min on ice. DNA concentration was measured with a NanoDropTM1000 instrument (Thermo Scientific, Karlsruhe, Germany).

Capture array design

Target regions for re-sequencing were selected from previously published PUTs of *P. menziesii* (Müller et al, 2012). A set of 134 candidate PUTs involved in drought response were identified by a BLASTX keyword search against NCBI's *nr* database and listed in Müller et al (2012). These drought-related candidates as well as all PUTs with a sequence longer than 449 bp were selected as target regions. We selected the longest PUT of a given gene, if a gene encoded for multiple transcripts because of alternative splicing. After selection, 57,110 PUTs with a total length of 46 Mbp were retained.

The sequence capture array with probes complementary to target regions was designed by NimbleGen (Roche NimbleGen, Madison, WI, USA) using a proprietary algorithm. Each probe differed by at most three base insertions, deletions, or mismatches from the target PUTs. Circa 2.1 million probes with an average length of 75 bp (minimum 50 bp, maximum 120 bp) were synthesized by NimbleGen. The resulting array was named 120412_Pseudotsuga_TM_EZ_HX1.

Library preparation

Sequence capture with the array was carried out with the NimbleGen SeqCap EZ Developer Kit (Roche NimbleGen, Madison, WI, USA). DNA library preparation and capturing were performed according to the manual (NimbleGen, 2011) with modifications described in Appendix 2 of Burgess (2011). For multiplexing, we used the TruSeq DNA Sample Preparation Kit v2 Set A (Illumina, San Diego, CA, USA).

Sequencing libraries were made from 1 μg of genomic DNA (gDNA) of each tree after shear-

ing with a Covaris S2 instrument (Covaris Inc., Woburn, MA, USA). Indexed libraries were prepared by end repairing, A-tailing, and ligation of Illumina adapters according to the manufacturer's protocol. Eight cycles of a pre-capture ligation mediated PCR (LM-PCR) were carried out using 20 μl of each indexed library. Twelve Illumina indices were used to tag the 72 individuals before combining them into six pooled libraries of twelve individuals each with a total amount of 1.1 μg DNA per library. In the hybridization step, 20 μl SeqCap EZ Developer Reagent (NimbleGen) were added to prevent unspecific hybridization. Streptavidin Dynabeads were used for DNA washing and sequence capturing. After a second LM-PCR with 18 cycles, amplified DNA was cleaned with a Qiagen QIAquick PCR Purification Kit (Qiagen, Chatsworth, CA, USA). The concentration and size distribution of libraries were assessed with an Agilent DNA 1000 Chip (Agilent Technologies, Waldbronn, Germany).

Sequencing and preprocessing

Each of the six pooled and enriched libraries was sequenced in a single lane of an Illumina HiSeq1000 sequencer with paired-end sequencing of 100 bp at the Kompetenzzentrum Fluoreszente Bioanalytik (KFB), Regensburg, Germany. Reads were quality-checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and filtered using custom Python scripts. The first three as well as the last seven bases of the reads were removed due to a general drop of quality in those regions. Sequences were excluded if they contained at least one 'N' (undefined base) or if more than 5 % of the bases had quality scores below 20. The cutadapt program (Martin, 2011) excluded reads with adapter contamination in the first 13 bases of the Illumina adapter (`cutadapt -O 10 -b AGATCGGAAGAGC -e 0.05`). Only read pairs that passed all filtering steps were kept for further analyses.

SNP detection

Mapping was performed with pBWA (Peters et al, 2012) by allowing five mismatches per read. Only reads with a mapping score equal or greater than 1 (`samtools view -q 1`) were retained. Coverage was calculated with BEDTools (Quinlan and Hall, 2010) and custom scripts. PLINK v1.07 (Purcell et al, 2007) was used to convert .vcf files to .bed files. SNPs were called with SAMtools v0.1.18 (Li et al, 2009) (`samtools mpileup -D -g -C 50 -S -A`) followed by bcftools (`bcftools view -bvcg`) and vcfutils (`vcfutils.pl varFilter -D 10000`). The

resulting .vcf file was filtered for variants with a minimum variant distance bias of 0.015, a minimum strand bias of 0.01, a minimum end distance bias of 0.05, and a minimum quality of 30 using vcf-annotate (Danecek et al, 2011, 2012). These filtering steps are expected to remove most PCR duplicates (Supplementary Info. B). Heterozygous SNPs called by this approach consisted of at least one reference and one alternative allele per individual. SNPs for which <10 reads in the total sample confirmed the variant were excluded from further analysis. Furthermore, SNPs with missing information in at least one individual were also excluded, resulting in a data set without missing data in all 71 individuals.

Population structure

Population structure was inferred with the SNP data. A discriminant analysis of principal components (DAPC) was performed with the R package *adegenet* by including provenance information (Jombart, 2008, Jombart et al, 2010). In DAPC, a principal components analysis (PCA) is followed by a discriminant analysis (DA). For the DA, a sufficient number of principal components (PCs) needs to be retained, but too many PCs cause over-fitting. We used the broken-stick criterion to retain 25 significant PCs (Frontier, 1976, Jackson, 1993). Furthermore, two discriminant functions were retained. A principal coordinate analysis (PCoA) was carried out based on pairwise F_{ST} values among the five provenances. PCoA and pairwise F_{ST} calculations were performed with VCFtools v0.1.11 and the R-package *ape* (Danecek et al, 2011, Paradis et al, 2004). The population structure of provenances was further inferred with ADMIXTURE v1.23 (Alexander et al, 2009) with the hypothetical number of populations K ranging from 2 to 7, random seeds, and the convergence criterion of log-likelihood increase between two consecutive iterations set to 10^{-6}.

Population genetic parameters

Nucleotide diversity π, tests of Hardy-Weinberg equilibrium (HWE), Tajima's D (Tajima, 1989), and F_{ST} values were calculated from .vcf files with VCFtools (Danecek et al, 2011). Nucleotide diversity π was calculated for each PUT by dividing the sum of the π values of each polymorphic site by the number of bases with non-zero coverage in all individuals. VCFtools uses the bi-allelic version of Weir and Cockerham's F_{ST} to calculated F_{ST} values for each SNP as well as for each PUT (Weir and Cockerham, 1984). Since F_{ST} was estimated according to Weir and

Cockerham (1984), negative values were possible. Tajima's D values were calculated per PUT using the method of Carlson et al (2005). Deviations of SNP genotypes from HWE were based on an exact test (Wigginton et al, 2005), which excludes sites with a p value below a threshold of 10^{-5}.

Identification of strongly differentiated genes

Genes with high levels of genetic differentiation between populations were identified using the outlier detection approaches implemented in LOSITAN (Antao et al, 2008), BayeScan v2.1 (Foll and Gaggiotti, 2008), and Bayenv2 (Günther and Coop, 2013). Two population models were investigated. In the 5-populations model, each provenance was considered a separate population, whereas in the 2-populations model, all coastal provenances were considered a single population and the AR provenance a separate population. LOSITAN uses the *fdist* F_{ST} method for outlier detection (Beaumont and Nichols, 1996). We performed five runs with 100,000 simulations with a 'neutral' mean F_{ST}. The infinite alleles mutation model was applied in all runs. The confidence interval was set to 0.95, the false discovery rate to 0.1, and the subsample size to 14. The number of expected populations was set to 2 or 5, depending on the model under consideration.

For both population models, we performed three runs of BayeScan with standard parameters, but the prior odds of the neutral model set to 100 as recommended in the manual for the observed SNP number. Higher prior odds eliminate false positives, but may increase the proportion of false negatives. For outlier detection, we used a false discovery rate of 0.1. The statistical power of BayeScan results is likely reduced because we consider only models of two and five populations resulting in the loss of candidate loci and the identification of extreme outlier loci only (Foll and Gaggiotti, 2008, Helyar et al, 2011).

Bayenv2 was used to compute the $X^T X$ test statistic, which is similar to F_{ST} (Günther and Coop, 2013), but accounts for the variance-covariance structure among populations. Bayenv2 first calculates a covariance matrix from a set of unlinked markers. To obtain these markers, we compared PUTs to the protein sequences of *Pinus taeda* v1.0 (http://loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/) with BLASTX (version 2.2.28+) and searched for synonymous SNPs. One synonymous SNP per PUT was randomly selected for the set used to calculate the covariance matrix. Different covariance matrices were calculated for the 2- and 5-populations models. The matrices were obtained after 100,000 iterations. Subse-

quently, SNP-wise X^TX values were also calculated after 100,000 iterations using the respective covariance matrix. Bayenv2 does not calculate p values, therefore X^TX values were ranked empirically.

Multiple Monte Carlo significance tests based on Tajima's D were used to identify PUTs that reject a standard neutral model. Tests were performed separately for the coastal and the interior population as well as for the joint population. The standard neutral model at each PUT was given by the standard coalescent with scaled mutation rate estimated by Watterson's estimate $\hat{\theta} = S/\sum_{i=1}^{n-1} i^{-1}$, in which S is the number of observed mutations in the PUT. Data were simulated with `ms` (Hudson, 2002). For each PUT, the significance at level α was tested by checking whether the observed value lies below the $(\alpha/2)$ or above the $(1-\alpha/2)$ quantile of the Tajima's D distribution under the neutral model. Monte Carlo p values for these tests were computed with the fixed h method (Besag and Clifford, 1991) with h = 20 and a maximal number of 1,000,0000 simulations. Corrections for multiple tests were done by controlling a family-wise error with a Bonferroni-Holm correction and a false discovery rate control with the Benjamini-Hochberg method. The overall significance level and the false discovery rate were both set to 0.1. Controlling for multiple testing was performed for each population and the whole population separately. The same analysis was performed also for the neutral coalescent with recombination allowed between adjacent sites as modeled in `ms`. A per-site recombination rate of 4.4×10^{-3} per generation was used (Jaramillo-Correa et al, 2010).

Comparison of observed genetic diversity with neutral models

To assess whether a standard neutral model is appropriate to describe the data, approximate Bayesian computation (ABC) (Beaumont, 2010) was used to fit two neutral demographic models to a set of summary statistics observed on the data. The demographic models reflect the two main varieties of Douglas-fir with bi-directional migration. We used an island (*I*) model (Figure 3.2a) and a population-split (*PS*) model (Figure 3.2b). The *I* model was included as it is the simplest model with population structure and the *PS* model to investigate if the divergence time between the two populations needs to be taken into account. Within prior parameter ranges for the models, ABC was applied to find the parameter sets that fitted best to the observed data based on coalescent simulations of neutral mutations.

Both models were defined for PUT sequence alignments of different lengths. We assumed the

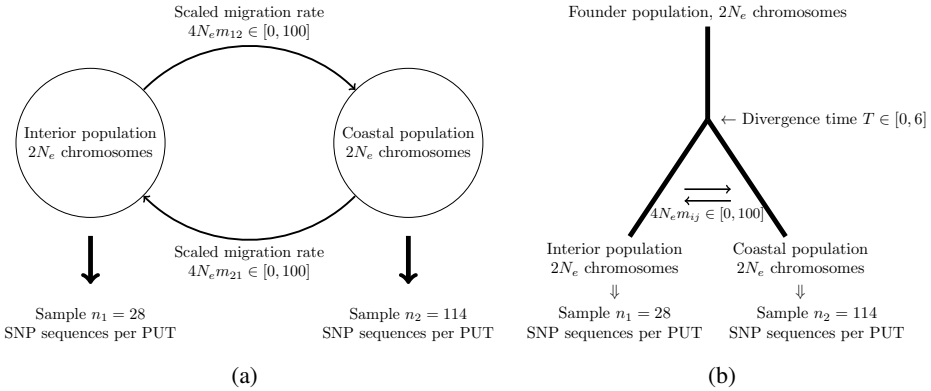


Figure 3.2.: (a) *I* model used for estimating model parameters by fitting simulated to observed summary statistics with ABC. (b) *PS* model used for estimating model parameters by fitting simulated to observed summary statistics with ABC.

Table 3.1.: Parameter ranges and specifications for the demographic models *I*, *PS*, I_1, $I_{\leq 2}$. All models were defined with uniform priors on parameter ranges. 'freq=k' indicates that all SNPs with absolute frequency *k* were removed from simulated sequences.

	θ	ρ	$4N_em_{12}$	$4N_em_{21}$	T	freq=1	freq=2
Range	$[0, 0.007]$	$[0, 0.02]$	$[0, 100]$	$[0, 100]$	$[0, 6]$		
Model	all	all	all	all	*PS*	I_1, $I_{\leq 2}$	$I_{\leq 2}$

same effective population size N_e in both populations (i.e., the *I* model has a total population size of $2N_e$ individuals) and in the ancestral population (*PS* model), asymmetric migration, and the same mutation and recombination rates in all PUTs. We also assumed that PUTs are unlinked and recombination occurs only within PUTs with 100 recombining units within each PUT. For each PUT, 114 SNP sequences corresponding to 57 diploid individuals from the coastal population and 28 SNP sequences (14 individuals) from the interior population were sampled. Variable parameters were the per-site mutation rate, θ, and recombination rate, ρ, two migration rates between coastal and interior populations, $4N_em_{12}$ and $4N_em_{21}$, and population divergence time, T, in the *PS* model. Parameter values were drawn as flat priors from parameter ranges given in Table 3.1. For θ and ρ, ranges include previous genome-wide estimates of $\theta = 5.886 \times 10^{-3}$ and $\rho = 4.375 \times 10^{-3}$ to $\rho = 4.434 \times 10^{-3}$ from Douglas-fir (Jaramillo-Correa et al, 2010). Parameter ranges for the scaled migration rates $4N_em_{12}$ and $4N_em_{21}$ allow intermediate to very high levels of migration and gene flow. The divergence time, T, was drawn from $[0,6]$ coalescent time units because already with intermediate migration rates, 99 % of any sampled pair from

both populations converged to the recent common ancestor before time unit 6.

The impact of low-frequency mutations was assessed by simulating the I model two more times after removing all polymorphisms segregating in one (model I_1) or up to two individuals (model $I_{\leq 2}$). For each model, 500,000 simulations of summary statistics were conducted with `msABC` (Pavlidis et al, 2010), using the length of randomly selected PUTs as input. To reduce computational time, statistics were simulated for only 1,000 PUTs. For the observed data and for each simulation, the means and variances of the following statistics across all PUTs were computed and used as summary statistics: Tajima's D within the interior, within the coastal, and within the whole population, Watterson's θ estimator, nucleotide diversity (not corrected for PUT length) as well as percentages of shared, fixed, and private alleles in the two populations. If simulated genes were not polymorphic, we treated Tajima's D and percentages of shared, fixed, and private alleles as missing values, analogous to msABC.

We performed an ABC model comparison with the package *abc* based on the rejection algorithm to assess which model describes the data better (Csilléry et al, 2012). All simulations from all models compared were pooled and the 0.5 % of simulations with the smallest Euclidean distance of their summary statistics with the observed summary statistics were retained. The Bayes factor $P(\text{Model } A|\text{data})/P(\text{Model } B|\text{data})$ between two models A and B given the observed summary statistics was approximated by the ratio of simulations retained from each model. An approximate sample from the posterior parameter distributions within each model given the observed data was constructed via ABC. The approximate sample from the posterior distributions was given by the parameters of the 1 % of simulations from the chosen model with the smallest Euclidean distance of their summary statistics to the observed ones. Using this sample, a posterior predictive check was performed by simulating the summary statistics under the model using the parameters from the posterior sample and comparing them to the observed values. The ABC model comparison and the construction of the posterior sample was conducted with the R package *abc* using the rejection algorithm. For posterior predictive checks, the summary statistics were simulated with `msABC`.

3.4. Results

Sequencing and sequence capturing

We re-sequenced 72 Douglas-fir trees from five provenances with a sequence capture array. Four of the five provenances were coastal provenances (Conrad Creek, CR; Cameron Lake, LA; Santiam River, RI; Timber, TI) and one was an interior provenance (Salmon Arms, AR). The set of sequenced provenances was determined by the availability of plant material at the German experimental sites.

The sequencing of 72 individuals resulted in 2.1 billion reads. After preprocessing and exclusion of a single outlier tree, which was identified by DAPC and phenotypic analyses (Supplementary Info. B), 1.6 billion reads were retained for further analyses. Read counts ranged from 12.1 to 44.7 million reads per individual (Supplementary Figure B.1) and from from 280 (LA) to 370 (CR) million reads per provenance (Supplementary Figure B.2). Between 32 and 52 % of reads per individual were mapped against the target PUTs. Reads from single individuals covered between 43 and 72 % of target regions at least once. Individuals from different provenances did not differ substantially in the proportion of target sequences not covered (Supplementary Figure B.3), which could have biased the population genetic analysis. Approximately 17 % of target nucleotides were covered by reads from all individuals (Figure 3.3). In total, 90 % of target nucleotides had a coverage of at least 10× over all 71 individuals and 97 % were covered by reads from at least one individual (for mean coverage per individual see Online resource 2). Only 116 out of 57,110 ($\sim$0.2 %) target PUTs were not covered at all and were probably not captured.

SNP calling

SNP calling identified 79,910 SNPs in 15,277 PUTs without missing data, i.e., SNPs were covered in all 71 individuals (Supplementary Figure B.4) of which 52,190 SNPs were shared between the five provenances, 1,530 SNPs were polymorphic only in the interior, and 9,642 SNPs only in the coastal provenances (Table 3.2). No SNP was fixed in any provenance. A total of 2,542 SNPs were not polymorphic in the data set and polymorphic only in comparison to the PUT reference sequence.

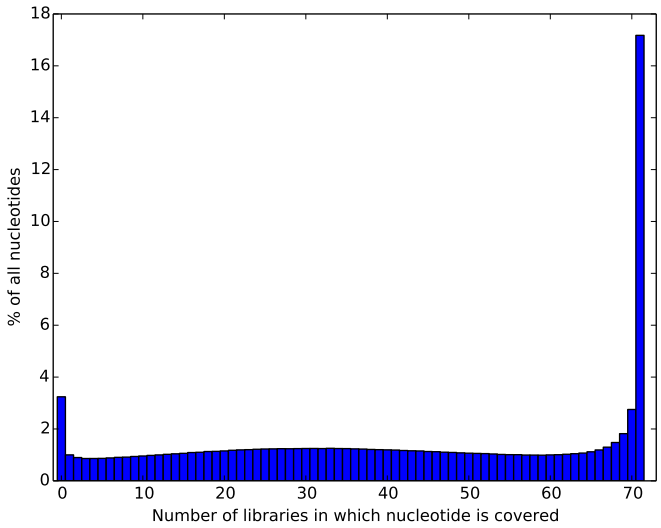


Figure 3.3.: Percentage of target nucleotides covered by reads from a given number of individuals. For example, 3.2 % of nucleotides are not covered and 17.2 % are covered with reads from all individuals.

Table 3.2.: Numbers of private SNPs in the provenances and the two varieties.

Coastal					Interior
CR	LA	RI	TI	Total	AR
348	345	279	327	9,642	1,530

Population structure

To quantify differences in the expected heterozygosity among provenances, we calculated the fixation index F_{ST} for each PUT with at least one SNP. The mean weighted average of F_{ST} values indicated little differentiation between all provenances ($F_{ST} = 0.0065$, s.d. = 0.0272) and even less differentiation between coastal provenances ($F_{ST} = 0.0015$, s.d. = 0.0188). The mean weighted average F_{ST} value of the 2-populations model was 0.0128 (s.d. = 0.0473, Supplementary Figure B.5). Pairwise F_{ST} values between the five provenances were low and ranged from 0.00024 to 0.01134 (Table 3.3). A PCoA of the F_{ST} distance matrix revealed a clear differentiation of the interior AR provenance from the coastal provenances on the first axis (Supplementary Figure B.6a), which explained 97 % of the observed variation, whereas the second axis explained only 3 %. The DAPC results were highly similar to the PCoA (Figure 3.4a, Supplementary Figure B.7). SNPs with high loading values that contribute most to the differentiation between

groups in a DAPC also showed high SNP-specific F_{ST} values (Supplementary Figure B.8). The ADMIXTURE analyses with values of K= 2 to K = 5 confirmed the differentiation between interior and coastal provenances, and a high level of admixture among coastal provenances (Figure 3.5).

Table 3.3.: Pairwise F_{ST} values per PUT and standard deviations for each pair of provenances.

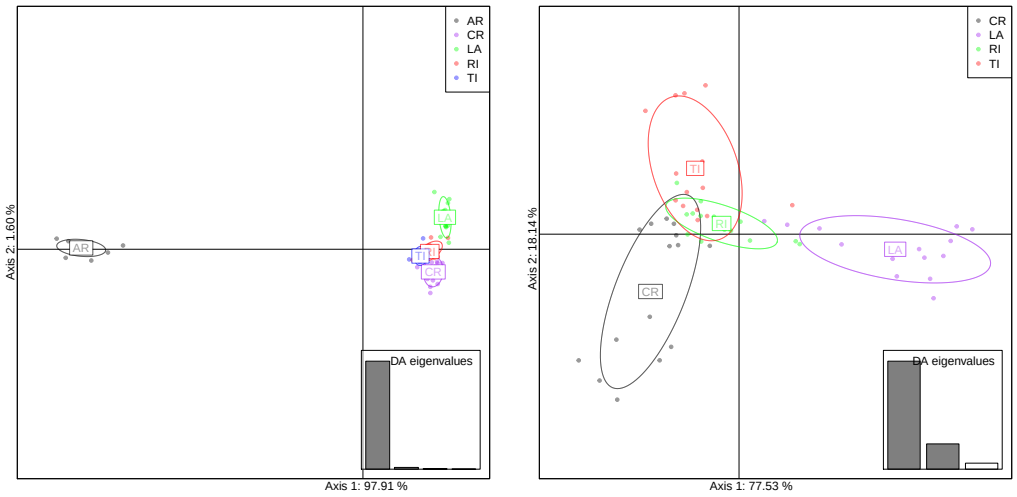
		Interior	Coastal		
		AR	CR	LA	RI
Coastal	CR	0.01134 ±0.0419			
	LA	0.010455 ±0.0416	0.0021 ±0.0288		
	RI	0.010477 ±0.0409	0.00058 ±0.0258	0.001 ±0.027	
	TI	0.00982 ±0.0403	0.00126 ±0.0271	0.00107 ±0.0279	0.00024 ±0.0256

To further characterize the population structure within coastal provenances, we performed DAPC and PCoA with only the four coastal provenances (Figure 3.4b and Supplementary Figure B.6b). In the DAPC analysis, LA showed some degree of differentiation from the other coastal populations, whereas in the PCoA, all provenances were clustered around zero. A high level of admixture between all coastal provenances for all values of K was observed with ADMIXTURE (Supplementary Figure B.9).

In summary, all methods indicated a clear genetic differentiation between coastal and interior provenances and little differentiation between coastal provenances either due to a high level of gene flow or other types of admixture.

Patterns of genetic diversity

To investigate patterns of polymorphisms among provenances, we first tested each SNP for deviation from Hardy-Weinberg equilibrium (HWE) using an exact test. A SNP was considered to significantly deviate from HWE if the p value was $\leq 10^{-5}$, a cutoff value frequently used in genome-wide associations for HWE testing (Pare, 2010). Accordingly, a total of 9,340 SNPs (11.7 %) deviated from HWE. The mean nucleotide diversity π per SNP for all individuals was 0.0032 (s.d. = 0.0033), and the average values for each provenance were lower but very



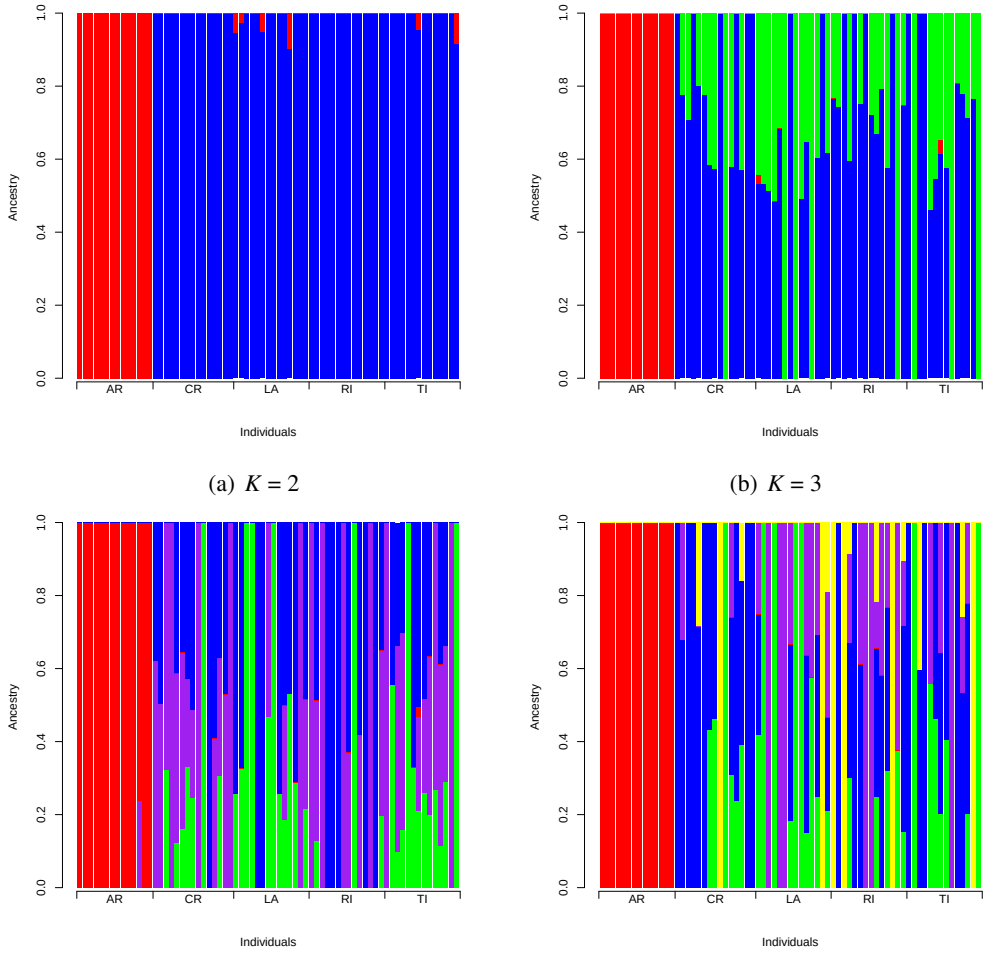
(a) (b)

Figure 3.4.: (a) DAPC of 71 interior and coastal trees. (b) DAPC of 57 coastal trees. AR - Salmon Arms, CR - Conrad Creek, LA - Cameron Lake, RI - Santiam River, TI - Timber.

similar to each other (Table 3.4; Supplementary Figure B.10). A comparison of π values of known genes with the corresponding PUTs revealed similar to slightly lower π values in our data (Supplementary Info. B). The mean Tajima's D value over all PUTs and provenances was 0.43 (s.d. = 0.78) indicating a minor excess of high-frequency SNPs. In contrast, single provenances showed mean Tajima's D values between 0.04 and 0.06 (Table 3.4, Supplementary Figure B.11), which was close to the neutral expectation. In summary, the five provenances exhibited a low level of intrapopulation genetic diversity, but also the expected signature of outcrossing, essentially neutrally evolving populations.

Table 3.4.: Mean values of nucleotide diversity, π, and Tajima's D over all provenances and within each provenance.

	All	AR	CR	LA	RI	TI
π	0.0032 $\pm$ 0.0033	0.0022 $\pm$ 0.002	0.0022 $\pm$ 0.0019	0.0023 $\pm$ 0.002	0.002 $\pm$ 0.0018	0.0022 $\pm$ 0.0019
Tajima's D	0.425 $\pm$ 0.782	0.059 $\pm$ 0.71	0.063 $\pm$ 0.721	0.06 $\pm$ 0.711	0.044 $\pm$ 0.716	0.051 $\pm$ 0.714



(a) $K = 2$ (b) $K = 3$

(c) $K = 4$ (d) $K = 5$

Figure 3.5.: Results of ADMIXTURE runs with $K = 2$ to $K = 5$. For all values of K the interior provenance is clearly separated from the coastal provenances. A strong admixture of coastal trees is seen for $K > 3$. Each individual is represented by a stacked column of ancestry percentages. AR - Salmon Arms, CR - Conrad Creek, LA - Cameron Lake, RI - Santiam River, TI - Timber.

Identification of highly differentiated outlier genes

We identified strongly differentiated genes based on the 2- and 5-populations models using three different outlier detection tests. The test results are summarized in Table 3.5 and the significant SNPs are listed in Online resources 3 and 4. More SNPs were identified under the 2- than the 5-populations model with both LOSITAN and BayeScan. There was a large overlap between LOSITAN, BayeScan, and Bayenv2 predictions, because 67 % (2-populations model) and 91 % (5-populations model) of the significant SNPs identified in both LOSITAN and BayeScan were

among the top 100 SNPs (based on a ranking by X^TX values) identified by Bayenv2. This is reflected by a highly significant correlation between F_{ST} and the X^TX values under both models (2-populations: Spearman rank $\rho = 0.38$; p value $< 2.2 \times 10^{-16}$; 5-populations: $\rho = 0.54$, p-value $< 2.2 \times 10^{-16}$). The sequences of PUTs with significant SNPs in both LOSITAN and BayeScan were searched against the NCBI *nr* database with an e-value cut-off of 10^{-6}. Fifty-two of 67 PUTs (78 %) under the 2-populations and 63 of 87 PUTs (72 %) under the 5-populations model showed significant hits and are listed in Online resources 5 and 6. The gene annotations indicate a broad functional diversity, including highly conserved genes and rapidly evolving disease resistance genes.

Highly differentiated genes can be viewed as candidate target genes for local adaptation to different environments. To further test for non-neutral evolution, we determined the deviation of observed Tajima's D values under a neutral 2-populations model. No PUT rejected the null hypothesis of neutral evolution with a FDR of 0.1 or a family-wise error rate of 0.1. We also did not find any significant deviations when the coastal population and the interior population were treated as single populations.

Table 3.5.: Summary of outlier tests for strongly differentiated SNPs. L - LOSITAN, B - BayeScan.

Outlier test	2-Populations model		5-Populations model	
	SNPs	PUTs	SNPs	PUTs
L	947	861	234	202
B	153	133	78	69
L ∩ B	97	87	75	67
L ∩ B ∩ top 100 X^TX	65	58	68	62
L ∩ B ∩ top 200 X^TX	93	84	71	64

Genetic diversity and differentiation of candidate drought-related genes

Among *a priori* defined 131 candidate genes with a putative function in drought response, only 17 were polymorphic for a total of 58 SNPs. SNP-wise F_{ST} values of this set were not significantly different from the remaining genome-wide SNPs (Wilcoxon test, p = 0.1335). Similarly, X^TX values did not differ from the genome-wide SNPs under the 2-populations (Wilcoxon test; p = 0.87) and 5-populations models (p = 0.42). Only one of 58 SNPs was among the top 100 X^TX values of the genome-wide distribution of SNPs under the 2-populations model, and no

SNP under the 5-populations model. On the other hand, drought-related genes appear more strongly differentiated on a gene-wise level because average F_{ST} values of the 17 PUTs with SNPs (mean $F_{ST} = 0.022$) were significantly larger than the average of the remaining 14,979 PUTs with SNPs (mean $F_{ST} = 0.0065$; Wilcoxon test: $p = 0.0062$). A permutation test that 10,000 times randomly selected 17 PUTs with the same number of SNPs and comparable heterozygosity as the drought related PUTs and a subsequent Wilcoxon ranksum test confirmed this result. Only 85 tests resulted in p values <0.006, which supported the hypothesis that F_{ST} values of drought-related PUTs are significantly larger than average.

Comparison of neutral demographic models

We used ABC to fit two neutral models of two populations connected by bi-directional asymmetric migration to the observed data (Figure 3.2, Table 3.1). The *I* model showed a marginally better fit than the *PS* model (Bayes factor = 2.73). The fitted *PS* model is very similar to the *I* model because of high divergence times and migration rates, and therefore both models show very similar posterior parameter distributions (Figure 3.6). Model comparison with even larger parameter ranges did not improve the Bayes factor between both models (Supplementary Info. B).

Since the total population size in the *I* model is $2N_e$ individuals, the scaled mutation and recombination rates have to be doubled for comparison with empirical estimates. The doubled mean of the scaled mutation rate in the fitted *I* model is 0.002, which is lower than the estimated rate of 0.006 from Jaramillo-Correa et al (2010), but agrees better with our empirical estimates of nucleotide diversity.

However, the best-fitting *I* model did not reproduce all summary statistics observed in the 1,000 randomly selected PUTs in the posterior predictive check analysis (Figure 3.7). Observed and simulated variances of the measured diversity statistics were similar under the *I* model, but mean values were quite different. Removing all rare polymorphisms with absolute frequency 1 and/or 2 (models I_1 and $I_{\leq 2}$) substantially increased the fit to the data compared to all other models. In a ABC model comparison of all models, only simulations from models I_1 and $I_{\leq 2}$ were kept, but $I_{\leq 2}$ fitted the data considerably better (Bayes factor = 93.09) although posterior predictive checks revealed that omitting these polymorphisms from the *I* model still does not reproduce the different summary statistics perfectly (Figure 3.7).

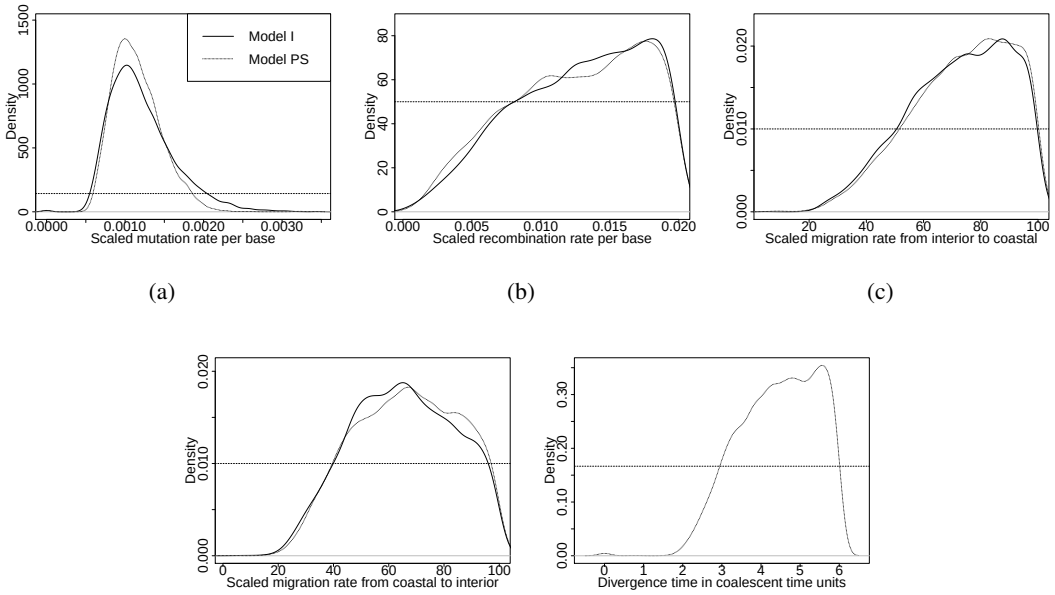


Figure 3.6.: Posterior distributions of the (a) mutation rate θ, (b) recombination rate ρ, (c) migration rate from interior to coastal population, (d) migration rate from coastal to interior population, and (e) divergence time. Straight lines show posteriors for model *I*, dotted lines posteriors for model *PS*. Prior distributions are added as dashed horizontal lines. All parameters are scaled by $4N_e$. N_e is the population size of the underlying neutral Wright-Fisher model.

3.5. Discussion

Quality of targeted sequence capturing

A proportion of essentially 100 % of captured target sequences and 97 % of target nucleotides covered at least once by sequence data indicate that targeted sequence capture is a useful approach for re-sequencing of the large and repetitive Douglas-fir genome. In the total sample, 90 % of target nucleotides were covered at least 10×. For comparison, a recent sequence capture study in barley with a comparable target region size (62 million bases) reported that 79 to 93 % of the target bases in multiplexed samples showed at least a 10× coverage (Mascher et al, 2013). In barley, on average, 50 % of reads per library could be mapped to the reference, whereas only 35 % of reads of our libraries could be mapped, which can be explained by the lack of a genome reference sequence in Douglas-fir. For the array design, we were restricted to PUTs as references that mainly cover exonic regions because they were derived from mRNA.

Captured genomic fragments may contain intronic regions if the probe was located next to the start or end of a gene or exon-intron boundaries, but reads consisting of intronic regions can not be mapped to reference PUTs. As a consequence, the proportion of unmapped reads is higher than in other species for which a whole genome reference sequence is already available. The restriction to coding regions may bias some results like genome-wide estimates of nucleotide diversity or selection acting on promoter regions or other non-coding regions, which cannot be investigated with our data. Alternative approaches like RADseq sample the genome more randomly including non-coding regions, but this method is also biased as it tends to underestimate nucleotide diversity (Arnold et al, 2013).

Population structure

The Salmon Arms (AR) provenance originated from a region where the coastal and interior varieties may interbreed (Kohnle et al, 2012), which is supported by low F_{ST} values between AR and the four coastal provenances. Nevertheless, the data indicate a clear genetic differentiation between AR and the coastal provenances, and very little differentiation within coastal provenances. The low levels of F_{ST} among pairs of coastal provenances suggest a nearly complete panmixis or high migration rates. These results confirm earlier studies reporting a low differentiation among coastal populations (Aagaard et al, 1998, Eckert et al, 2009d, Krutovsky et al, 2009, Viard et al, 2001), which were based on smaller marker numbers than our study. The DAPC analysis, however, revealed a subtle genetic differentiation among coastal provenances, in contrast to PCoA and ADMIXTURE. The inconsistent results for coastal provenances may have several explanations. First, the absence of a population structure in coastal provenances may result from an insufficient number of polymorphisms used for the analysis. Our analysis was based on 79,910 SNPs, but the ADMIXTURE manual recommends at least 100,000 polymorphisms for populations with F_{ST} values <0.01 (Alexander et al, 2013). Second, the genetic differentiation in the DAPC analysis may result from data over-fitting, which results in nearly perfect matches to any cluster. We tried to avoid over-fitting by applying the broken-stick method to retain the optimal number of PCs for the discriminant analysis (Jackson, 1993). Third, substantially fewer private SNPs were observed in coastal provenances compared to the interior provenance, which also reduced the power to differentiate among coastal provenances. The low number of private SNPs may result from a higher level of gene flow between coastal provenances than between

coastal and interior provenances. The latter hypothesis is consistent with a study of chloroplast and mitochondrial DNA that observed limited gene flow between coastal and interior varieties (Wei et al, 2011) and another study that identified strong gene flow in coastal provenances based on allozyme and microsatellite markers (Krutovsky et al, 2009). Since strong gene flow counteracts genetic differentiation, footprints of (potentially adaptive) genetic differentiation among coastal provenances may be weak. Unfortunately, our sample was unbalanced with respect to the number of interior and coastal provenances, and we could not test whether a similar pattern is present in interior provenances.

Nucleotide diversity

In comparison to other plant species, conifers have a low nucleotide diversity despite the large census population sizes of major conifer species (Heuertz et al, 2006, Mosca et al, 2012). Estimates of nucleotide diversity per site, π, in four conifers ranged from 0.0013 in *Pinus cembra* to 0.0081 in *Pinus mugo* (Mosca et al, 2012). For Douglas-fir, π in coding regions of 18 Douglas-fir genes was estimated as 0.00456 (Krutovsky and Neale, 2005). Another study estimated π as 0.0076 for synonymous and 0.002 for non-synonymous sites in 121 genes (Eckert et al, 2009d). We estimated π as 0.0032 for five provenances and from 0.0020 to 0.0023 for individual provenances. Those values are at the low end of values observed in conifers, but are comparable to known values for Douglas-fir. Our sample originated from a small section of the total distribution range of Douglas-fir and the sampling of species-wide diversity may be incomplete. On the other hand, our sample covers a larger geographic range (North to South: 800 km, East to West: 450 km) than the sample of Krutovsky and Neale (2005) (North to South: 600 km, East to West: 250 km). Another explanation for the low diversity is the SNP calling procedure, which may have been too conservative because at least ten reads over the total sample were required to confirm the variant nucleotide. The proportion of rare and private SNPs was likely reduced, which lowered estimated values of nucleotide diversity. This issue is difficult to resolve because SNP calling in next generation sequencing data strongly depends on the algorithm and the parameters used, and represents a compromise between controlling for type I and type II errors.

Genetic differentiation as indicator of adaptive divergence

Previous studies reported only a small number of candidate genes for positive selection in Douglas-fir and other conifer species (Eckert et al, 2009d, Krutovsky and Neale, 2005, Palmé et al, 2008). Our results are consistent with these findings because the outlier tests identified only a small set of strongly differentiated genes. The 58 genes classified as outliers by all three methods (LOSITAN, BayeScan, and Bayenv2) can be considered as high-confidence candidate genes for positive selection because the proportion of false positives can be reduced if the output of several methods is combined (Lotterhos and Whitlock, 2014). Based on their annotation, a cold shock protein (PMUSI_28593) or a CC-NBS-LRR resistance-like protein (PMUSI_10962) are promising candidates for further analyses.

An important phenotypic difference between interior and coastal provenances is drought tolerance. Unexpectedly, none of the *a priori* defined drought-related genes included in the array design was among genes harboring highly differentiated SNPs in a comparison of the three outlier tests. On the other hand, the mean F_{ST} values of the polymorphic drought-related genes were significantly larger than the average suggesting that they evolve differently from the remaining genes despite their low level of polymorphism.

A high level of differentiation in individual genes does not necessarily imply that they were targets of positive selection. To further test for footprints of selective sweeps or genetic hitchhiking, we compared Tajima's D values of polymorphic PUTs with values expected under a standard neutral model, both within the coastal and interior population and within the whole population, but did not find any significant deviations after correction for multiple testing. The same result was found if the model additionally allowed for recombination. These results suggest that positive selection in response to adaptive differentiation does not contribute significantly to observed patterns of standing genetic variation in Douglas-fir for biological or technical reasons. Given the polygenic nature of many adaptive traits, multigenic selection on these is expected, which limits the power to detect selective sweeps in individual genes. The high migration rate between coastal and interior provenances in combination with recombination may lead to a rapid loss of selection signals over time and further reduce detection power. Furthermore, adaptation may result from changes in gene expression rather than protein sequence evolution (He et al, 2012). Since our capture array was based on transcriptome data and did not cover regulatory sequences, we could not investigate the genetic diversity of regulatory regions.

Comparison of observed diversity with neutral models

A further explanation for the absence of Tajima's *D* outliers is an inappropriate standard neutral model. Given the observed population structure in coastal and interior provenances, we first assessed whether a panmictic model is justified. Both the *I* and *PS* models showed a comparable fit to the data and high migration rates consistent with the overall low F_{ST} values. The posterior distribution of divergence time in the *PS* model was shifted towards high values (i.e., deep divergence time) which essentially reduces the *PS* model to the *I* model. Given the high posterior probabilities for high migration rates in the fitted *I* model, we conclude that the panmictic model will not bias the results considerably (Supplementary Info. B).

In the posterior predictive checks, however, the *I* and *PS* models showed a poor fit to the distribution of means of summary statistics (Figure 3.7), and they did not capture all relevant processes influencing genetic diversity. For example, the neutral model assumed constant per-site recombination and mutation rates to reduce the computing time of the ABC analysis. Although variable mutation and recombination rates cause more variation between genes and thus higher variances of summary statistics, constant rates likely do not strongly bias our results because the variation observed in the data was within the range of variation in the fitted model for all summary statistics.

A reduced model fit could also be caused by low-frequency SNPs. We observed fewer low-frequency SNPs than predicted by the neutral models, and models without low frequency SNPs (I_1 and $I_{\leq 2}$) fitted the observed data much better. Douglas-fir is an outcrossing species with a large genome size. Therefore, balancing or directed selection likely affects only few genes because linkage disequilibrium between genes decays rapidly because of recombination, and an ABC analysis based on means and variances should not be strongly influenced by these types of selection. In contrast, purifying selection, which was reported for coastal Douglas-fir (Eckert et al, 2009d) and other conifers (Grivet et al, 2011, Palmé et al, 2009), affects many genes. It causes reduced frequencies of deleterious alleles and leads to a higher proportion of low-frequency polymorphisms, which we did not observe. Possible explanations are a bias against rare polymorphisms in the SNP calling procedure or a demographic factor such as a recent bottleneck.

Since Tajima's *D* is strongly influenced by low-frequency polymorphisms (Achaz, 2008), their absence may explain the lack of significant outliers in our data despite the possible action of

positive selection. Other neutrality tests like Fay and Wu's H (Fay and Wu, 2000) are much less influenced by rare polymorphisms. Unfortunately, we could not use this statistic, due to the lack of an outgroup in our data set. Given the strong effect of low-frequency variants on demographic modeling, we considered F_{ST}-based and related outlier methods as more reliable approaches in the current study because they are less affected by low-frequency variants. However, these are not explicit tests of positive selection, and the adaptive value of identified candidate genes needs to be characterized by other approaches.

3.6. Conclusion

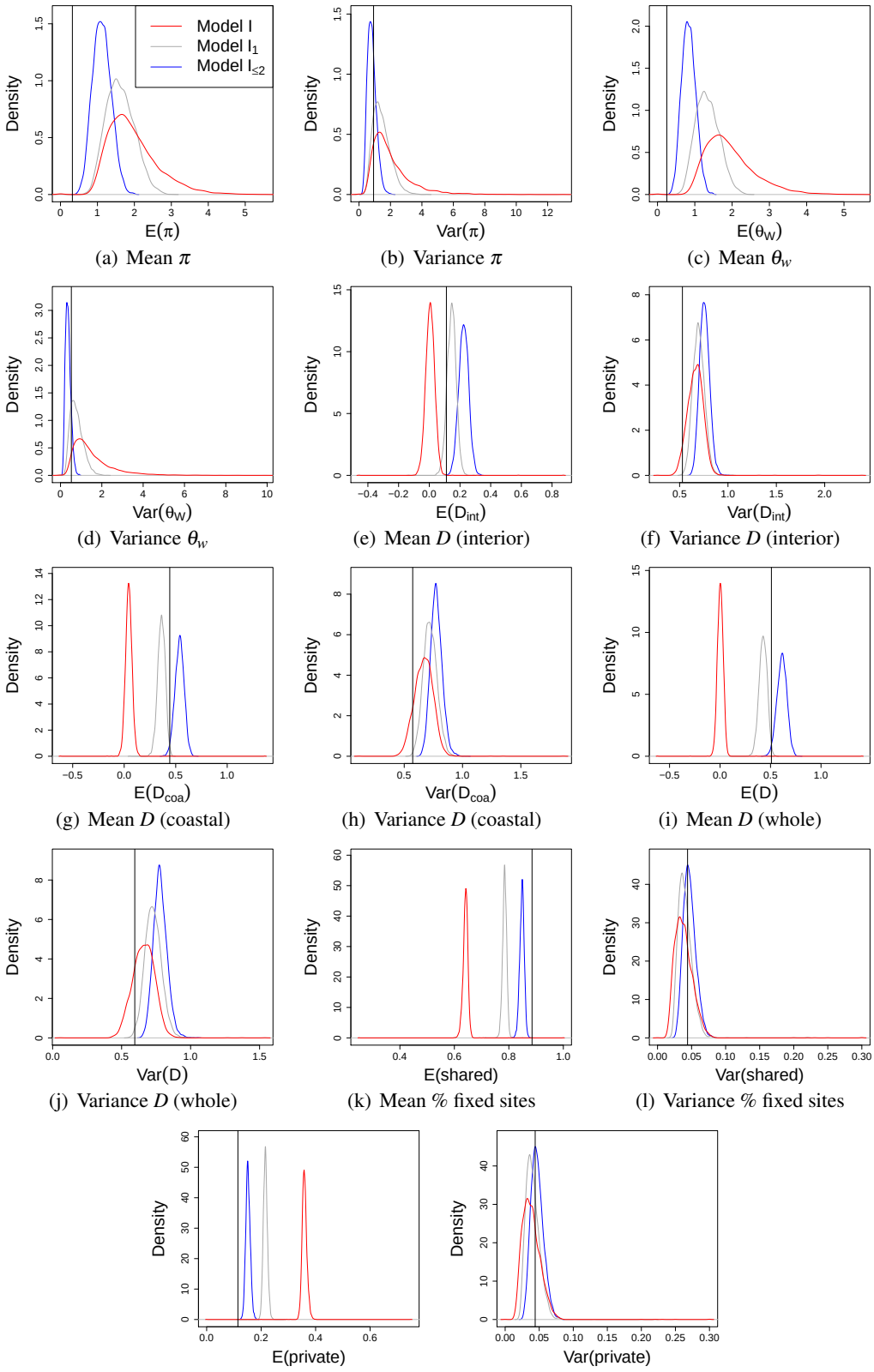
We demonstrated that sequence capture is a suitable method for analyzing genetic variation in conifers with large and complex genomes. A high coverage of targeted regions was achieved, and the population genetic analysis confirmed the differentiation between coastal and interior Douglas-fir provenances despite a significant level of gene flow. We also identified candidate genes with potential footprints of selection. The present study was based on a small number of provenances that do not represent the full spectrum of genetic diversity of Douglas-fir. Increasing the sample sizes and diversity of provenances to achieve a more balanced sampling design will probably improve the ability to detect genomic footprints of local adaptation. We also expect that sequence capture-based genotyping will contribute to mapping the genotype-phenotype relationship for important traits like drought tolerance or disease resistance, and to the genome-based selection of Douglas-fir provenances for tree breeding and other applications.

Acknowledgements

This study was funded by DFG grant SCHM1354/3-1 to KJS as part of the collaborative project 'DougAdapt' and by funding from the Forest Research Institute (FVA) Baden-Württemberg to HW. We thank Elisabeth Kokai-Kota and Anna-Maria Weißer for technical support in the library preparation. Christina Lanz and Detlef Weigel (Max-Planck-Institute of Developmental Biology, Tübingen) kindly let us use their DNA shearing facility for library construction.

Data archiving statement

A detailed description how to process the data and custom scripts used in this study are available at http://doi:10.5061/dryad.14ns8. Raw data is available at http://www.ebi.ac.uk/ena/data/view/PRJEB5165 with the accession number PRJEB5165.



(a) Mean π (b) Variance π (c) Mean θ_w

(d) Variance θ_w (e) Mean D (interior) (f) Variance D (interior)

(g) Mean D (coastal) (h) Variance D (coastal) (i) Mean D (whole)

(j) Variance D (whole) (k) Mean % fixed sites (l) Variance % fixed sites

(m) Mean % shared sites (n) Variance % fixed sites

Figure 3.7.: Posterior predictive checks of summary statistics for model I (red curve) and models I_1 and $I_{\leq 2}$ (grey and blue curve). Mean and variances of the used statistics are shown for 1,000 PUTs. The densities are estimated from values of simulations of the ABC fitted models, and the vertical lines show observed values. θ_w = Watterson's estimator, π = nucleotide diversity, D = Tajima's D.

4. Comparison of genotyping-by-sequencing and sequence capture for population structure inference in Douglas-fir

Thomas Müller, Elisabeth Kokai-Kota and Karl J. Schmid

[1]Department of Crop Biodiversity And Breeding Informatics, University of Hohenheim, Stuttgart, Germany

Submitted to Molecular Ecology Resources

4.1. Abstract

Genotyping-by-sequencing (GBS) is a cost-effective method to genotype species with large and complex genomes like Douglas-fir (*Pseudotsuga menziesii*). We performed a single- and a double-digest GBS using DNA of 96 Douglas-fir trees from nine coastal and interior provenances to compare single-digest GBS, double-digest GBS, and targeted sequence capture of putative unique transcripts (PUTs) derived from transcriptome sequencing. Polymorphisms were called by mapping reads against reference PUTs and by performing a *de novo* analysis with *Stacks*. With both, the reference- and the *de novo*-based SNP detection, more SNPs with less than 50% missing data and a higher mean coverage per SNP were found in double-digest than in single-digest GBS data. However, most SNPs without missing data were identified by sequencing sequence capture libraries, although at a much higher cost per individual. Population structure inference with methods like discriminant analysis of principal components (DAPC) or ADMIXTURE based on the single-digest GBS, double-digest GBS, and sequence capture data identified the same subgroups in the sample (coastal, northern, and southern interior), and were consistent with the geographic origin of the provenances. This result demonstrated that GBS is a useful method for population structure analysis.

Since double-digest GBS performed better than single-digest GBS with respect to the number of high-quality SNPs identified, it is the recommended approach for population structure analysis in non-model species with or without a sequenced reference genome.

4.2. Introduction

Most conifer species tend to have large genome sizes, and the re-sequencing of whole genomes on a routine basis for population genomic studies is not yet available. In an applied context, the long generation times of conifers require a reliable genotypic characterization of tree provenances for reforestation or breeding purposes. For these reasons, genotyping assays based on single nucleotide polymorphisms (SNPs) were developed for several conifers including Douglas-fir (Howe et al, 2013), maritime pine (Chancerel et al, 2011), sugar pine and loblolly pine (Eckert et al, 2009c, Jermstad et al, 2011). Methods to analyze restriction site associated DNA markers by sequencing (RADseq) such as genotyping-by-sequencing (GBS, Elshire et al, 2011) are cost-efficient alternatives to SNP arrays for genotyping large samples with thousands of markers. In GBS, the genome is fragmented by restriction enzymes, and the digested fragments are size selected after adapter ligation. Therefore, only a fraction of the genome is sequenced.

GBS was initially applied to maize and barley (Elshire et al, 2011) and subsequently in crops like potato (Uitdewilligen et al, 2013), switchgrass (Lu et al, 2013), or wheat (Poland et al, 2012a). The combination of a rare and a frequent cutting restriction enzyme in GBS (Poland et al, 2012b) results in fewer genome fragments that are sequenced at a higher average coverage, which leads to a lower proportion of missing data. GBS with two restriction enzymes has already been applied to genotype conifer species with large genomes like lodgepole pine (*Pinus contorta*) and white spruce (*Picea glauca;* Chen et al, 2013).

Although GBS and related RADseq methods currently experience a great interest in genotyping and population genomics (Narum et al, 2013), they also attracted some criticism because they contain a lot of missing data and are biased in population genetic parameter estimation (Arnold et al, 2013, Gautier et al, 2013). The assumed advantages of GBS motivated us despite the biases to apply GBS to Douglas-fir populations and to compare the results of GBS with targeted sequence capture for inferring the population structure and estimate descriptive population parameters.

Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) is one of the most important tree species world wide with a large economic value due to its wood quality and growth parameters. The natural habitat of Douglas-fir extends over several thousand kilometers in North America where two main varieties are present. Coastal Douglas-fir (*P. menziesii* var. *menziesii*) populates coastal areas from British Columbia to central California and interior Douglas-fir (*P. menziesii* var. *glauca*) occurs along the Rocky mountains from British Columbia to Mexico. The interior Douglas-fir was previously divided into a northern and a southern subgroup based on allozyme and RAPD markers (Aagaard et al, 1998, Li and Adams, 1989). In British Columbia, a *menziesii–glauca* transition zone was described with some reciprocal introgression between the varieties (Eckert et al, 2009a, Kohnle et al, 2012). As a consequence, coastal populations are expected to be more similar to northern interior than to southern interior populations (Aagaard et al, 1998, Li and Adams, 1989).

Since the genome size of Douglas-firs is approximately 19 Gbp (Ahuja and Neale, 2005), whole genome sequencing is still prohibitively expensive for population genomic analysis. Cheaper alternatives are RNASeq (Howe et al, 2013, Müller et al, 2012) and sequence capture approaches (Müller et al, 2015a) which were already applied to small numbers of individuals in Douglas-fir. Nevertheless, these methods are still very expensive for genotyping large population samples compared to GBS.

We performed a single digest (SD, one restriction enzyme) and a double digest (DD, two restriction enzymes) GBS of DNA from 96 Douglas-fir trees from four coastal, three northern interior, and two southern interior populations. Five of these provenances were previously characterized by targeted sequence capture (Müller et al, 2015a). Therefore, we restricted the comparison between GBS and sequence capture to these five provenances. We identified SNPs by mapping the reads against a reference consisting of putative unique transcripts (PUTs) from transcriptome sequencing, and by a *de novo* assembly of the reads with the *Stacks* package (Catchen et al, 2013b, 2011). Although we obtained many more SNPs without missing data in the targeted sequence capture, SD-GBS and DD-GBS gave similar results in the inference of population structure and differentiated between coastal, northern interior, and southern interior populations. Since DD-GBS produced more SNPs, fewer missing data, and higher mean coverage per SNP than SD-GBS, we propose DD-GBS for large-scale, cost-efficient genotyping of Douglas-fir populations for population genomic analyses.

4.3. Material and methods

Plant material

DNA of several trees from five interior and four coastal provenances (Table 4.1, Figure 4.1) was extracted as described in Müller et al (2015a) (for AR, CR, LA, RI, and TI) and in Müller et al (2012) (for BC1, BC6, CO, and NM).

Table 4.1.: Origin of trees.

Provenance	Origin	Experiment code	Variety	Number of trees
Conrad Creek	Washington, USA	CR	coastal	15
Cameron Lake	British Columbia, Canada	LA	coastal	14
Santiam River	Oregon, USA	RI	coastal	13
Timber	Oregon, USA	TI	coastal	13
Salmon Arms	British Columbia, Canada	AR	interior	13
Twin Lake	British Columbia, Canada	BC1	interior	8
Prince George	British Columbia, Canada	BC6	interior	7
Fort Collins	Colorado, USA	CO	interior	6
Raton	New Mexico, USA	NM	interior	6

Library preparation and sequencing

In genotyping-by-sequencing (GBS) one or more restriction enzymes (RE) are used to obtain reduced representation libraries which reduce the sequencing costs per individual. We performed two GBS experiments, one using ApeKI as RE (single digest, SD; Elshire et al, 2011) and a second using ApeKI and HindIII as REs simultaneously (double digest, DD; Poland et al, 2012b). ApeKI and HindIII are both type II REs cutting at defined positions within their recognition sites. ApeKI cuts the sequence GCWGC (W codes for A or T), if the 3'C nucleotides in both strands are unmethylated, leaving a 3 bp overhang (CWG). Since genic regions in higher plant genomes are generally hypomethylated, whereas repetitive regions are hypermethylated, the use of ApeKI should reduce the sequencing of repetitive regions (Nelson et al, 2008), although it is not known whether this is relevant for Douglas-fir. In contrast to ApeKI, HindIII is a rare cutter, which recognizes the sequence AAGCTT and cuts with a 4 bp overhang (AGCT).

Ninety-six barcodes (obtained from Elshire et al (2011) and the online service http://www.deenabio.com/services/gbs-adapters) of length 4 – 8 bp were generated (Supplementary Table

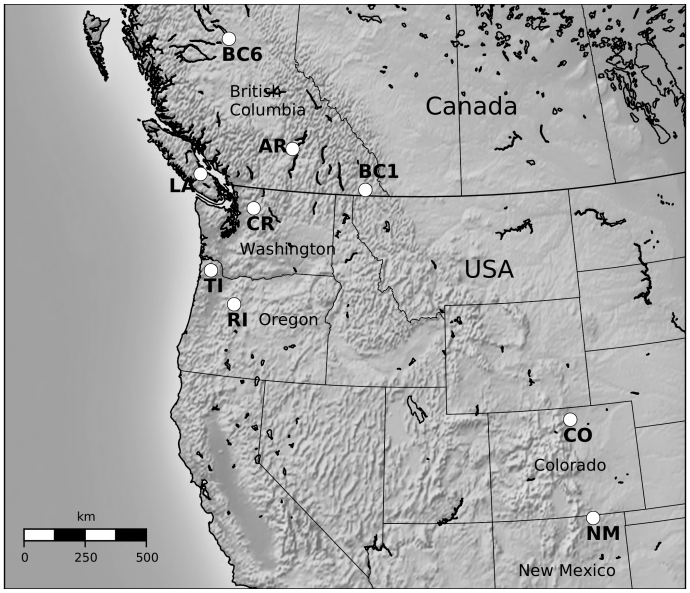


Figure 4.1.: Map of North America showing the origin of the provenances. AR - Salmon Arms, BC1 - Twin Lake, BC6 - Prince George, CO - Fort Collins, NM - Raton, CR - Conrad Creek, LA - Cameron Lake, RI - Santiam River, TI - Timber.

C.1). Complementary barcode oligonucleotides were synthesized for each barcode (forward: 5'–ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxx–3', reverse: 5'–CWGyyyyAG ATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT–3'; xxxx stands for a barcode of length 4 – 8, yyyy for the reverse complement of the barcode), diluted in 10 mM Tris 8.0 to 50 μM and annealed using 50 μl of each in a thermocycler (95°C, 2 min; ramp down to 25°C by 0.1°C/s; 25°C, 30 min; 4°C hold). RE specific common adapter oligonucleotides (ApeKI-forward: 5'–C WGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG–3', HindIII-forward: 5'–AGCT AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG–3', reverse: 5'–CTCGGCATTCCT GCTGAACCGCTCTTCCGATCT–3') were prepared in the same way as the barcode adapter. Barcode and common adapters were quantified using Qubit 2.0, Invitrogen, CA, USA, and diluted with water to a final concentration of 0.3 ng/μl. ddH_2O was added to genomic DNA to obtain concentrations of 600 ng / 51 μl. For SD-GBS, 17 μl of gDNA-ddH_2O mix (i.e., 200 *ng* of gDNA) of each tree were combined with 2 μl NEB Buffer 3 (10X) and 1 μl ApeKI (4U/μl) to a total volume of 20 μl and incubated at 75°C for two hours. For DD-GBS 16 μl of gDNA-ddH_2O mix (i.e., 188 *ng* of gDNA) of each tree was combined with 2 μl NEB Buffer 2 (10X), 1 μl ApeKI (4U/μl), and 1 μl HindIII (20U/μl) to a total volume of 20 μl and incubated at 37°C

for 2 hours followed by incubation at 75°C for 2 hours. The following steps were performed with both GBS libraries: For each sample 20 μl digested DNA, 10 μl common adapter (ApeKI for SD, HindIII for DD), 4 μl ddH_2O, 5 μl NEB Ligase Buffer(10X), and 1 μl NEB T4 DNA ligase (400 CELU/μl) were combined with 10 μl of one of the 96 barcode adapters, whereby each sample received a different barcode (but the same in both GBS approaches). After incubation at 22°C for 60 minutes, ligase was deactivated by incubating at 65°C for 30 minutes. Samples were then cooled to 4°C. For both GBS approaches, 5 μl of digested DNA-adapter fragments of each barcode were pooled together and cleaned with QIAquick PCR Purification Kit (Qiagen, Chatsworth, CA, USA) and eluted in 50 μl Qiagen elution buffer. Next, a PCR reaction for each of the two pooled samples was performed using 10 μl of pooled DNA, 25 μl NEB 2x Taq Master Mix, 2 μl PCR primer mix (forward: 5'–AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCCGATCT–3', reverse: 5'–CAAGCAGAAGACGG CATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT–3'; 25pmol/ μl each primer), and 13 μl dH_2O. The following PCR protocol was used: 5 minutes at 72°C; 30 seconds at 98°C; 18 cycles of: 10 seconds at 98°C, 30 seconds at 65°C, 30 seconds at 72°C; and a final extension for 5 minutes at 72°C. PCR reactions were cleaned up with QIAquick PCR Purification Kit and eluted in 30 μl elution buffer. Libraries were validated with Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany). All primers and oligonucleotides were ordered from Metabion, Steinkirchen, Germany.

Sequencing was performed on an Illumina HiSeq 1000 at the Kompetenzzentrum Fluoreszente Bioanalytik (KFB), Regensburg, Germany. Each library was sequenced on one lane, using paired-end sequencing with a length of 100 bp per read.

Data preparation

After removing PhiX reads, reads were demultiplexed according to their barcodes using a custom Python script. Due to low quality values at certain positions of the reads (assessed with FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), reads were preprocessed in the following manner. The first read of each pair (R1 reads) of the SD-GBS were cut at the end to receive a length of 85 bp. The second read of each pair (R2 reads) were cut to a total length of 80 bp, by trimming five bp at the start and the remaining bp at the end. Reads of the DD-GBS showed slightly better quality values. R1 and R2 reads of DD-GBS were cut to a length of 90

bp. Nucleotides were cut at the end in the case of R1 reads, while the first seven and the last three bp of R2 reads were cut. Reads were excluded if they contained at least one 'N' (undefined base) or if more than 10% of the bases had quality scores below 20 after trimming. Both reads of a pair had to pass the filtering steps to be included in the following analyses.

Comparison with SeqCap data

SeqCap data were obtained from a sequence capture study, in which targeted re-sequencing of 57,110 putative unique transcripts (PUTs) in 72 trees from five provenances was performed and sequenced on six lanes of a HiSeq1000 (Müller et al, 2015a). In contrast to this study, we also identified SNPs with missing data and we used SAMtools v.0.1.19 to search for SNPs (Li et al, 2009). By using a more recent SAMtools version and different filtering options, we found a slightly different number of SNPs without missing data in the SeqCap data than Müller et al (2015a) (75,347 vs. 79,910 SNPs).

Since we compared the results of the GBS approaches with data from a sequence capture study (Müller et al, 2015a), we mapped SD-GBS, DD-GBS, and SeqCap reads against the 57,110 PUTs that served as target regions in the sequence capture with a total length of 46 Mbp using pBWA (Peters et al, 2012) allowing for five mismatches per read. SNP-calling was performed with SAMtools v.0.1.19 (`samtools mpileup -D -g -C 50 -S -A -E`) followed by bcftools (`bcftools view -bvcg`) and vcfutils (`vcfutils.pl varFilter -D 30000`). Reads with a mapping score below 1 were excluded (`samtools view -q 1` Li et al, 2009).

A stringent SNP filtering was performed, using VCFtools v0.1.12a (Danecek et al, 2011), vcf-annotate, and custom python scripts because uninformative SNPs may bias the results (Roesti et al, 2012). In the first step, only sites with at least ten variant reads were kept. Second, genotypes of individuals with less than two reads were set to missing. Next, monomorphic sites consisting only of homozygous genotypes were removed. Finally, only SNPs with less than 50% missing data were kept. The final set of SNPs could include sites with less than ten variant reads, because it may be possible that some genotypes of a SNP were set to missing in the second step. Since the number of variant reads is given per site and not per individual in the .vcf file, it was not possible to determine the exact number of variant reads after step two.

SNP-calling was performed twice for each GBS approach: once with all samples (referred to as SDall and DDall) and once only with samples also present in the targeted sequence capture

study (AR, CR, LA, RI, and TI) (Müller et al, 2015a) with the exception of one AR, one RI, and one TI tree, for which not enough DNA material was available (referred to as SDcap and DDcap). The same SNP-calling approach was applied to the sequence capture data (referred to as SeqCap).

Data analysis

Pairwise F_{ST} values were calculated for each data set using VCFtools. PLINK v1.07 (Purcell et al, 2007) was used to convert .vcf files to .bed files. Population structure was analyzed with ADMIXTURE v.1.23 (Alexander et al, 2009), discriminant analysis of principal components (DAPC Jombart et al, 2010), and a principal coordinate analysis (PCoA) of pairwise F_{ST} values. ADMIXTURE, a Bayesian method, applies maximum likelihood estimation using models similar to models used by STRUCTURE (Pritchard et al, 2000). We used the R package *adegenet* to perform DAPCs with the SNP data and provenance information (Jombart, 2008, Jombart et al, 2010) to test whether the detected SNPs can be used to differentiate between the provenances (missing data points were not imputed). The number of principal components to keep in the principal component analysis step of the DAPC was determined with the `optim.a.score` function of the *adegenet* package. Furthermore, we retained three discriminant functions. PCoA and neighbor-joining trees were plotted using the R-package *ape* (Paradis et al, 2004). *adegenet* was further used to calculate the amount of missing data. For visualization we used the Python library *matplotlib* (Hunter, 2007).

De novo analysis using *Stacks*

We applied several components of the software pipeline *Stacks* v.0.20 to analyze the data *de novo*, i.e., without a reference sequence (Catchen et al, 2013b, 2011). *Stacks* was written for RE based data such as GBS to generate genetic maps and perform population genomic analysis. We applied *Stacks* only on R1 reads of SD and DD data, because paired end reads cannot be processed directly. Data preparation was performed as described above. In the first step of the pipeline, `ustacks` was run with both data sets (referred to as STACKS_SD and STACKS_DD), clustering all reads of an experiment into stacks, requiring a minimum coverage of three reads per stack. Stacks with a maximum distance of two were merged, and SNPs were identified with a maximum likelihood framework (Hohenlohe et al, 2010). Next, `cstacks` constructed a so

called catalog for the samples consisting of consensus loci of all stacks of all libraries. The allowed maximum distance between stacks combined to one locus was set to three. `sstacks` was used to detect polymorphisms of stacks found in `ustacks` against the catalogs of `cstacks`. To correct genotype calls of individual samples based on the given results, `rxstacks` was run, followed by a final run of `cstacks` and `sstacks`. Results of the pipeline were written in .vcf file format using the *Stacks* tool `populations`. The resulting .vcf files were analyzed as described above.

4.4. Results

More SNPs were identified with DD-GBS than SD-GBS

After quality filtering, 251 million and 249 million reads were retained with SD-GBS and DD-GBS (Figure 4.2 and Supplementary Figure C.1). Even though some individual trees were covered by a very low number of reads, we decided against removing those trees because we wanted to keep as many individuals as possible in the GBS data which were also included in the SeqCap data (Supplementary Figure C.2; four and one tree with less than 100,000 reads in SD and DD data set, respectively).

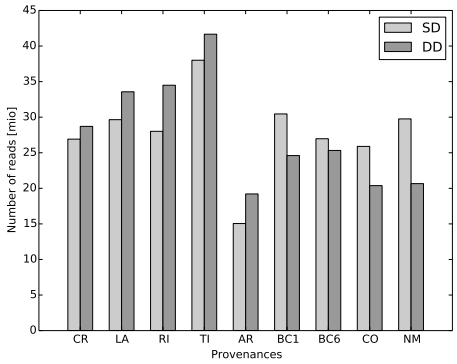
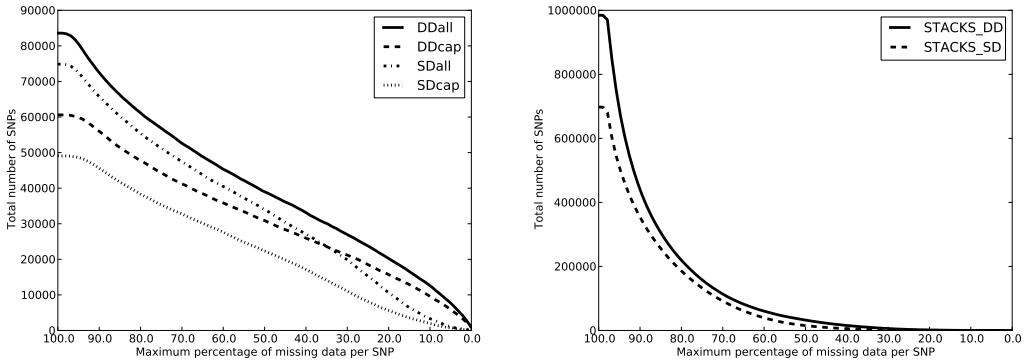


Figure 4.2.: Number of reads per provenance in SD- and DD-GBS after preprocessing.

The fraction of reads mapping against the PUT reference was similar for both approaches (SD: 11.4%, DD: 11.8%). We detected more SNPs in the DD than the SD data (Figure 4.3 (a), Table 4.2). Mean coverage was higher and percentage of missing data points was lower within DD data compared to SD data (Table 4.3, Figure 4.4). The effect of missing data per SNP on total

number of SNPs is shown in Figure 4.3. The lower the maximum percentage of missing data per SNP, the lower was the number of called SNPs. Nevertheless, we found considerably more SNPs without missing data in DD approaches (Table 4.2). To improve the quality of the data and to remove SNPs which are probably uninformative due to many missing data, we considered only SNPs with <50% missing data in the following analyses. Most SNPs (with or without missing data) and the lowest proportion of missing data were present in the SeqCap data, because the reads of that experiment mapped to all reference PUTs and not only to a small fraction like the GBS reads did (Table 4.2, Figure 4.4).

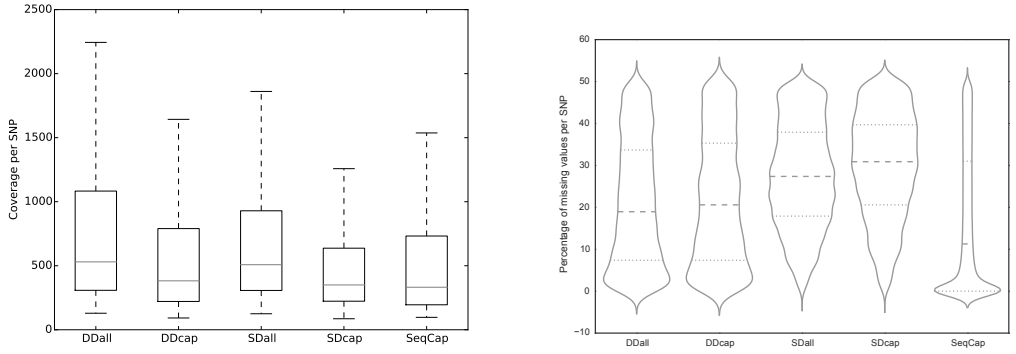


(a) (b)

Figure 4.3.: Total number of SNPs depending on maximum percentage of missing data per SNP. a) SNPs called after mapping the reads against the reference PUTs, b) *de novo* approach using *Stacks*.

Table 4.2.: Number of SNPs detected with different data sets and filtering criteria. SeqCap: data from a sequence capture experiment; SDall, DDall, SDcap, DDcap: SD- and DD-GBS data processed with a set of PUTs as reference using all libraries (all) or the subset of five provenances present in SeqCap (cap).

Max. percentage of missing data per SNP	Number of SNPs				
	SeqCap	SDall	DDall	SDcap	DDcap
100%	395,772	74,867	83,585	49,083	60,601
$\leq$ 50%	293,434	33,663	38,733	22,354	30,895
0%	75,347	21	896	21	934



(a) (b)

Figure 4.4.: (a) Read coverage per SNP and per data set. Outliers are not plotted. (b) Distribution of percentages of missing values per SNP per data set (plotted with python library *seaborn*). Dotted lines represent the quartiles. DD - double digest GBS, SD - single digest GBS, all - all individuals, cap - only individuals also present in the sequence capture were used, SeqCap - sequence capture.

Table 4.3.: Mean coverage per SNP across all individuals and mean percentage of missing data points per SNPs (sem = standard error of the mean). SeqCap: data from a sequence capture experiment; SDall, DDall, SDcap, DDcap: SD- and DD-GBS data processed with a set of PUTs as reference using all libraries (all) or the subset of five provenances present in SeqCap (cap).

	Mean coverage	sem	Mean percentage missing	sem
SeqCap	717.6	1.97	16.34	0.03
SDcap	653.4	6.91	29.90	0.08
DDcap	804.6	7.89	21.93	0.09
SDall	929.0	7.83	27.57	0.07
DDall	1097.1	9.60	21.16	0.08

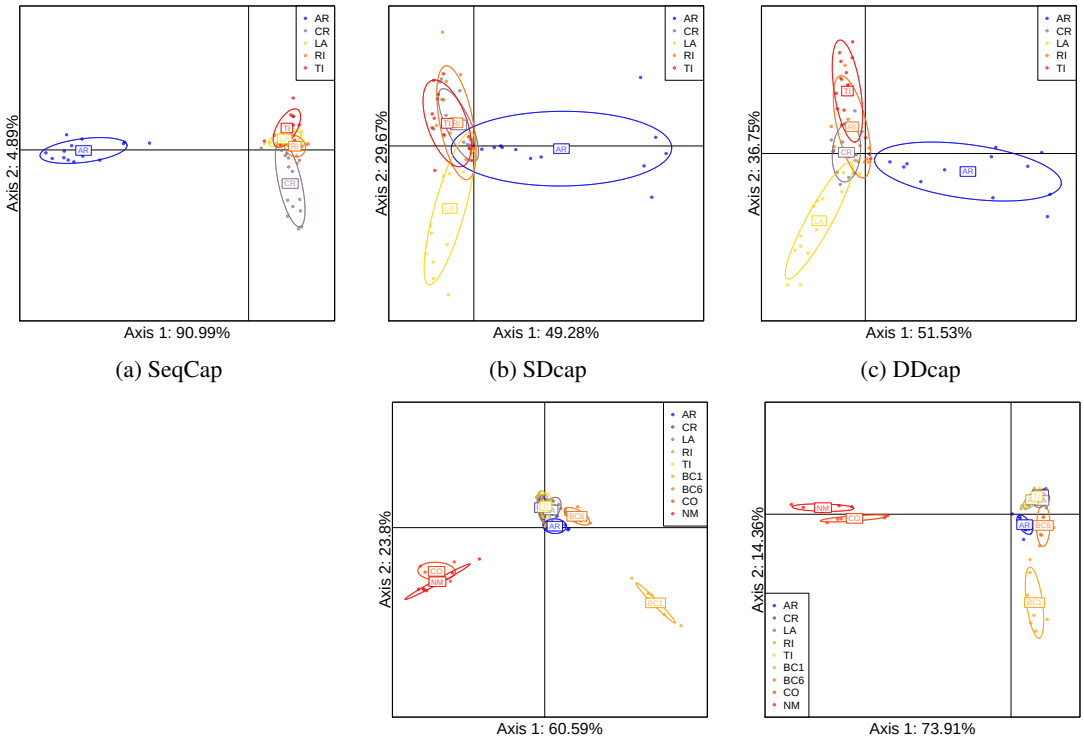
GBS approaches reveal population structure

We performed DAPC on SDcap and DDcap data and compared the results to the SeqCap data (Figure 4.5 (a) – (c)). The SeqCap data showed a clear separation between AR and the coastal provenances, which was not as clear in the SDcap and DDcap data, where the AR provenance is distributed mainly along the x-axis, with a small overlap to the coastal cluster in SDcap and no overlap in DDcap.

An ADMIXTURE analysis with $K = 2$ indicated admixture between coastal and interior provenances in the SDcap and DDcap data, while SeqCap data showed a clear separation between coastal and interior provenances (Figure 4.6 (a) – (c)). This may be caused by the signifi-

cantly lower number of SNPs in the GBS data because ADMIXTURE needs a large number of SNPs for structure detection if F_{ST} values are low, as is the case with our provenances.

We also conducted a DAPC analysis with the SDall and DDall data (Figure 4.5 (d) and (e)). In both GBS data sets the coastal provenances grouped together and were not separated. The northern interior AR and BC6 provenances located close to the coastal group and the remaining northern interior BC1 was located further away. The southern interior CO and NM were also grouped together. It seems that DDall result reflected the location of the provenances better than SDall, because BC1 provenance is located closer to the other two northern interior provenances. ADMIXTURE runs with the number of estimated populations K set to 3 confirmed these findings for SDall and DDall data, even grouping two BC6 individuals into the coastal cluster (Figure 4.6).

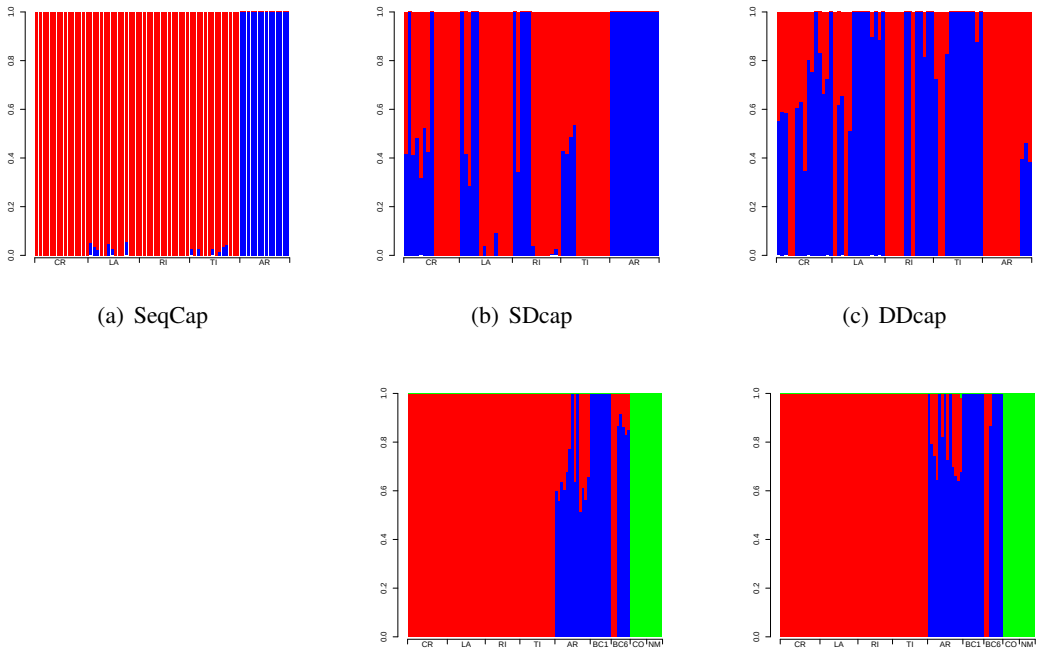


(a) SeqCap (b) SDcap (c) DDcap

(d) SDall (e) DDall

Figure 4.5.: Results of DAPC using SNP data from (a) SeqCap, (b) SDcap, (c) DDcap, (d) SDall, and (e) DDall.

Pairwise F_{ST} values for each data set calculated with VCFtools were used to perform a PCoA (Figure 4.7). Results of SeqCap, SDcap, and DDcap data were comparable to each other, sepa-



(a) SeqCap (b) SDcap (c) DDcap

(d) SDall (e) DDall

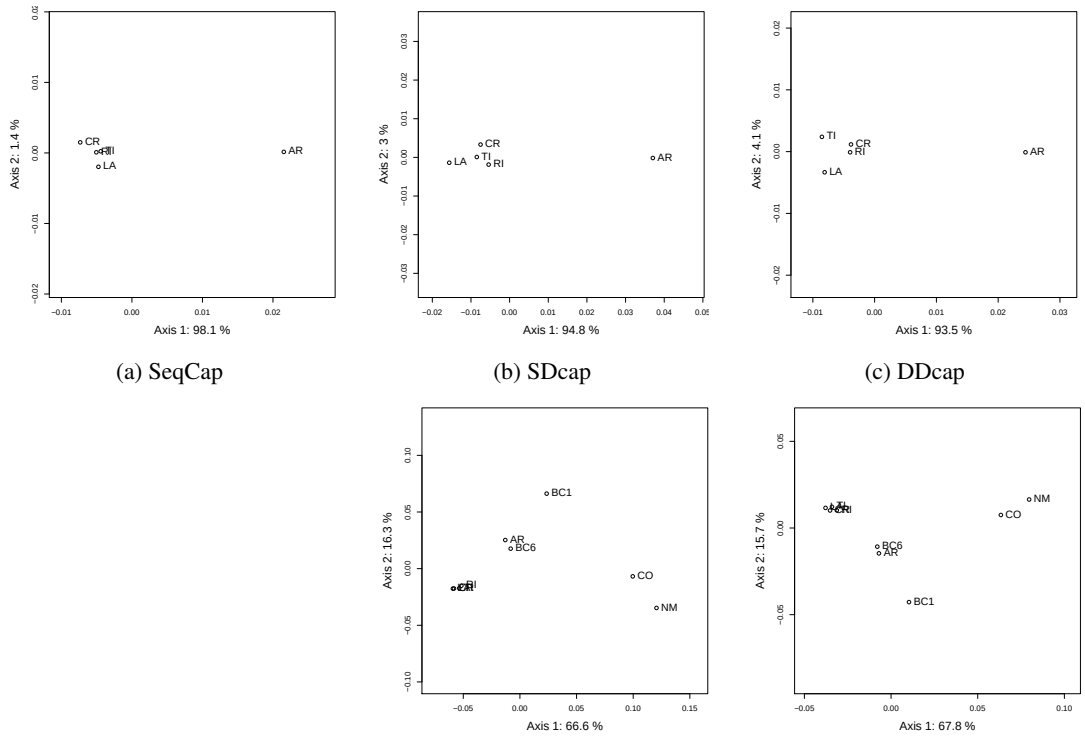
Figure 4.6.: Results of ADMIXTURE with $K = 2$ using SNP data from (a) SeqCap, (b) SDcap, and (c) DDcap, and results of ADMIXTURE with $K = 3$ using SNP data from (d) SDall and (e) DDall.

rating the interior from the coastal provenances. SDall and DDall results were similar to DAPC results of those data sets. AR and BC6 provenances were located between BC1 and the coastal provenances, while the southern interior provenances were again grouped together.

We constructed neighbor-joining trees of SDall and DDall based on the genetic distances (Supplementary Figure C.4). In both trees, there was a cluster of coastal as well as a cluster of southern interior provenances. Furthermore, BC1 and BC6 trees were located close to each other. AR trees were located between coastal and interior provenances.

De novo analysis with *Stacks*

A *de novo* analysis of SD- and DD-GBS data with *Stacks* detected more SNPs in DD than in SD data. The results of *Stacks* contained a large proportion of missing data within the SNPs (Figure 4.3 (b), Supplementary Table C.2), for instance, DD and SD data resulted in 984,472 and 697,616 SNPs with missing data, but only 31,333 and 14,542 SNPs remained, after excluding



(a) SeqCap (b) SDcap (c) DDcap

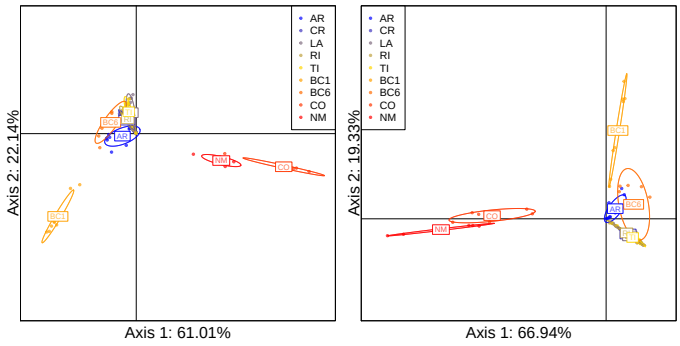
(d) SDall (e) DDall

Figure 4.7.: PCoA of pairwise F_{ST} values of each data set. (a) SeqCap, (b) SDcap, (c) DDcap, (d) SDall, (e) DDall.

SNPs with more than 50% missing values. A reason for the large proportion of missing values may be the libraries with extremely low coverage. To improve data quality we used in further analyses again only SNPs with $\leq$ 50% missing data over all individuals. As a consequence, all polymorphisms were removed in one AR individual (BA27) after this filtering step in the SD data, and it is missing in the analyses of SD data. Compared to the SDall and DDall analyses, for which reference sequences were used, fewer SNPs were detected with $\leq$ 50% missing data (SDall 33,663 vs. STACKS_SD 14,542; DDall 38,733 vs. STACKS_DD 31,333).

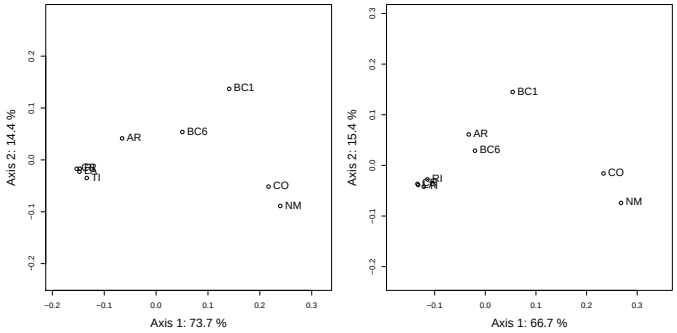
Nevertheless, DAPC and PCoA of pairwise F_{ST} values based on *Stacks* results were generally consistent with the results of the mapping approach (Figure 4.8, Figure 4.9).

Two neighbor-joining trees were build using STACKS_SD and STACKS_DD data (Supplementary Figure C.4). In both trees formed the coastal and the southern interior provenances a group. The southern interior group was close to the northern interior BC1 and BC6 provenances (with the exception of two BC6 individuals). Individuals of AR provenance were within the



(a) SD (b) DD

Figure 4.8.: Results of DAPC using data sets analyzed with *Stacks*. (a) SD and (b) DD.



(a) SD (b) DD

Figure 4.9.: PCoA of pairwise F_{ST} values using data sets analyzed with *Stacks*. (a) SD and (b) DD.

coastal and the northern interior group in the STACKS_SD data, but between the coastal and the two interior groups in STACKS_DD data. Therefore, it seems that the geographic distribution of provenances is better reflected in STACKS_DD result.

4.5. Discussion

We used single- and double-digest GBS to identify new SNPs in different provenances of Douglas-fir and to compare both GBS methods with a sequence capture approach (Müller et al, 2015a) in their ability to detect population structure. Because of a high level of gene flow within coastal and between coastal and northern interior Douglas-fir provenances (Krutovsky et al, 2009, Wei et al, 2011) the genotyping of many individuals may be required to achieve sufficient power for population structure inference (Fumagalli, 2013). We tested whether GBS is suitable for

population structure analysis of Douglas-fir and found that it is a suitable method for inferring population subdivision in Douglas-fir.

GBS analysis of the Douglas-fir genome

In this study we performed GBS using one restriction enzyme (ApeKI) and a combination of two restriction enzymes (ApeKI and HindIII). ApeKI cuts frequently, mainly in genic regions, because of its methylation sensitivity, while HindIII is a methylation insensitive rare cutter. Computer analysis of the putative unique transcripts (PUTs) suggested that both restriction enzymes are suitable for GBS in Douglas-fir, which was confirmed in test digests that produced sufficient amounts of DNA within the required length distribution.

As expected, the higher sequence coverage per fragment resulted in a higher total number of SNPs and a lower proportion of missing data in the double-digest compared to the single-digest GBS. Therefore, double-digest GBS seems to have advantages over single-digest GBS with respect to SNP calling in large and mostly unknown genomes. The ongoing sequencing of the Douglas-fir genome (Neale et al, 2013) will allow the utilization of other restriction enzymes that may, for example, preferentially target genic regions.

Despite the rapidly expanding use of GBS and other RADseq methods, many questions regarding the advantages and disadvantages of the RADseq library preparation method remain (Andrews et al, 2014, Puritz et al, 2014). Furthermore, RADseq and related methods have inherent biases in the estimation of different genetic diversity parameters like nucleotide diversity π (Arnold et al, 2013, Gautier et al, 2013). No strong consensus on best practices and optimal RADseq variants for certain scientific questions has emerged yet, but simulation studies or method comparisons in empirical studies can be used to evaluate GBS data (Arnold et al, 2013, Puritz et al, 2014). In our study, one concern were the highly variable number of reads obtained from different individuals (Supplementary Figure C.2) because some barcode adapters resulted in only a low number of reads. We observed the problem in other experiments as well which suggests that low quality of the barcode adapters and not of the genomic DNA explains this problem. We consider this to be a minor problem, because we were able to genotype a sufficient number of individuals per provenance to infer the population structure in our sample.

Another issue is the high proportion of missing data because it can not be determined whether missing data resulted from a random sampling effect or the loss of a restriction site. The pro-

portion of missing data can be reduced by replication, because a two or five fold replication of a sample can increase SNP coverage up to 60% or 90%, respectively (Poland and Rife, 2012). The random sampling of the genome with RADseq methods needs to be considered in the modeling of demographic history (Arnold et al, 2013). Targeted approaches such as sequence capture are advantageous to infer the demographic history, because sequenced regions tend to be longer and have a lower proportion of missing data. On the other hand, a focus on particular genomic regions in targeted sequence capture, like exons, may also bias genome-wide estimates of population parameters such as nucleotide diversity.

Detection of population structure with GBS

Our main goal was to evaluate GBS for population structure inference in Douglas-fir compared to targeted sequence capture. Only a small proportion of the available GBS reads were utilized in the reference-based SNP-calling, since only a fraction of the reads mapped to reference PUTs. This was expected, since the genome is digested randomly in GBS. Nevertheless, a differentiation in coastal and interior provenances was possible. But the differentiation was not as strong with the GBS than with the sequence capture data, especially with methods that are based on individuals like DAPC and ADMIXTURE. Individuals with a low coverage cannot be separated well and should be removed from analysis (Chen et al, 2013).

Since some barcodes produced low read counts in both GBS methods, DAPC results improved slightly after they were removed (Supplementary Figure C.3). PCoA results based on pairwise F_{ST} values, which rely on population measurements and not on single individuals, showed a clear separation of AR and the coastal provenances with SeqCap, SDcap, and DDcap data with all barcodes.

Compared to our previous sequence capture study, we included four more provenances in the present GBS analyses. Two provenances originated from the northern and two from the southern interior distribution range. The complete GBS data (SDall and DDall) showed a subdivision of interior Douglas-fir provenances into a northern and a southern group, which confirms previous studies based on RAPD and allozyme markers (Aagaard et al, 1998, Li and Adams, 1989), and demonstrates a high level of population differentiation in the interior relative to the coastal provenances.

Both GBS approaches produced similar results with the different inference methods tested.

ADMIXTURE and PCoA based on pairwise F_{ST} values separated coastal, northern interior, and southern interior provenances. Two interior BC6 trees were assigned to the coastal cluster, and the interior AR provenance showed some admixture with coastal provenances in the ADMIXTURE analyses. With DAPC, the northern interior provenances were separated, and AR and BC6 were more closely located to the coastal cluster, which may result from high gene flow between northern interior and coastal provenances (Krutovsky et al, 2009, Wei et al, 2011). The neighbor-joining trees of the genetic distances also showed some BC6 and AR individuals clustered with the coastal trees (Supplementary Figure C.4).

In summary, all population structure inference methods suggest that northern interior provenances are more related to coastal provenances than to southern interior provenances and that AR and BC6 may represent transition provenances. Both GBS methods detected this population structure although only coding sequences (PUTs) were used as reference and individuals with low read counts were not removed. The results are consistent with the origin of the AR and BC6 provenances from a transition zone, in which coastal and interior Douglas-fir may interbreed (Kohnle et al, 2012). However, no method differentiated between the coastal provenances based on the GBS data, but we obtained the same result with the sequence capture data that contained many more SNPs (Müller et al, 2015a). This may be explained with the high level of gene flow between coastal provenances (Krutovsky et al, 2009) or that a sample size of 13 to 15 individuals per provenance is too low for a correct inference of closely related populations (Fumagalli, 2013).

De novo analysis with *Stacks*

Stacks is a tool for the analysis of RADseq and GBS data and has been used for different species (e.g., Catchen et al, 2013a, Pujolar et al, 2014, Wu et al, 2013). We identified considerably more SNPs in *de novo* analysis with *Stacks* compared to the mapping against the PUT references. This was expected, because we could map only 12% of the sequence reads against the reference and use them for SNP calling, whereas in the *de novo* analysis, all reads were available for SNP detection. However, the proportion of missing values was very high among SNPs called by *Stacks*, probably because of low read counts in some individuals. Using the same threshold for missing data per SNP as in the reference-based approaches we called about 50% fewer SNPs with the single-digest GBS, and 20% fewer SNPs with the double-digest GBS than in the reference-based

SNP detection.

The inferred population structure was similar with the *de novo* analyses of single- and double-digest data in comparison to the reference-based analysis. The neighbor-joining tree constructed with the double-digest data reflects the geographical origin of Douglas-fir better than the single-digest data, since in the latter the AR individuals were located within the coastal cluster. This may reflect a high proportion of missing data due to the low read counts in the single-digest GBS data of some individuals. The neighbor-joining trees clustered the two BC6 individuals again within the coastal trees, as in the referenced based approaches (Supplementary Figure C.4). Multiple independent ADMIXTURE runs with SNP data obtained from single- and double-digest *de novo*-analyses gave inconsistent results, probably due to an insufficient number of SNPs as $\geq 100,000$ SNPs are recommended for populations with low F_{ST} values as in Douglas-fir (Alexander et al, 2009).

To summarize, the population structure inference based on *de novo*-analyses produced similar outputs than the reference-based analyses and suggests that a reference sequence is not required for a RADseq-based population structure analysis.

4.6. Conclusion

GBS (in particular double-digest GBS) is a cost-efficient method for investigating the population structure of Douglas-fir. Single- and double-digest GBS methods distinguished between coastal, northern interior, and southern interior provenances. The results were similar to those obtained with sequence capture data. Nevertheless, we were not able to detect a significant population differentiation among coastal provenances, probably due to small sample sizes. Double-digest is preferable to single-digest GBS, because a comparable number of reads produced more SNPs with a higher sequence coverage.

It is important to recognize the limitations of GBS, or RADseq methods in general, in comparison to sequence capture or whole-genome re-sequencing for characterizing natural populations. Nevertheless, in species with a large genome, a high level of gene flow, large population sizes, and obligate outcrossing such as Douglas-fir, GBS or other RADseq methods can be used as a component in a two-stage approach for genomic analysis. For example, GBS can identify a subset of individuals from a large sample for subsequent sequence capture or whole-genome

re-sequencing to investigate footprints of selection at high resolution. Furthermore, SNP markers identified with GBS can be used to develop SNP arrays with reduced ascertainment bias for genome-wide association studies or genomic selection in tree breeding.

Authors contributions

TM and KJS designed the study and wrote the manuscript. EKK performed lab-work. Data was analyzed by TM.

Acknowledgements

This work was funded by a grant of the DFG (SCHM1354-3) as part of the collaborative project 'DOUGADAPT - Adaptation of forest trees to climate change - Diversity of drought responses in Douglas-fir provenances' and an endowment of the Stifterverband für die Deutsche Wissenschaft.

Data Accessibility

Raw data is available at http://www.ebi.ac.uk/ena/data/view/PRJEB7733 and http://www.ebi.ac.uk/ena/data/view/PRJEB7737 with the accession numbers PRJEB7733 and PRJEB7737. Custom scripts and further data are available at http://datadryad.org (DOI YET UNKOWN).

5. General discussion

The main goals of this thesis were to identify a reference transcriptome of Douglas-fir, to analyze the genotypic variation in several provenances, and, if possible, to identify drought-related genes under selection. Furthermore, the population structure of several provenances was inferred. For these purposes, several next-generation sequencing experiments were performed.

5.1. Assembly of reference sequences and identification of drought candidate genes

Because no Douglas-fir reference genome was available, the first task was to establish reference sequences, which can then be used to detect polymorphisms in the genome of different individuals (Chapter 2, Müller et al, 2012). The genome of Douglas-fir with a size of approximately 19 Gbp is too large for cost-efficient whole genome sequencing (Ahuja and Neale, 2005). Transcriptome sequencing is a convenient alternative, in which the coding regions of a genome are sequenced, and which was already applied in several non-model organisms (Novaes et al, 2008, Parchman et al, 2010, Pauchet et al, 2009). With transcriptome sequencing we assembled 170,859 putative unique transcripts (PUTs) from twelve pooled cDNA libraries, which were constructed from wood and needle tissue of seedlings from coastal and interior provenances subjected to drought stress experiments. Because PUTs also contain alternatively spliced transcripts, the number of PUTs exceeds the number of 30,000 to 50,000 expected genes (Rigault et al, 2011). Another reason for the large number of PUTs is that genes are represented by several PUTs, due to incomplete assembly of the genes. This corresponds with the result that the average length of the PUTs was roughly half the expected average gene length in eukaryotes (Xu et al, 2006). Nevertheless, the comparison of functional annotations of the PUTs with *Arabidopsis thaliana* and *Picea sitchensis* transcriptome data showed similar results, suggesting that the

PUT set is a good representation of the Douglas-fir transcriptome.

With the PUTs as reference sequences we searched the data for genetic variation in the form of single nucleotide polymorphisms (SNPs). Therefore, we used three different SNP detection tools and reported large differences in their results. A number of 27,688 SNPs, however, were identified by all tools and were considered as highly reliable. Since at the time when the study was conducted only around 1,300 SNPs had been identified in Douglas-fir (Eckert et al, 2009b,d), our result enlarged the number of known SNPs considerably. As more studies have been performed in the meantime to identify SNPs in Douglas-fir, nowadays hundred-thousands of SNPs are known (Howe et al, 2013, Müller et al, 2015a).

Most of the highly reliable SNPs segregated in both coastal and interior provenances indicating a high level of gene flow or shared ancestral polymorphisms. We expected to see less shared polymorphisms between coastal and southern interior varieties. The large amount of shared polymorphisms was probably caused by the composition of the cDNA libraries, because the libraries consisted of pooled samples from several individuals and provenances. As a consequence of the pooling, the interior Douglas-fir cDNA libraries contained individuals from northern and southern interior provenances. Because northern interior individuals probably originated from the coastal-interior transition zone (the transition zone cannot be defined exactly), they are genetically more similar to coastal varieties than the southern interior individuals. This probably increased the amount of shared polymorphisms between both varieties (Kohnle et al, 2012). We also found a higher degree of genetic diversity within the coastal variety, which may be influenced by the composition of reads, because more reads of the coastal cDNA libraries were available than of interior cDNA libraries. If the composition of reads introduced a bias at all, it was not very strong, because the results were consistent with earlier studies of genetic diversity and gene flow in Douglas-fir (Aagaard et al, 1998, Li and Adams, 1989).

Drought tolerance is an important phenotypic difference between coastal and interior Douglas-fir provenances (Pharis and Ferrell, 1966). Using BLASTX keyword searches, we identified a set of 134 potentially drought related PUTs, which were present exclusively in drought stressed seedlings (Altschul et al, 1990). The small number of identified drought related PUTs may reflect that the response to drought is mainly driven by up- or down-regulation of genes, which are also expressed in control plants. Some of the PUTs could of course also result from a sampling artifact, since it is possible that PUTs were not found within not stressed seedlings by chance.

The 134 identified drought candidate PUTs were included in a subsequently performed sequence capture study.

5.2. Re-sequencing studies

5.2.1. Sequence capture

In targeted sequence capture, oligonucleotides complementary to target regions are used to capture only those regions from fragmented DNA (Grover et al, 2012). The target regions of our study consisted of 57,110 PUTs, including the 134 drought related PUTs identified before (Chapter 3, Müller et al, 2015a). The experiment was conducted with 72 trees from one northern interior and four coastal provenances, which were part of the international provenance trial in Germany (Kenk and Thren, 1984). Unfortunately, no southern interior provenance was available from the field trials.

Whole genome sequences are usually used to ensure the suitability of the designed oligonucleotides, for instance by testing if they match repetitive regions of the genome. Because no whole genome reference was available, the oligonucleotides were only tested within the Douglas-fir PUTs. Nevertheless, sequence capture was successful, since almost all target regions were captured in at least one individual. Since only coding regions where available as references, it was not possible to map reads to flanking non-coding regions like introns. Therefore, the total amount of mapped reads was lower than in comparable studies (average per library of 35% compared to an average of 50% in barley), but results were similar in terms of coverage per base (Mascher et al, 2013). SNP detection using one SNP calling method, but stringent filtering criteria, resulted in 79,910 SNPs without missing data.

Three methods based on allelic differentiation were used to identify positive selection within the PUTs. The 58 PUTs identified by all methods were considered as high-confidence candidates for positive selection. The small number is consistent with other studies, which also reported only small numbers of probably positive selected genes in conifer species (Eckert et al, 2009d, Krutovsky and Neale, 2005, Palmé et al, 2008). Seventeen of the drought related candidate PUTs were polymorphic, but none was among the candidates for positive selection. Nevertheless, average F_{ST} values, representing genetic differentiation, of drought related PUTs were significantly larger than the average values of the remaining PUTs, suggesting that drought re-

lated PUTs evolve differently. We further tested the polymorphic PUTs for non-neutral evolution by comparing their Tajima's *D* values with values expected under a standard neutral model, but no Tajima's *D* outliers were found. Because Tajima's *D* values are strongly influenced by low-frequency variants (Achaz, 2008), the absence of rare variants after the stringent SNP calling probably caused the lack of Tajima's *D* outliers. Further explanations for the lack of signals of selection could be the reduction of detection power due to multigenic selection or a rapid loss of selection signals caused by high migration rates. In addition, it was not possible to test if adaptation resulted from changes in gene expression (He et al, 2012), because the target regions did not cover regulatory regions.

Because the target regions of this sequence capture study consisted only of coding regions, selection acting on non-coding regions as well as the influence of non-coding regions on for example nucleotide diversity were missed. Nevertheless, the calculated mean nucleotide diversity in the sequence capture is within the range of mean nucleotide diversity values reported for Douglas-fir and other conifers (Eckert et al, 2009d, Krutovsky and Neale, 2005, Mosca et al, 2012).

5.2.2. Genotyping-by-sequencing

Genotyping-by-sequencing (GBS) is a cost-efficient and easy-to-multiplex method that can be used to genotype many individuals of species with a large and complex genome like Douglas-fir. The genome is fragmented using one or more restriction enzymes in GBS (Elshire et al, 2011, Poland et al, 2012b). The amount of fragments which are sequenced is reduced considerably compared to whole genome sequencing because eventually only fragments with a specific length are used for sequencing. We performed a single-digest and a double-digest GBS experiment and compared the results with each other and with the results of the sequence capture experiment (Chapter 4, Müller et al, 2015b). The GBS experiments were performed with material from the same trees as the sequence capture and, in addition, with material from 14 trees from two northern and two southern interior provenances from the transcriptome study (Chapter 2, Müller et al, 2012).

SNPs were identified by mapping the reads of both GBS approaches against the target regions of the sequence capture and by a *de novo* analysis of the reads. Both GBS approaches revealed a large number of polymorphisms. More SNPs with a higher mean coverage and less missing

data were detected by the double-digest than by the single-digest GBS in the reference-based as well as in the *de novo*-based approach.

Nevertheless, both GBS data sets contained a large amount of missing data in both the reference and the *de novo*-based approaches, probably because some libraries contained a low number of reads after sequencing. Since the barcodes of those libraries also yielded low numbers of reads in other experiments, we assume that a low quality of the barcode adapters and not of the used Douglas-fir DNA was the reason. Furthermore, it is impossible to determine whether missing reads from individuals are random results or due to a loss of restriction sites caused for example by a polymorphism in those sites. To circumvent at least the problem of random missing data, one or more replications of the experiments could be performed, which could increase the SNP coverage up to 60 to 90% and also decrease the amount of missing data considerably (Poland and Rife, 2012). Nevertheless, GBS data needs to be interpreted with caution, because GBS and related RADseq methods tend to have inherent biases for instance in the estimation of different genetic diversity parameters (Arnold et al, 2013, Gautier et al, 2013).

5.2.3. Population structure inference with sequence capture and GBS data

Population structure within sequence capture and GBS data was inferred through methods using population parameters like pairwise F_{ST} values and methods using SNP information like ADMIXTURE (Alexander et al, 2009). All applied methods were able to infer the expected coastal-interior structure of the five provenance of the sequence capture experiment. The separation of the five provenances in coastal and interior variety was not as strong in single- and double-digest GBS data applying methods relying on SNP data. The low coverage of some individuals resulted in a low number of detected polymorphisms, and as a consequence the individuals could not be separated properly. Nevertheless, methods based on population parameters showed a clear separation of the interior and the coastal provenances in both GBS data.

The population structure of five interior and four coastal provenances was inferred using the data obtained from reference- and *de novo*-based SNP calling of single- and double digest GBS data. All data sets and analyses showed a separation of coastal and interior variety and a further subdivision of the interior variety in northern and southern interior provenances. This confirms results of studies based on RAPD and allozyme markers (Aagaard et al, 1998, Li and Adams,

1989) and suggests a high level of population differentiation in the interior compared to the coastal provenances.

Two of the northern interior provenances originated from the transition zone, in which coastal and interior Douglas-fir may interbreed (Kohnle et al, 2012). This was reflected for instance by the DAPC results in which those two provenances were located closer to the coastal provenances than the third northern interior provenance.

None of the population structure analyses, neither with sequence capture nor with any GBS data set, differentiated the coastal provenances, probably due to the high gene flow among coastal provenances (Krutovsky et al, 2009). In the GBS data sets this result may be influenced by the large amount of missing data. Since it was not possible to infer structure in the coastal provenances in the sequence capture experiment, where a large number of SNPs without missing data were identified, it seems reasonable that the number of individuals per provenance was too low to detect genetic differentiation. In general, population structure can be inferred most accurate with large sample sizes (Fumagalli, 2013). Therefore, it is necessary to analyze more individuals of the coastal provenances to detect their population structure.

5.2.4. Sequence capture vs. GBS

Both methods are cost-efficient alternatives to whole genome sequencing to get sequence information and to genotype individuals. We were able to infer the expected population structure with both methods.

To be able to perform a targeted sequence capture, sequence information needs to be available in advance to define the target regions and to construct the complementary oligonucleotides. GBS samples the genome randomly with restriction enzymes, and no reference sequence is needed in advance. This was confirmed by our *de novo*-analyses of GBS reads, which returned similar results as a reference-based approach and a sequence capture experiment.

Sequence capture yields a large amount of SNPs with higher coverage and less missing data, but is more expensive than GBS. But the estimation of genetic diversity parameters can be biased with GBS data (Arnold et al, 2013, Gautier et al, 2013). Because in sequence capture longer contiguous fragments are obtained the estimation of those parameters may be less biased, but the selection of for example coding regions as targets may also introduce a bias. In modeling the demographic history, sequence capture with its non-random genome sampling is advantageous

to GBS methods, because the random sampling of the genome in the latter needs to be taken into account (Arnold et al, 2013). Even though the use of GBS and further RADseq methods is increasing, there is an ongoing debate on advantages and disadvantages of RADseq library preparation methods (Andrews et al, 2014, Puritz et al, 2014).

To summarize, the choice of experiment depends on the analyses which need to be performed. If reference sequence information is available and costs are negligible, targeted sequence capture is the best choice for most analyses. If GBS is performed and the genome is large, complex, and mostly unknown, double-digest should be favored over single-digest GBS. With double-digest GBS more SNPs with a higher coverage and less missing data can be identified, which improves downstream analyses.

5.3. Conclusion

Due to its ability to adapt to different climatic conditions and habitats, Douglas-fir is a highly interesting species for researchers, but a reference genome is still missing. We established the first set of putative unique transcripts representing the transcriptome of Douglas-fir, and detected a large number of single nucleotide polymorphisms. We identified drought related candidate genes, but we found no signs of positive selection amongst them. Nevertheless, a different set of candidate genes, which may be related to local adaptation, has been identified. Furthermore, we showed the suitability of sequence capture and genotyping-by-sequencing methods to infer population structure in species with complex, large, and mostly unknown genomes. The reference transcriptome, the identified SNPs, and the vast amount of phenotypic data collected in the "DougAdapt" project can be used for whole genome sequencing and tree breeding projects, and will be used in genomic selection and genome wide association studies to identify SNPs with associations to important phenotypes. The generated data represents a fundamental and enormous resource of information to further analyze the adaptability of Douglas-fir.

Bibliography

454 Life Science (2009) Genome Sequencer FLX System Software Manual, version 2.3

Aagaard JE, Krutovskii KV, Strauss SH (1998) RAPDs and allozymes exhibit similar levels of diversity and differentiation among populations and races of Douglas-fir. Heredity 81(1):69–78, DOI 10.1046/j.1365-2540.1998.00355.x

Aas G (2008) Die Douglasie (*Pseudotsuga menziesii*) in Nordamerika: Verbreitung, Variabilität und Ökologie. LWF - Wissen 59

Achaz G (2008) Testing for neutrality in samples with sequencing errors. Genetics 179(3):1409–1424, DOI 10.1534/genetics.109.104042

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 28(20):e87–e87, DOI 10.1093/nar/28.20.e87

Ahuja MR, Neale DB (2005) Evolution of genome size in conifers. Silvae Genet 54(3):126–137

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9):1655–64, DOI 10.1101/gr.094052.109

Alexander DH, Novembre J, Lange K (2013) ADMIXTURE 1.23 Software Manual

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–10, DOI 10.1016/S0022-2836(05)80360-2

Andrews KR, Hohenlohe PA, Miller MR, Hand BK, Seeb JE, Luikart G (2014) Trade-offs and utility of alternative RADseq methods: Reply to Puritz *et al*. Mol Ecol 23(24):5943–5946, DOI 10.1111/mec.12964

Andrews SF, Flanagan LB, Sharp EJ, Cai T (2012) Variation in water potential, hydraulic characteristics and water source use in montane Douglas-fir and lodgepole pine trees in southwestern Alberta and consequences for seasonal changes in photosynthetic capacity. Tree Physiol 32(2):146–160, DOI 10.1093/treephys/tpr136

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a F_{st}-outlier method. BMC Bioinform 9:323, DOI 10.1186/1471-2105-9-323

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol 22(11):3179–90, DOI 10.1111/mec.12276

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25(1):25–29, DOI 10.1038/75556

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst 41:379–406, DOI 10.1146/annurev-ecolsys-102209-144621

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B 263(1377):1619–1626, DOI 10.1098/rspb.1996.0237

Besag J, Clifford P (1991) Sequential Monte Carlo *p*-values. Biometrika 78(2):301–304, DOI 10.1093/biomet/78.2.301

Blanca J, Chevreux B (2012) sff_extract. URL http://bioinf.comav.upv.es/sff_extract/index.html

Bradshaw H, Stettler R (1993) Molecular genetics of growth and development in *Populus*. I. Triploidy in hybrid poplars. Theor Appl Genet 86(2-3):301–307, DOI 10.1007/BF00222092

Breese MR, Liu Y (2013) NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics 29(4):494–6, DOI 10.1093/bioinformatics/bts731

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. J Geophys Res 111(D12106), DOI 10.1029/2005JD006548

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268(1):78–94, DOI 10.1006/jmbi.1997.0951

Burgess D (2011) Multiplex sequence capture for targeted resequencing of candidate gene panels in cancer. Roche Nimblegen, Inc.

Campbell RK (1979) Genecology of Douglas-fir in a watershed in the Oregon cascades. Ecology 60(5):1036–1050, DOI 10.2307/1936871

Campbell RK, Sugano AI (1979) Genecology of bud-burst phenology in Douglas-fir: Response to flushing temperature and chilling. Bot Gaz 140(2):223–231, DOI 10.1086/337079

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43(10):956–963, DOI 10.1038/ng.911

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15(11):1553–65, DOI 10.1101/gr.4326505

Carter R, Klinka K (1990) Relationships between growing-season soil water-deficit, mineralizable soil nitrogen and site index of coastal Douglas fir. Forest Ecol Manag 30(1–4):301–311, DOI 10.1016/0378-1127(90)90144-Z

Catchen J, Bassham S, Wilson T, Currey M, O'Brien C, Yeates Q, Cresko WA (2013a) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. Mol Ecol 22(11):2864–83, DOI 10.1111/mec.12330

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013b) Stacks: an analysis tool set for population genomics. Mol Ecol 22(11):3124–40, DOI 10.1111/mec.12354

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3 (Bethesda) 1(3):171–82, DOI 10.1534/g3.111.000240

Chancerel E, Lepoittevin C, Provost GL, Lin YC, Jaramillo-Correa JP, Eckert AJ, Wegrzyn JL, Zelenika D, Boland A, Frigerio JM, Chaumeil P, Garnier-Géré P, Boury C, Grivet D, González-Martínez SC, Rouzé P, Peer YVd, Neale DB, Cervera MT, Kremer A, Plomion C (2011) Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. BMC Genomics 12(1):368, DOI 10.1186/1471-2164-12-368

Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Rep 11(2):113–116, DOI 10.1007/BF02670468

Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA (2013) Mining conifers' megagenome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genet Genomes 9(6):1537–1544, DOI 10.1007/s11295-013-0657-1

Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008:12, DOI 10.1155/2008/619832

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676, DOI 10.1093/bioinformatics/bti610

Coops NC, Coggins SB, Kurz WA (2007) Mapping the environmental limitations to growth of coastal Douglas-fir stands on Vancouver Island, British Columbia. Tree Physiol 27(6):805–815, DOI 10.1093/treephys/27.6.805

Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol 3(3):475–479, DOI 10.1111/j.2041-210x.2011.00179.x

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–8, DOI 10.1093/bioinformatics/btr330

Danecek P, Nellåker C, McIntyre RE, Buendia-Buendia JE, Bumpstead S, Ponting CP, Flint J, Durbin R, Keane TM, Adams DJ (2012) High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. Genome Biol 13(4):26, DOI 10.1186/gb-2012-13-4-r26

Darwin C (1859) On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life. J. Murray

Darychuk N, Hawkins B, Stoehr M (2012) Trade-offs between growth and cold and drought hardiness in submaritime Douglas-fir. Can J For Res 42(8):1530–1541, DOI 10.1139/x2012-092

Dean CA (2007) Genotype and population performances and their interactions for growth of coastal Douglas-fir in western Washington. For Sci 53(4):463–472

Dlugosch KM, Lai Z, Bonin A, Hierro J, Rieseberg LH (2013) Allele identification for transcriptome-based population genomics in the invasive plant *Centaurea solstitialis*. G3 (Bethesda) 3(2):359–367, DOI 10.1534/g3.112.003871

Douglas D (1914) Journal Kept by David Douglas During His Travels in North America 1823-1827. Together with a Particular Description of Thirty-Three Species of American Oaks and Eighteen Species of Pinus. London: William Wesley & Son

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci USA 100(15):8817–8822, DOI 10.1073/pnas.1133470100

Du B, Jansen K, Junker LV, Eiblmeier M, Kreuzwieser J, Gessler A, Ensminger I, Rennenberg H (2014) Elevated temperature differently affects foliar nitrogen partitioning in seedlings of diverse Douglas fir provenances. Tree Physiol 34(10):1090–1101, DOI 10.1093/treephys/tpu074

Ducić T, Parladé J, Polle A (2008) The influence of the ectomycorrhizal fungus *Rhizopogon subareolatus* on growth and nutrient element localisation in two varieties of Douglas fir (*Pseudotsuga menziesii* var. *menziesii* and var. *glauca*) in response to manganese stress. Mycorrhiza 18(5):227–39, DOI 10.1007/s00572-008-0174-5

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, St Clair JB, Neale DB (2009a) Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. Genetics 182(4):1289–302, DOI 10.1534/genetics.109.102350

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, St Clair JB, Neale DB (2009b) Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*, pinaceae). I. cold-hardiness related traits. Genetics 182(4):1289–302, DOI 10.1534/genetics.109.102350

Eckert AJ, Pande B, Ersoz ES, Wright MH, Rashbrook VK, Nicolet CM, Neale DB (2009c) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda l.*). Tree Genet Genomes 5(1):225–234, DOI 10.1007/s11295-008-0183-8

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD, Tearse BR, Krutovsky KV, Neale DB (2009d) Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*). Genetics 183(1):289–98, DOI 10.1534/genetics.109.103895

El-Lakany MH, Sziklai O (1970) Variation in nuclear characteristics in selected western conifers and its relation to radiosensitivity. Radiat Bot 10(5):421–427, DOI 10.1016/S0033-7560(70)80004-7

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):e19,379, DOI 10.1371/journal.pone.0019379

Ensminger I, Schmidt L, Lloyd J (2008) Soil temperature and intermittent frost modulate the rate of recovery of photosynthesis in Scots pine under simulated spring conditions. New Phytol 177(2):428–42, DOI 10.1111/j.1469-8137.2007.02273.x

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8(8):610–618, DOI 10.1038/nrg2146

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3):1405–1413

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34(3):e22–e22, DOI 10.1093/nar/gnj023

Fink AH, Brücher T, Krüger A, Leckebusch GC, Pinto JG, Ulbrich U (2004) The 2003 European summer heatwaves and drought – synoptic diagnosis and impacts. Weather 59(8):209–216, DOI 10.1256/wea.73.04

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180(2):977–93, DOI 10.1534/genetics.108.092221

Frontier S (1976) Étude de la decroissance des valeurs propres dans une analyze en composantes principales: comparison avec le modèle de baton brisé. J Exp Mar Biol Ecol 25:67–75, DOI 10.1016/0022-0981(76)90076-9

Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. PLoS ONE 8(11):e79,667, DOI 10.1371/journal.pone.0079667

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186(1):207–18, DOI 10.1534/genetics.110.114397

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schön CC,

Sun Q, Walter H, Martin OC, Falque M (2011) A large maize (*Zea mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS ONE 6(12):e28,334, DOI 10.1371/journal.pone.0028334

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet JM, Estoup A (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol 22(11):3165–78, DOI 10.1111/mec.12089

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11(5):759–769, DOI 10.1111/j.1755-0998.2011.03024.x

González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006) DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172(3):1915–26, DOI 10.1534/genetics.105.047126

Grivet D, Sebastiani F, Alía R, Bataillon T, Torre S, Zabal-Aguirre M, Vendramin GG, González-Martínez SC (2011) Molecular footprints of local adaptation in two mediterranean conifers. Mol Biol Evol 28(1):101–16, DOI 10.1093/molbev/msq190

Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. Am J Bot 99(2):312–9, DOI 10.3732/ajb.1100323

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. Genetics 195(1):205–20, DOI 10.1534/genetics.113.152462

Halliday WED, Brown AWA (1943) The distribution of some important forest trees in Canada. Ecology 24(3):353–373, DOI 10.2307/1930537

Hamanishi ET, Campbell MM (2011) Genome-wide responses to drought in forest trees. Forestry 84(3):273–283, DOI 10.1093/forestry/cpr012

Hamrick JL, Godt MJ, Sherman-Broyles SL (1992) Factors influencing levels of genetic diversity in woody plant species. New Forests 6(1):95–124, DOI 10.1007/BF00120641

Hanewinkel M, Cullmann DA, Schelhaas MJ, Nabuurs GJ, Zimmermann NE (2013) Climate change may cause severe loss in the economic value of European forest land. Nature Clim Change 3(3):203–207, DOI 10.1038/nclimate1687

Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T, Jeddeloh JA, Jia G, Springer NM, Vance CP, Stupar RM (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol 155(2):645–55, DOI 10.1104/pp.110.166736

He F, Zhang X, Hu J, Turck F, Dong X, Goebel U, Borevitz J, de Meaux J (2012) Genome-wide analysis of *cis*-regulatory divergence between species in the *Arabidopsis* genus. Mol Biol Evol 29(11):3385–95, DOI 10.1093/molbev/mss146

Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, Clark AG, Nielsen R (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res 18(7):1020–1029, DOI 10.1101/gr.074187.107

Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, Cariani A, Maes GE, Diopere E, Carvalho GR, Nielsen EE (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol Ecol Resour 11 Suppl 1:123–36, DOI 10.1111/j.1755-0998.2010.02943.x

Henry IM, Nagalakshmi U, Lieberman MC, Ngo KJ, Krasileva KV, Vasquez-Gross H, Akhunova A, Akhunov E, Dubcovsky J, Tai TH, Comai L (2014) Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. Plant Cell 26(4):1382–1397, DOI 10.1105/tpc.113.121590

Hermann RK (1981) Die Gattung *Pseudotsuga* - Ein Abriß ihrer Systematik, Geschichte und heutigen Verbreitung. Forstarchiv 52(6):204–212

Hermann RK, Lavender DP (1990) *Pseudotsuga menziesii* (Mirb.) Franco. In: Silvics of North America: 1. Conifers, Burns, RM and Honkala, BH (tech. coords.), U.S. Department of Agriculture, Agriculture Handbook 654, Washington DC, pp 527–540

Hermann RK, Lavender DP (1999) Douglas-fir planted forests. New Forests 17:53–70, DOI 10.1023/A:1006581028080

Hess M, Wildhagen H, Ensminger I (2013) Suitability of Illumina deep mRNA sequencing for reliable gene expression profiling in a non-model conifer species (*Pseudotsuga menziesii*). Tree Genet Genomes 9(6):1513–1527, DOI 10.1007/s11295-013-0656-2

Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics 174(4):2095–105, DOI 10.1534/genetics.106.065102

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6(2):e1000,862, DOI 10.1371/journal.pgen.1000862

Howe GT, Yu J, Knaus B, Cronn R, Kolpak S, Dolan P, Lorenz WW, Dean JFD (2013) A SNP resource for Douglas-fir: de novo transcriptome assembly and SNP detection and validation. BMC Genomics 14:137, DOI 10.1186/1471-2164-14-137

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18:337–8, DOI 10.1093/bioinformatics/18.2.337

Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9(3):90–95, DOI 10.1109/MCSE.2007.55

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37(suppl 1):D211–215, DOI 10.1093/nar/gkn785

IPCC (2007) Intergovernmental Panel on Climate Change - Fourth Assessment Report

Jackson DA (1993) Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. Ecology 74(8):2204–2214, DOI 10.2307/1939574

Jansen K, Sohrt J, Kohnle U, Ensminger I, Gessler A (2013) Tree ring isotopic composition, radial increment and height growth reveal provenance-specific reactions of Douglas-fir towards environmental parameters. Trees 27(1):37–52, DOI 10.1007/s00468-012-0765-9

Jaramillo-Correa JP, Verdú M, González-Martínez SC (2010) The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. BMC Evol Biol 10(1), DOI 10.1186/1471-2148-10-22

Jermstad KD, Eckert AJ, Wegrzyn JL, Delfino-Mix A, Davis DA, Burton DC, Neale DB (2011) Comparative mapping in *Pinus*: sugar pine (*Pinus lambertiana* dougl.) and loblolly pine (*Pinus taeda* l.). Tree Genet Genomes 7(3):457–468, DOI 10.1007/s11295-010-0347-1

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405, DOI 10.1093/bioinformatics/btn129

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94, DOI 10.1186/1471-2156-11-94

Kenk G, Thren M (1984) Ergebnisse verschiedener Douglasienprovenienzversuche in Baden-Württemberg. Teil I: Der Internationale Douglasien-Provenienzversuch 1958. Allg Forst-Jagdztg 155:165–184

Kleinschmit J, Bastien JC (1992) IUFRO's role in Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) tree improvement. Silvae Genet

Kleinschmit J, Svolba J, Weisgerber H, Jestaedt M, Dimpflmeier R, Ruetz W, Dieterich H (1979) Ergebnisse aus dem internationalen Douglasien-Herkunftsversuch von 1970 in der Bundesrepublik Deutschland. Silvae Genet 28:5–6

Kofler R, Pandey RV, Schlötterer C (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27(24):3435–3436, DOI 10.1093/bioinformatics/btr589

Kohnle U, Hein S, Sorensen FC, Weiskittel AR (2012) Effects of seed source origin on bark thickness of Douglas-fir (*Pseudotsuga menziesii*) growing in southwestern Germany. Can J For Res 42(2):382–399, DOI 10.1139/x11-191

Konnert M, Ruetz W, Schirmer R (2008) Fragen zum forstlichen Vermehrungsgut bei Douglasie. LWF - Wissen 59(22–26)

Kownatzki D, Kriebitzsch WU, Bolte A, Liesebach H, Schmitt U, Elsasser P (2011) Zum Douglasienanbau in Deutschland: ökologische, waldbauliche, genetische und holzbiologische Gesichtspunkte des Douglasienanbaus in Deutschland und den angrenzenden Staaten aus naturwissenschaftlicher und gesellschaftspolitischer Sicht. Braunschweig : Johann Heinrich von Thünen-Institut (vTI), Bundesforschungsinstitut für Ländliche Räume, Wald und Fischerei

Krutovsky K, St Clair J, Saich R, Hipkins V, Neale D (2009) Estimation of population structure in coastal Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco var. *menziesii*] using allozyme and microsatellite markers. Tree Genet Genomes 5(4):641–658, DOI 10.1007/s11295-009-0216-y

Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. Genetics 171(4):2029–41, DOI 10.1534/genetics.105.044420

Kumar S, Blaxter M (2010) Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11(1):571, DOI 10.1186/1471-2164-11-571

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–60, DOI 10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–9, DOI 10.1093/bioinformatics/btp352

Li P, Adams WT (1989) Range-wide patterns of allozyme variation in Douglas-fir (*Pseudotsuga menziesii*). Can J For Res 19(2):149–161, DOI 10.1139/x89-022

Little EL Jr (1971) Atlas of United States trees, volume 1 – Conifers and important hardwoods. U.S. Department of Agriculture Forest Service Miscellaneous Publication 1146

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. Mol Ecol 23(9):2178–2192, DOI 10.1111/mec.12725

Lowell EC, Maguire DA, Briggs DG, Turnblom EC, Jayawickrama KJS, Bryce J (2014) Effects of silviculture and genetics on branch/knot attributes of coastal Pacific Northwest Douglas-fir and implications for wood quality—A synthesis. Forests 5(7):1717–1736, DOI 10.3390/f5071717

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE (2013) Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genet 9(1):e1003,215, DOI 10.1371/journal.pgen.1003215

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24(3):133–141, DOI 10.1016/j.tig.2007.12.007

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380, DOI 10.1038/nature03959

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17(1):10–12, DOI 10.14806/ej.17.1.200

Martinez-Meier A, Sanchez L, Dalla-Salda G, Gallo L, Pastorino M, Rozenberg P (2009) Ring density record of phenotypic plasticity and adaptation to drought in Douglas-fir. Forest Ecol Manag 258(5):860–867, DOI 10.1016/j.foreco.2009.03.021

Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, Ayling S, Steuernagel B, Pfeifer M, D'Ascenzo M, Akhunov ED, Hedley PE, Gonzales AM, Morrell PL, Kilian B, Blattner FR, Scholz U, Mayer KFX, Flavell AJ, Muehlbauer GJ, Waugh R, Jeddeloh JA, Stein N (2013) Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. Plant J 76(3):494–505, DOI 10.1111/tpj.12294

Meehl Ga, Tebaldi C (2004) More intense, more frequent, and longer lasting heat waves in the 21st century. Science 305(5686):994–7, DOI 10.1126/science.1098704

Menzies A, Newcombe CF, Forsyth J, Vancouver G (1923) Menzie's journal of Vancouver's voyage, April to October, 1792. Provincial Archives of British Columbia

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10:219, DOI 10.1186/1471-2164-10-219

Mosca E, Eckert AJ, Liechty JD, Wegrzyn JL, La Porta N, Vendramin GG, Neale DB (2012) Contrasting patterns of nucleotide diversity for four conifers of Alpine European forests. Evol Appl 5(7):762–75, DOI 10.1111/j.1752-4571.2012.00256.x

Mukhopadhyay P, Basak S, Ghosh TC (2008) Differential selective constraints shaping codon usage pattern of housekeeping and tissue-specific homologous genes of rice and Arabidopsis. DNA Res 15(6):347–356, DOI 10.1093/dnares/dsn023

Müller T, Ensminger I, Schmid KJ (2012) A catalogue of putative unique transcripts from Douglas-fir (*Pseudotsuga menziesii*) based on 454 transcriptome sequencing of genetically diverse, drought stressed seedlings. BMC Genomics 13(1):673, DOI 10.1186/1471-2164-13-673

Müller T, Freund F, Wildhagen H, Schmid KJ (2015a) Targeted re-sequencing of five Douglas-fir provenances reveals population structure and putative target genes of positive selection. Tree Genet Genomes 11(1):1–17, DOI 10.1007/s11295-014-0816-z

Müller T, Kokai-Kota E, Schmid KJ (2015b) Comparison of genotyping-by-sequencing and sequence capture for population structure inference in Douglas-fir. Mol Ecol Resour Manuscript submitted for publication.

Myhre S, Tveit H, Mollestad T, Lægreid A (2006) Additional Gene Ontology structure for improved biological reasoning. Bioinformatics 22(16):2020–2027, DOI 10.1093/bioinformatics/btl334

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol 22(11):2841–7, DOI 10.1111/mec.12350

NCBI data repository (2012) NCBI data repository. URL http://www.ncbi.nlm.nih.gov/protein?term=picea%20sitchensis

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11(2):149–155, DOI 10.1016/j.pbi.2007.12.004

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12(2):111–122, DOI 10.1038/nrg2931

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9(7):325–330, DOI 10.1016/j.tplants.2004.05.006

Neale DB, Langley CH, Salzberg SL, Wegrzyn JL (2013) Open access to tree genomes: the path to a better forest. Genome Biol 14(6):120, DOI 10.1186/gb-2013-14-6-120

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu LS, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, Dejong PJ, Yorke JA, Salzberg SL, Langley CH (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15(3):R59, DOI 10.1186/gb-2014-15-3-r59

Nelson W, Luo M, Ma J, Estep M, Estill J, He R, Talag J, Sisneros N, Kudrna D, Kim H, Ammiraju JSS, Collura K, Bharti AK, Messing J, Wing RA, SanMiguel P, Bennetzen JL, Soderlund C (2008) Methylation-sensitive linking libraries enhance gene-enriched sequencing of complex genomes and map DNA methylation domains. BMC Genomics 9:621, DOI 10.1186/1471-2164-9-621

NimbleGen (2011) NimbleGen SeqCap EZ Library SR User's Guide, version 3.0. Roche NimbleGen

Ning Z, Cox A, Mullikin J (2001) SSAHA: A fast search method for large DNA databases. Genome Res 11(10):1725–1729, DOI 10.1101/gr.194201

Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinform 11:187, DOI 10.1186/1471-2105-11-187

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9:312, DOI 10.1186/1471-2164-9-312

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hällman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Käller M, Luthman J, Lysholm F, Niittylä T, Olson A, Rilakovic N, Ritland C, Rosselló JA, Sena J, Svensson T, Talavera-López C, Theißen G, Tuominen H, Vanneste K, Wu ZQ, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Garcia Gil R, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497(7451):579–84, DOI 10.1038/nature12211

Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. Phil Trans R Soc B 365(1537):185–205, DOI 10.1098/rstb.2009.0219

Palmé AE, Wright M, Savolainen O (2008) Patterns of divergence among conifer ESTs and polymorphism in *Pinus sylvestris* identify putative selective sweeps. Mol Biol Evol 25(12):2567–77, DOI 10.1093/molbev/msn194

Palmé AE, Pyhäjärvi T, Wachowiak W, Savolainen O (2009) Selection on nuclear genes in a *Pinus* phylogeny. Mol Biol Evol 26(4):893–905, DOI 10.1093/molbev/msp010

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20(2):289–290, DOI 10.1093/bioinformatics/btg412

Parchman T, Geist K, Grahnen J, Benkman C, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11(1):180, DOI 10.1186/1471-2164-11-180

Pare G (2010) Genome-wide association studies–data generation, storage, interpretation, and bioinformatics. J Cardiovasc Transl Res 3(3):183–8, DOI 10.1007/s12265-010-9181-y

Pauchet Y, Wilkinson P, van Munster M, Augustin S, Pauron D, ffrench Constant RH (2009) Pyrosequencing of the midgut transcriptome of the poplar leaf beetle *Chrysomela tremulae* reveals new gene families in Coleoptera. Insect Biochem Molec Biol 39(5-6):403–413, DOI 10.1016/j.ibmb.2009.04.001

Pavlidis P, Laurent S, Stephan W (2010) msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. Mol Ecol Resour 10(4):723–7, DOI 10.1111/j.1755-0998.2010.02832.x

Pavy N, Parsons LS, Paule C, MacKay J, Bousquet J (2006) Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. BMC Genomics 7:174, DOI 10.1186/1471-2164-7-174

Peters D, Luo X, Qiu K, Liang P (2012) Speeding up large-scale next generation sequencing data analysis with pBWA. J Appl Bioinform Comput Biol 1(1), DOI 10.4172/jabcb.1000101

Pharis RP, Ferrell WK (1966) Differences in drought resistance between coastal and inland sources of Douglas fir. Can J Bot 44(12):1651–1659, DOI 10.1139/b66-177

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink JL (2012a) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5(3):103, DOI 10.3835/plantgenome2012.06.0006

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5(3):92, DOI 10.3835/plantgenome2012.05.0005

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012b) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2):e32,253, DOI 10.1371/journal.pone.0032253

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–59

Pujolar JM, Jacobsen MW, Als TD, Frydenberg J, Munch K, Jónsson B, Jian JB, Cheng L, Maes GE, Bernatchez L, Hansen MM (2014) Genome-wide single-generation signatures of local selection in the panmictic European eel. Mol Ecol 23(10):2514–28, DOI 10.1111/mec.12753

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly M, Sham P (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81(3):559–575, DOI 10.1086/519795

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE (2014) Demystifying the RAD fad. Mol Ecol 23(24):5937–5942, DOI 10.1111/mec.12965

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–2, DOI 10.1093/bioinformatics/btq033

Rehfeldt GE (1979) Ecological adaptations in Douglas-fir (*Pseudotsuga menziesii* var. *glauca*) populations. Heredity 43(3):383–397, DOI 10.1038/hdy.1979.89

Rehfeldt GE (1989) Ecological adaptations in Douglas-fir (*Pseudotsuga menziesii* var. *glauca*): a synthesis. Forest Ecol Manag 28(3–4):203–215, DOI 10.1016/0378-1127(89)90004-2

Reyer C, Lasch P, Mohren GMJ, Sterck FJ (2010) Inter-specific competition in mixed forests of Douglas-fir (*Pseudotsuga menziesii*) and common beech (*Fagus sylvatica*) under climate change - a model-based analysis. Ann For Sci 67(8):805p1–11, DOI 10.1051/forest/2010041

Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, Mackay JJ (2011) A white spruce gene catalog for conifer genome analyses. Plant Physiol 157(1):14–28, DOI 10.1104/pp.111.179663

Riggins CW, Peng Y, Stewart CN Jr, Tranel PJ (2010) Characterization of de novo transcriptome for waterhemp (*Amaranthus tuberculatus*) using GS-FLX 454 pyrosequencing and its application for studies of herbicide target-site genes. Pest Manag Sci 66(10):1042–52, DOI 10.1002/ps.2006

Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T (2009) A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). Mol Phylogenet Evol 53(3):826–34, DOI 10.1016/j.ympev.2009.08.014

Roesti M, Salzburger W, Berner D (2012) Uninformative polymorphisms bias genome scans for signatures of selection. BMC Evol Biol 12:94, DOI 10.1186/1471-2148-12-94

Rudloff Ev (1972) Chemosystematic studies in the genus Pseudotsuga. I. Leaf oil analysis of the coastal and Rocky Mountain varieties of the Douglas-fir. Can J Bot 50(5):1025–1040, DOI 10.1139/b72-126

Salem M, Rexroad CE, Wang J, Thorgaard GH, Yao J (2010) Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches. BMC Genomics 11:564, DOI 10.1186/1471-2164-11-564

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74(12):5463–7, DOI 10.1073/pnas.74.12.5463

Schober R (1972) Zur Gründung des Vereins der Forstlichen Versuchsanstalten Deutschlands vor 100 Jahren. Forstarchiv 43:221–227

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, Alkan C, Kidd JM, Sun Y, Drautz DI, Bouffard P, Muzny DM, Reid JG, Nazareth LV, Wang Q, Burhans R, Riemer C, Wittekindt NE, Moorjani P, Tindall EA, Danko CG, Teo WS, Buboltz AM, Zhang Z, Ma Q, Oosthuysen A, Steenkamp AW, Oostuisen H, Venter P, Gajewski J, Zhang Y, Pugh BF, Makova KD, Nekrutenko A, Mardis ER, Patterson N, Pringle TH, Chiaromonte F, Mullikin JC, Eichler EE, Hardison RC, Gibbs RA, Harkins TT, Hayes VM (2010) Complete Khoisan and Bantu genomes from southern Africa. Nature 463(7283):943–947, DOI 10.1038/nature08795

Schwappach A (1907) Über den Wert der verschiedenen Formen der Douglas-Fichte. Mitteilungen der Deutschen Dendrologischen Gesellschaft 16:122–125

SeqClean (2012) SeqClean. URL http://sourceforge.net/projects/seqclean/

Shagin DA, Rebrikov DV, Kozhemyako VB, Altshuler IM, Shcheglov AS, Zhulidov PA, Bogdanova EA, Staroverov DB, Rasskazov VA, Lukyanov S (2002) A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. Genome Res 12(12):1935–42, DOI 10.1101/gr.547002

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotech 26(10):1135–1145, DOI 10.1038/nbt1486

Shinozaki K, Yamaguchi-Shinozaki K (2007) Gene networks involved in drought stress response and tolerance. J Exp Bot 58(2):221–227, DOI 10.1093/jxb/erl164

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321(6071):674–679, DOI 10.1038/321674a0

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36(suppl 1):D1009–14, DOI 10.1093/nar/gkm965

Tajima F (1989) Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814):796–815, DOI 10.1038/35048692

Tsukada M (1982) *Pseudotsuga menziesii* (Mirb.) Franco: its pollen dispersal and later Quaternary history in the Pacific Northwest. Jpn J Ecol 32(2):159–187

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, dePamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313(5793):1596–1604, DOI 10.1126/science.1128691

Uitdewilligen JGAML, Wolters AMA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS ONE 8(5):e62,355, DOI 10.1371/journal.pone.0062355

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17(7):1636–47, DOI 10.1111/j.1365-294X.2008.03666.x

Viard F, El-Kassaby YA, Ritland K (2001) Diversity and genetic structure in populations of *Pseudotsuga menziesii* (Pinaceae) at chloroplast microsatellite loci. Genome 44(3):336–44, DOI 10.1139/gen-44-3-336

Wang W, Vinocur B, Shoseyov O, Altman A (2004) Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends Plant Sci 9(5):244–252, DOI 10.1016/j.tplants.2004.03.006

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63, DOI 10.1038/nrg2484

Watkinson JI, Sioson AA, Vasquez-Robinet C, Shukla M, Kumar D, Ellis M, Heath LS, Ramakrishnan N, Chevone B, Watson LT, Zyl Lv, Egertsdotter U, Sederoff RR, Grene R (2003) Photosynthetic acclimation is reflected in specific patterns of gene expression in drought-stressed loblolly pine. Plant Physiol 133(4):1702–1716, DOI 10.1104/pp.103.026914

Wei XX, Beaulieu J, Khasa D, Vargas-Hernández J, López-Upton J, Jaquish B, Bousquet J (2011) Range-wide chloroplast and mitochondrial DNA imprints reveal multiple lineages and complex biogeographic history for Douglas-fir. Tree Genet Genomes 7(5):1025–1040, DOI 10.1007/s11295-011-0392-4

Weigel D, Mott R (2009) The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol 10(5):107, DOI 10.1186/gb-2009-10-5-107

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38(6):1358–1370, DOI 10.2307/2408641

Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76(5):887–93, DOI 10.1086/429864

Williams LM, Oleksiak MF (2011) Ecologically and evolutionarily important SNPs identified in natural populations. Mol Biol Evol 28(6):1817–26, DOI 10.1093/molbev/msr004

Wright S (1951) The genetical structure of populations. Ann Eugen 15(4):323–54

Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, Khan MA, Tao S, Korban SS, Wang H, Chen NJ, Nishio T, Xu X, Cong L, Qi K, Huang X, Wang Y, Zhao X, Wu J, Deng C, Gou C, Zhou W, Yin H, Qin G, Sha Y, Tao Y, Chen H, Yang Y, Song Y, Zhan D, Wang J, Li L, Dai M, Gu C, Wang Y, Shi D, Wang X, Zhang H, Zeng L, Zheng D, Wang C, Chen M, Wang G, Xie L, Sovero V, Sha S, Huang W, Zhang S, Zhang M, Sun J, Xu L, Li Y, Liu X, Li Q, Shen J, Wang J, Paull RE, Bennetzen JL, Wang J, Zhang S (2013) The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res 23(2):396–408, DOI 10.1101/gr.144311.112

Xu L, Chen H, Hu X, Zhang R, Zhang Z, Luo ZW (2006) Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Mol Biol Evol 23(6):1107–1108, DOI 10.1093/molbev/msk019

Yang H, Ding Y, Hutchins LN, Szatkiewicz J, Bell TA, Paigen BJ, Graber JH, de Villena FPM, Churchill GA (2009) A customized and versatile high-density genotyping array for the mouse. Nat Methods 6(9):663–666, DOI 10.1038/nmeth.1359

Zdobnov EM, Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17(9):847–8, DOI 10.1093/bioinformatics/17.9.847

Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 21(2):236–239, DOI 10.1093/molbev/msh010

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. BioTechniques 30(4):892–7

Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA, Shagin DA (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res 32(3):e37, DOI 10.1093/nar/gnh031

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, Neale DB, Salzberg SL, Yorke JA, Langley CH (2014) Sequencing and assembly of the 22-Gb loblolly pine genome. Genetics 196(3):875–890, DOI 10.1534/genetics.113.159715

Nomenclature

c	Control
m	Mild stress
s	Severe stress
ABC	Approximate Bayesian computation
AR	Salmon Arms
ATP	Adenosine triphosphate
BC	British Columbia
BC1	Twin Lake
BC6	Prince George
BLAST	Basic local alignment search tool
bp	Base pair
c	Coastal
CA	California
cDNA	Complementary DNA
CDS	Coding DNA sequence
CO	Fort Collins
CR	Conrad Creek
CTAB	Cetyltrimethylammonium bromide
DA	Discriminant analysis
DAPC	Discriminant analysis of principal components
DCNC	Coastal Douglas-fir, needle tissue, no stress
DCNM	Coastal Douglas-fir, needle tissue, mild stress
DCNS	Coastal Douglas-fir, needle tissue, severe stress
DCWC	Coastal Douglas-fir, wood tissue, no stress
DCWM	Coastal Douglas-fir, wood tissue, mild stress
DCWS	Coastal Douglas-fir, wood tissue, severe stress
DD	Double digestion
ddNTP	di-deoxynucleotidetriphosphate
DINC	Interior Douglas-fir, needle tissue, no stress
DINM	Interior Douglas-fir, needle tissue, mild stress
DINS	Interior Douglas-fir, needle tissue, severe stress
DIWC	Interior Douglas-fir, wood tissue, no stress
DIWM	Interior Douglas-fir, wood tissue, mild stress

DIWS	Interior Douglas-fir, wood tissue, severe stress
DNA	Deoxyribonucleic acid
dNTP	deoxynucleotidetriphosphate
DSN	Duplex-specific nuclease
FDR	False discovery rate
g	Gram
Gbp	Gigabase pairs
GBS	Genotyping-by-sequencing
gDNA	Genomic DNA
GO	Gene ontology
HSP	High-scoring segment pair
HWE	Hardy-Weinberg equilibrium
i	Interior
IPCC	Intergovernmental Panel on Climate Change
kb	Kilobase pairs
km	Kilometer
l	Liter
LA	Cameron Lake
LM-PCR	Ligation mediated PCR
MA	Massachusetts
Mbp	Megabase pairs
MID	Multiplex identifier
min	Minute
MPa	Million pascal
mRNA	Messenger ribonucleic acid
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NL	Netherlands
NM	Raton
PC	Principal component
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PCR	Polymerase chain reaction
PE	Paired end
PUT	Putative unique transcript
RADseq	Restriction site associated DNA sequencing
RAPD	Randomly amplified polymorphic DNA
RE	Restriction enzyme
RI	Santiam River

RNA	Ribonucleic acid
s.d.	Standard deviation
SD	Single digestion
SNP	Single nucleotide polymorphism
TAIR	The Arabidopsis Information Resource
TI	Timber
USA	United States of America
USGS	United States Geological Survey
WA	Washington
WI	Wisconsin

Appendices

A. A catalogue of putative unique transcripts from Douglas-fir (*Pseudotsuga menziesii*) based on 454 transcriptome sequencing of genetically diverse, drought stressed seedlings

Supplementary Figures and Tables

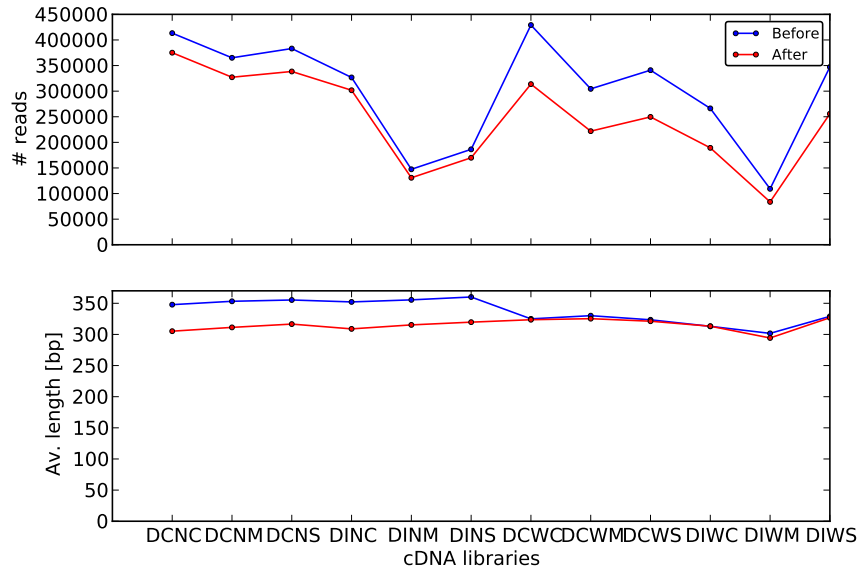


Figure A.1.: Characteristics of the libraries. Number of reads and average read length per library before and after the pre-processing steps.

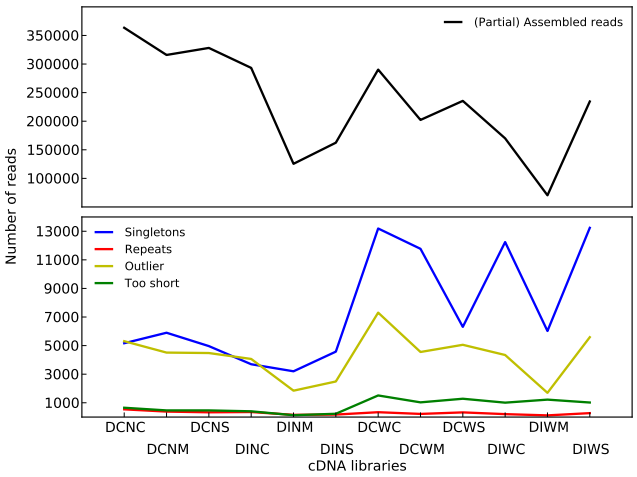


Figure A.2.: Read composition of the assembly. The origin as well as the number of reads assembled or otherwise marked by Newbler is illustrated.

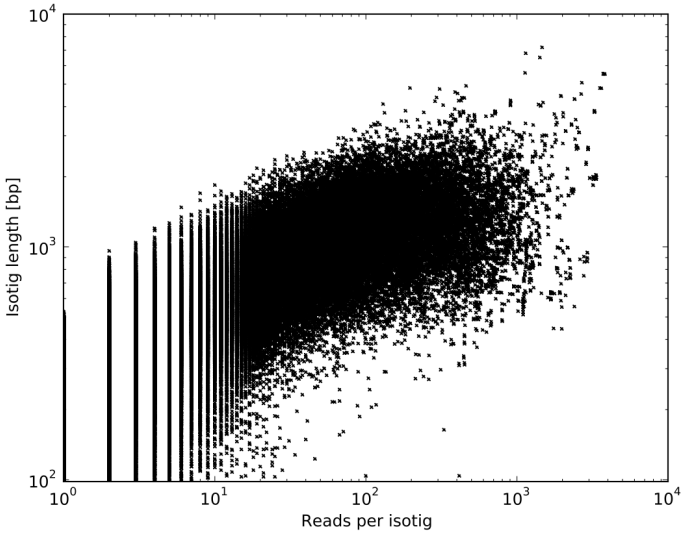


Figure A.3.: The log-log plot shows that the sequence length is depending on the number of reads assembled to the sequence.

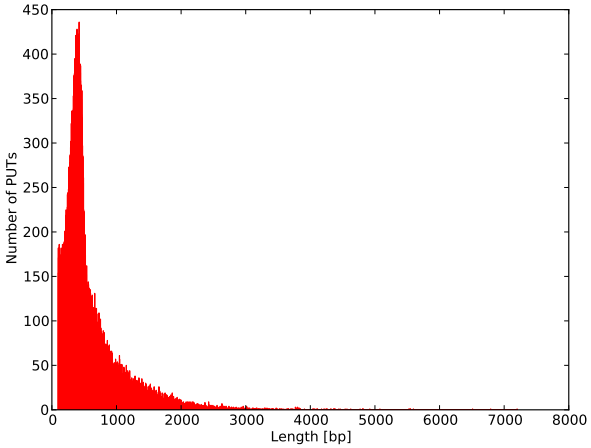


Figure A.4.: Number of isotigs per sequence length. Reads of all twelve cDNA libraries were assembled using Newbler.

Supplementary file 7: BLASTX keyword search results. This file lists in a tab separated style for each BLASTX keyword search hit the following information: keyword, isotig id, isotig group, hit id, hit definition, e-value. If there were more than one hit per keyword and isotig, only the best hit (i.e., the one with the smallest e-value) is listed. Because this file contains more than 2000 lines, please visit the online published version of the paper.

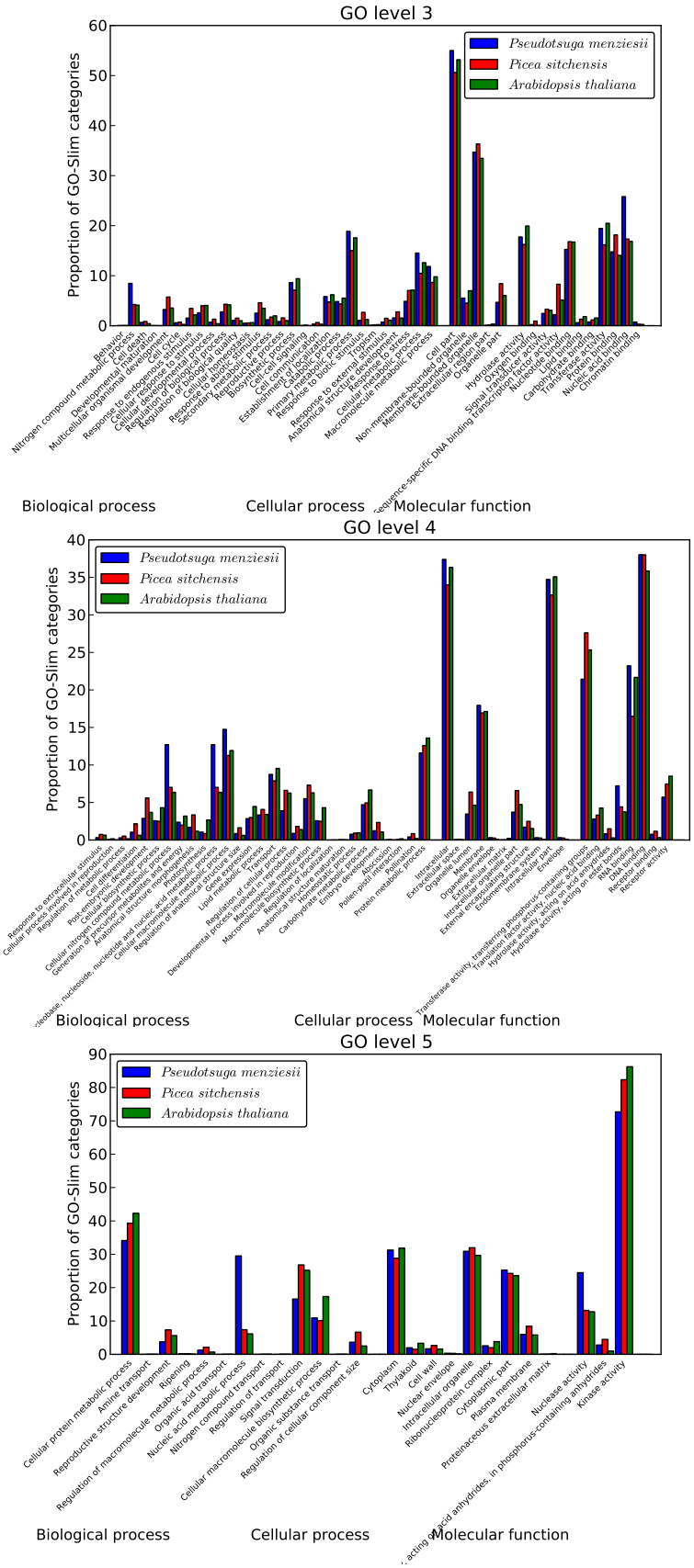


Figure A.5.: Comparison of the GO-Slim categories level 3 - 5. Comparison of the distribution of the GO-Slim categories of the Douglas-fir PUTs set versus *P. sitchensis* and *A. thaliana* at GO level 3 to 5. Transcriptome data of *P. sitchensis* and *A. thaliana* were obtained from NCBI and TAIR databases, respectively.

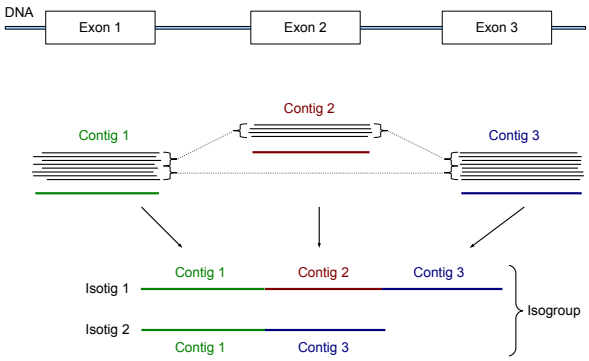


Figure A.6.: Schematic example of contigs, isotigs, and isogroups produced by Newbler. Single reads (black lines) are assembled to contigs. The dotted lines represent subsets of reads implying connections between the contigs. The red, blue, and green line represent the consensus sequence of the contigs. The isogroup consists of two isotigs and in total three different contigs.

Table A.1.: Number of identical BLASTX hits (only results from the keyword search are considered) of isotigs (non-singleton PUTs) from different set(s) of groups (see Figure 2.3). In the not listed combinations of sets, there were no identical BLASTX hits. *c* = control, *m* = mild stress, *s* = severe stress, *cm* = control and mild stress, *cs* = control and severe stress, *ms* = mild and severe stress, *cms* = control, mild and severe stress.

Group(s)	Number of equal BLASTX hits
c cms cs ms	1
cm ms	1
cms cs ms s	1
c cm cms cs m s	1
cs m	1
c cm cms s	1
c cm cms m	1
c cms cs m s	1
c cm	1
cms cs ms	1
m s	1
cm cms s	1
cms cs m s	1
cm cms ms	1
c cms ms	1
c cm cms cs m ms s	1
c s	1
cm s	1
cms m s	2
c cms cs	2
c cm cms cs	2
c cm cms m ms s	2
c cm cms	2
cm cms m	2
m ms	2
c cms cs s	3
c cms s	3
cms cs m	3
cms ms s	4
cs s	4
ms	4
cm cms cs	5
cms cs s	8
cms m	9
cms ms	13
m	14
cm cms	17
cm	19
cs	20
s	21
c	24
cms s	27
cms cs	29
c cms	30
cms	557

Table A.2.: Composition of the cDNA libraries.

Library	Provenance	# individuals
DCNC	Lowry Lake	6
	Twin Harbors	6
	Pend Oreille	5
DCNM	Lowry Lake	5
	Twin Harbors	5
	Pend Oreille	4
DCNS	Lowry Lake	5
	Twin Harbors	5
	Pend Oreille	4
DINC	Twin Lake	6
	Salmon Arms	6
	Prince George	6
	Fort Collins	6
	Raton	6
DINM	Twin Lake	5
	Salmon Arms	5
	Prince George	4
	Fort Collins	4
	Raton	5
DINS	Twin Lake	4
	Salmon Arms	4
	Prince George	4
	Fort Collins	4
	Raton	5
DIWC	Twin Lake	7
	Salmon Arms	3
	Prince George	4
	Fort Collins	3
	Raton	3
DIWM	Twin Lake	4
	Salmon Arms	4
	Prince George	4
	Fort Collins	3
	Raton	5
DIWS	Twin Lake	3
	Salmon Arms	4
	Prince George	4
	Fort Collins	3
	Raton	3
DCWC	Lowry Lake	7
	Twin Harbors	5
	Pend Oreille	5
DCWM	Lowry Lake	6
	Twin Harbors	6
	Pend Oreille	4
DCWS	Lowry Lake	5
	Twin Harbors	5
	Pend Oreille	4

Table A.3.: Origin of the provenances in detail.

Variety	Provenance	Country and province	Elevation	Climate	Experiment code	Nursery code
Coastal	Lowry Lake	Canada, BC	185 m	very moist	BC2	FDC 1294
Coastal	Twin Harbors	USA, WA	0-1000 m	very moist	WA1	FDC SP07-33
Coastal	Pend Oreille	USA, WA	2800-3500 m	montane/dry	WA2	FDC PI06-144
Interior	Fort Collins	USA, CO	2500 m	montane/dry	CO	FDI 123
Interior	Raton	USA, NM	2300 m	montane/dry	NM	FDI unknown
Interior	Prince George	Canada, BC	850 m	Northern limit of range/dry	BC6	FDI 44913
Interior	Salmon Arms	Canada, BC	850 m	dry	BC3	FDI 39924
Interior	Twin Lake	Canada, BC	1067 m	very dry	BC1	FDI 2053

B. Targeted re-sequencing of five Douglas-fir provenances reveals population structure and putative target genes of positive selection

Supplementary Figures

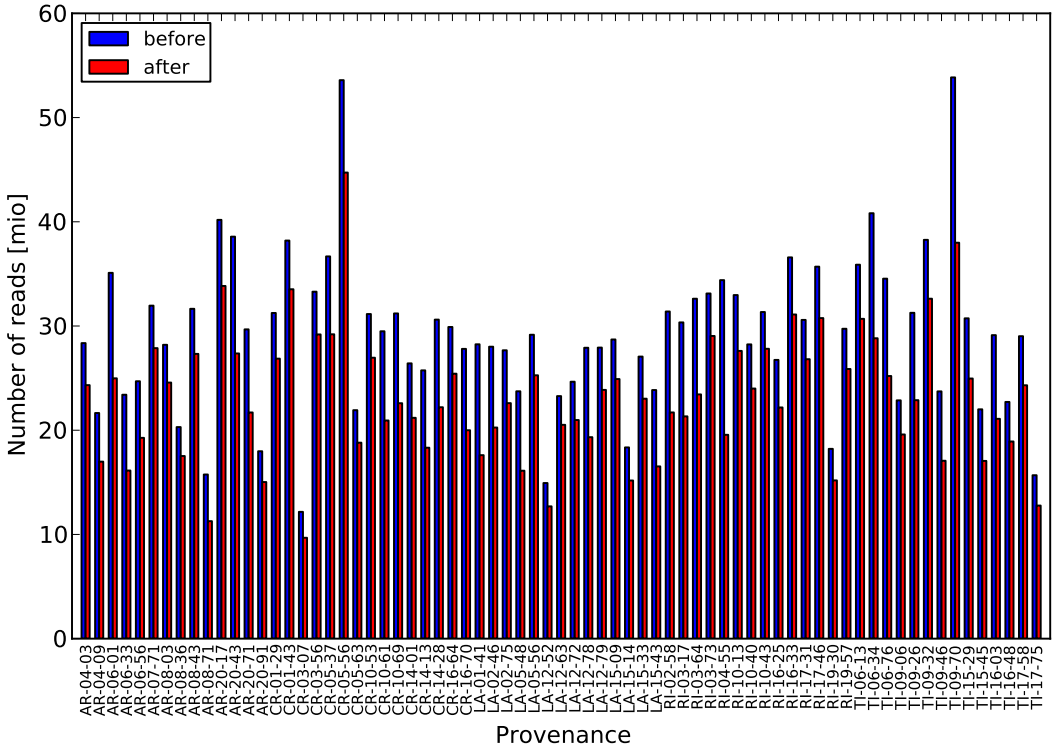
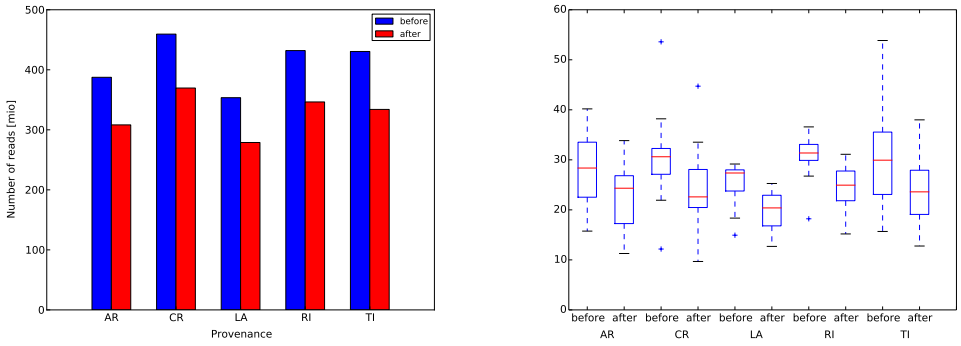


Figure B.1.: Number of reads per library before and after preprocessing.



(a) (b)

Figure B.2.: (a) Total number of reads per provenance and (b) number of reads of libraries per provenance before and after the preprocessing.

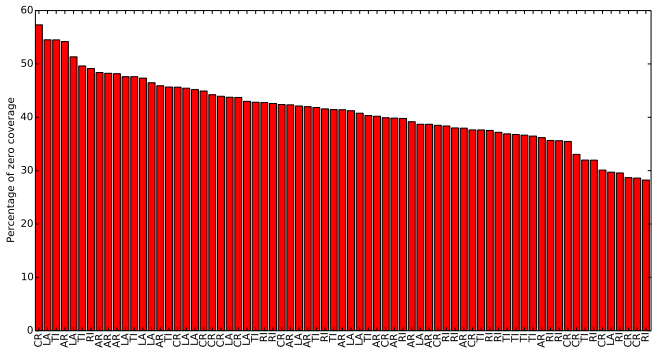


Figure B.3.: Proportion of target nucleotides with zero coverage per individuals.

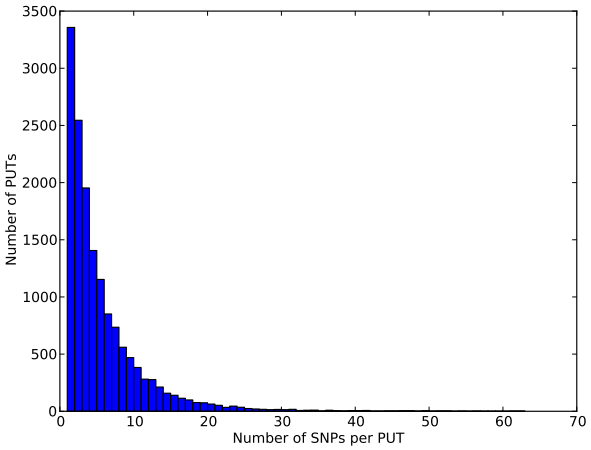
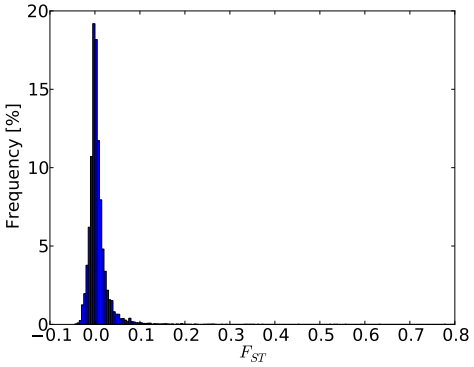
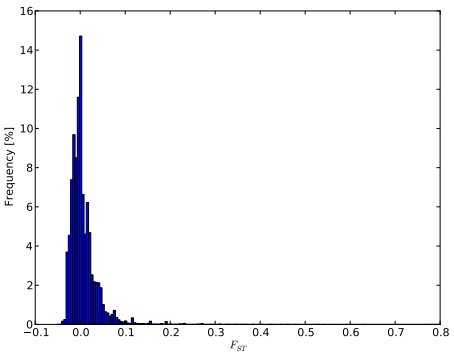


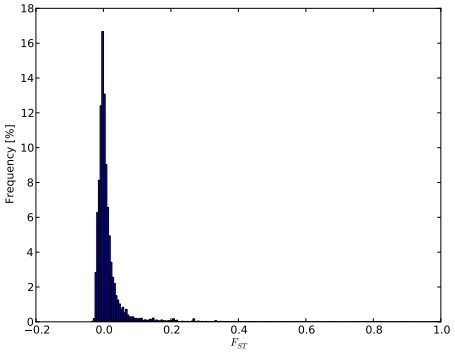
Figure B.4.: Distribution of SNPs per PUT over all PUTs.



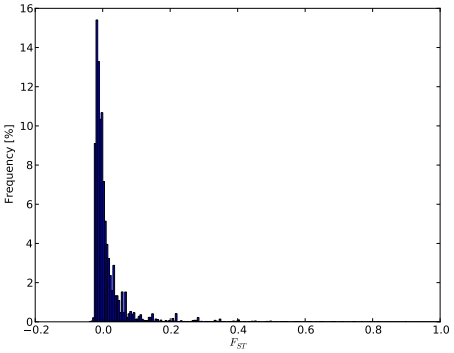
(a) Five provenances, PUTwise



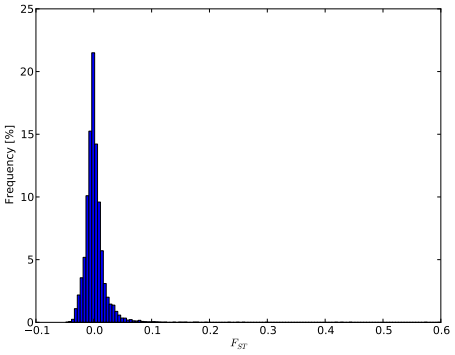
(b) Five provenances, SNPwise



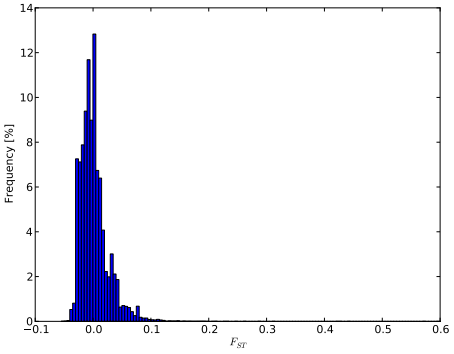
(c) Interior and coastal, PUTwise



(d) Interior and coastal, SNPwise

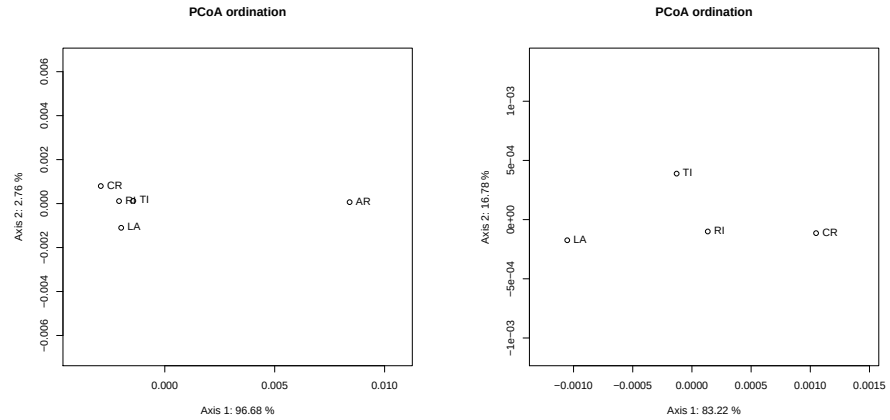


(e) Four coastal provenances, PUTwise



(f) Four coastal provenances, SNPwise

Figure B.5.: Distribution of F_{ST} values. (a) PUTwise F_{ST} values among all provenances, (b) SNPwise F_{ST} values among all provenances, (c) PUTwise F_{ST} values among interior and coastal provenances, (d) SNPwise F_{ST} values among interior and coastal provenances, (e) PUTwise F_{ST} values among the four coastal provenances, (f) SNPwise F_{ST} values among the four coastal provenances.



(a) PCoA - all provenances (b) PCoA - coastal provenances

Figure B.6.: Principal Coordinate Analysis (PCoA) of pairwise F_{ST} values per PUT of (a) five provenances and (b) four coastal provenances.

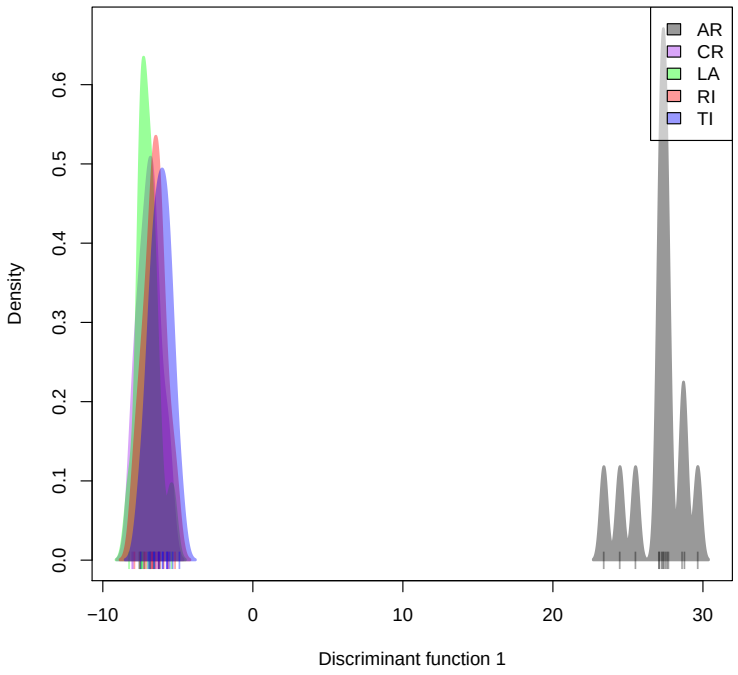


Figure B.7.: DAPC with two clusters identified by adegenet's find.clusters method. AR individuals cluster together and coastal individual cluster together.

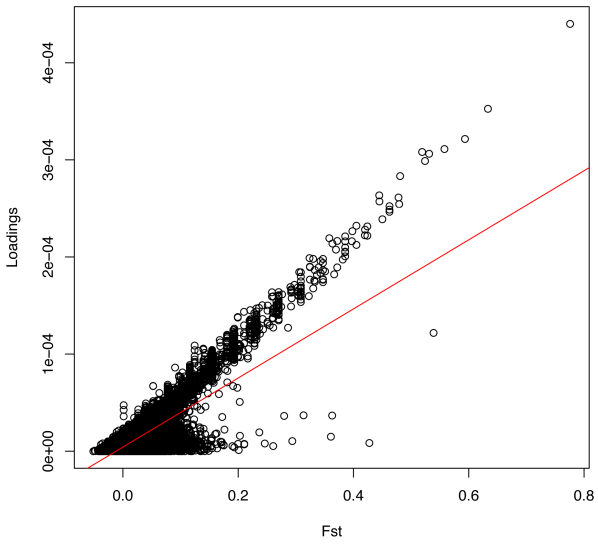


Figure B.8.: Positive correlation (red line) of F_{ST} values and loadings of the DAPC.

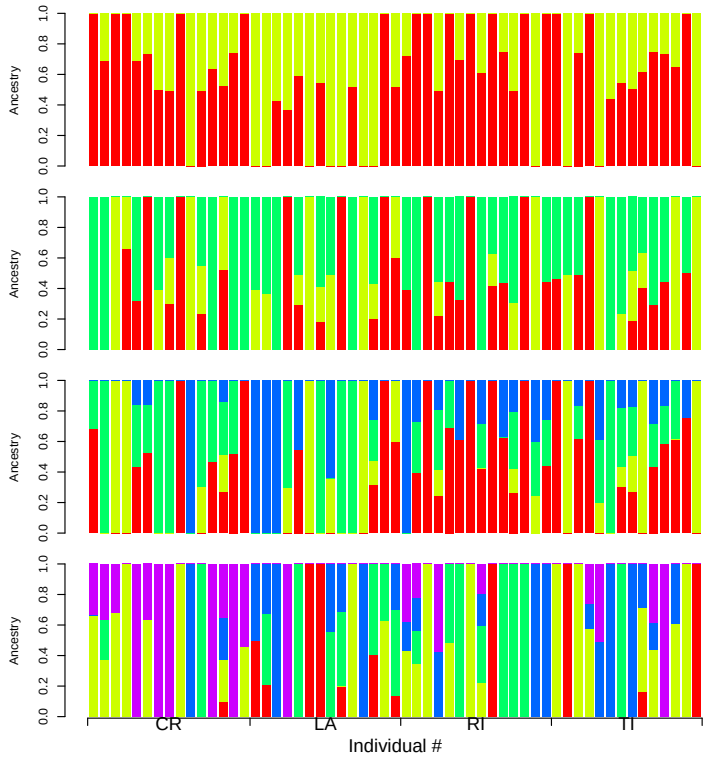
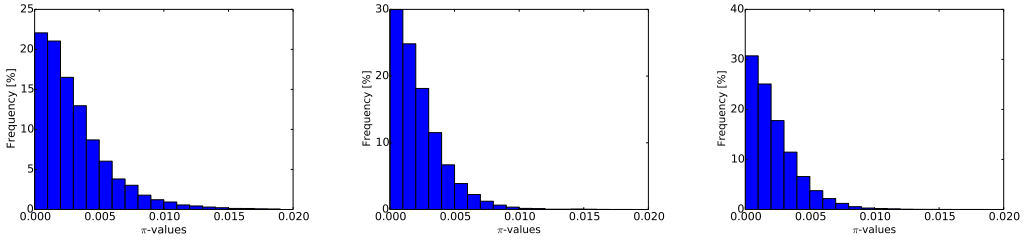


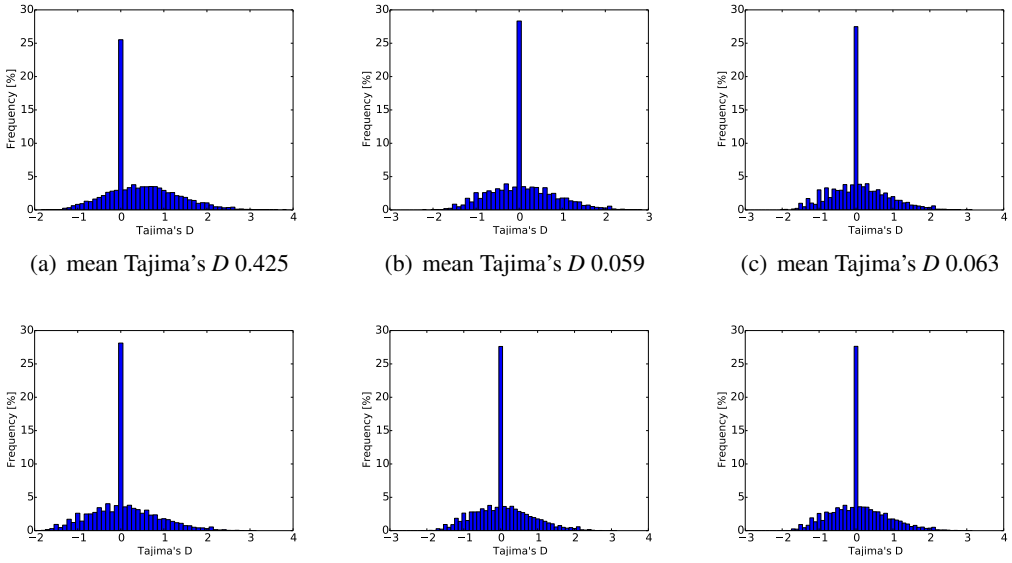
Figure B.9.: ADMIXTURE results with coastal provenance trees for K equals 2 to 5.



(a) mean π 0.0032 (b) mean π 0.0022 (c) mean π 0.0022

(d) mean π 0.0023 (e) mean π 0.002 (f) mean π 0.0022

Figure B.10.: Histogram of nucleotide diversity π per SNP values, calculated for (a) all provenances, (b) AR, (c) CR, (d) LA, (e) RI, and (f) TI.



(a) mean Tajima's D 0.425 (b) mean Tajima's D 0.059 (c) mean Tajima's D 0.063

(d) mean Tajima's D 0.06 (e) mean Tajima's D 0.044 (f) mean Tajima's D 0.051

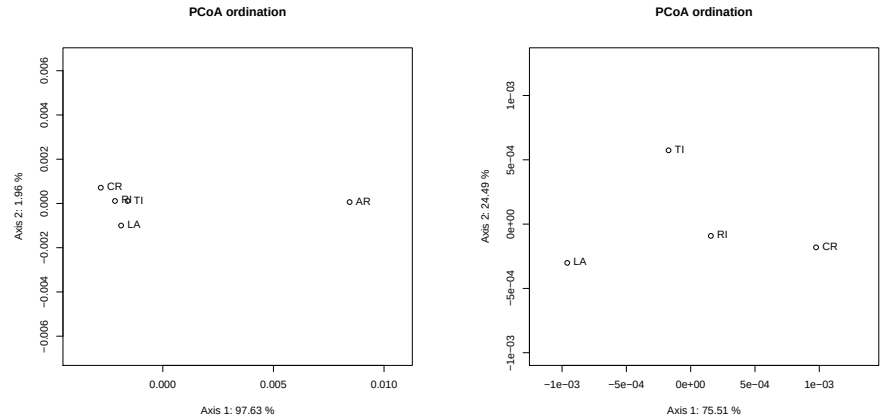
Figure B.11.: Histogram of Tajima's D values calculated for (a) all provenances, (b) AR, (c) CR, (d) LA, (e) RI, and (f) TI.

Supplementary Information

Supplementary Information 1

PCR duplicates

PCR duplicates can be the source for wrong SNP-calling and biased results. We wanted to test whether removing PCR duplicates substantially alters our results, which should not be the case, due to the applied filtering steps. We removed PCR duplicates using bamutils (`bam dedup -rmDups`) (Breese and Liu, 2013), SNP-calling and analysis were performed as described in the paper. Here, we present some of the results, to show that removing duplicates does not alter the results substantially in our case. Without removing the PCR duplicates, we identified 79,910 SNPs. If duplicates were removed we found 81,032 SNPs. 71,097 SNPs were found in both cases, 9,935 SNPs were only identified with duplicates, and 8,813 were identified only after removing PCR duplicates. Pairwise F_{ST} values were similar (Supplementary Figure B.6 and B.12, Table 3.3 and Supplementary Table B.1). Supplementary table B.2 shows that mean values of Tajima's D and π as well as the number of outlier PUTs did not differ with or without PCR duplicates. Furthermore, ADMIXTURE and DAPC results did not show differences (Figure 3.5, 3.4, Supplementary Figure B.13 and B.14). Since the differences were marginal and removing duplicates usually depends on mapping coordinates solely, we decided to use the data set without removed PCR duplicates.



(a) PCoA - all provenances

(b) PCoA - coastal provenances

Figure B.12.: Principal Coordinate Analysis (PCoA) of pairwise F_{ST} values per PUT without PCR duplicates of (a) five provenances and (b) four coastal provenances.

Table B.1.: Pairwise F_{ST} values and standard deviations for each pair of provenances after removing duplicates.

		Interior	Coastal		
		AR	CR	LA	RI
Coastal	CR	0.01126 ±0.0419			
	LA	0.01038 ±0.0415	0.00194 ±0.0286		
	RI	0.0106 ±0.0412	0.00053 ±0.0256	0.00099 ±0.0271	
	TI	0.01004 ±0.0406	0.00137 ±0.0275	0.00118 ±0.0277	0.00056 ±0.0261

Table B.2.: Mean Tajima's D, mean π and number of outlier found in three runs of BayeScan without PCR duplicates. The number in parentheses in the outlier column gives the number of outliers found only with or without PCR duplicates.

	Mean Tajima's D	Mean π	Outlier
PCR duplicates not removed	0.42 ± 0.78	0.0032 ± 0.0033	68 (4)
PCR duplicates removed	0.44 ± 0.78	0.0032 ± 0.0032	69 (5)

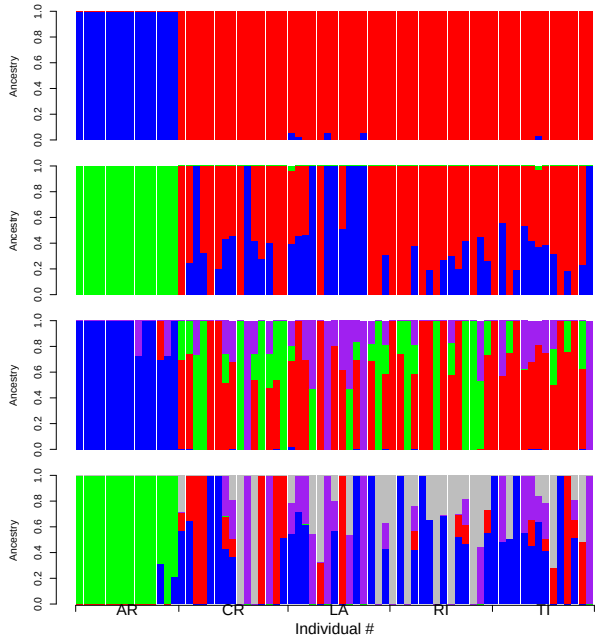


Figure B.13.: ADMIXTURE results for K equals 2 to 5 without PCR duplicates.

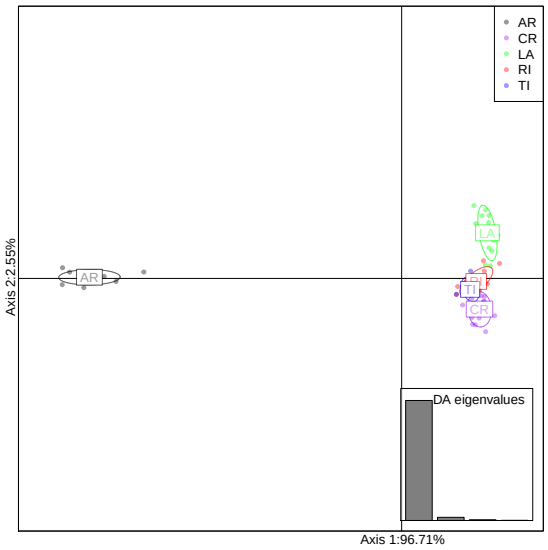


Figure B.14.: DAPC without PCR duplicates.

Supplementary Information 2

Outlier tree

A discriminant analysis of principal components (DAPC) showed an outlier of the interior trees (Sifi-AR-18-5) located more closely to the coastal than to the other interior trees (Figure B.15). Further, yet unpublished phenotypic data, showed, that this tree is also an outlier in other analysis compared with the other interior trees. The tree was located at the edge of an interior field, the neighbor field contained a coastal population (Darrington). Therefore, it is most likely that the doubtful tree is a coastal tree. The tree was excluded from further analysis and SNP detection was performed again without that tree.

Supplementary Information 3

Comparison of π-values

Krutovsky and Neale (2005) calculated π for 18 genes (Krutovsky and Neale, 2005). We downloaded the sequences from uniprot database and performed a BLASTX search of PUTs against those genes. For each PUT which was mapped in the sequence capture approach, we took the best hit. We then compared the *pi*-values of the PUTs with the *pi*-values given in (Krutovsky and Neale, 2005). If a query was hit by several PUTs, we took the mean *pi* of the PUTs for the comparison. For more information see Online resource 7.

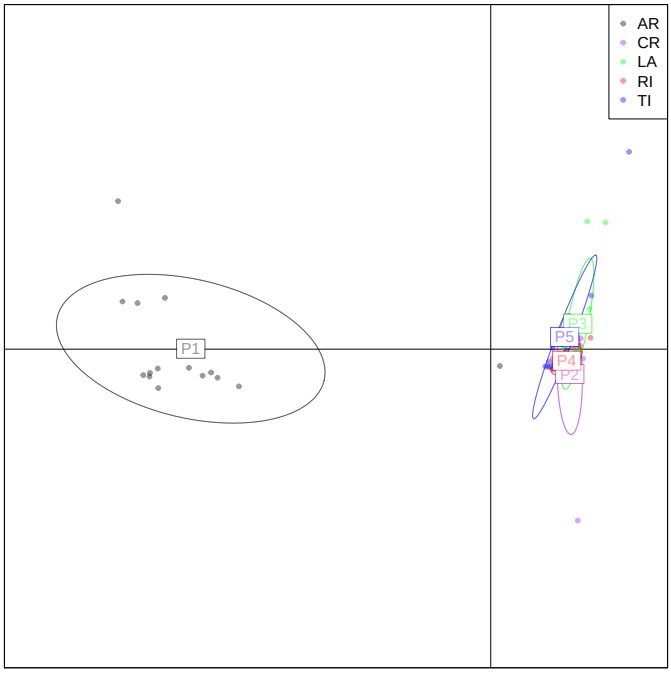


Figure B.15.: DAPC of 72 trees. One outlier of the interior trees is located next to the coastal cluster. This and other results lead to the conclusion, that this tree is a coastal tree.

Table B.3.: Comparison of π-values of 18 genes of Krutovsky and Neale (2005) and PUTs used in this study. n.a. means that no PUT which was captured mapped to the gene.

Full gene name	Gene	Mean π	Mean π (Krutovsky)	Mean Tajima's D	Tajima's D (Krutovsky)
Translation elongation factor-1 α-subunit	EFA1	0.00458	0.00274	0.762	0.656
Thiazole biosynthetic enzyme	TBE	0.00271	0.00516	0.397	0.723
Flavanone-3-hydroxylase	F3H1	n.a.	0.00528	n.a.	1.576
Flavanone-3-hydroxylase	F3H2	0.00285	0.00629	0.312	0.150
Formin-like protein AHF1	Formin-like	n.a.	0.00480	n.a.	1.498
α-tubulin	AT1	0.00309	0.00936	-0.187	0.037
Late embryogenesis abundant type 2 dehydrin-like protein	LEA2	0.00190	0.00647	0.244	0.862
Metallothionein-like protein	MT-like	0.00197	0.01334	-0.177	1.639
60S ribosomal protein L31a	60S-RPL31a	0.00067	0.01011	-1.096	0.479
Late embryogenesis abundant EMB11-like protein	LEA-EMB11	n.a.	0.01378	n.a.	0.593
40S-Ribosomal Protein 3a Protein	40S-RPS3a	0.00344	0.00601	0.319	0.336
Polyubiquitin	PolyUBQ	0.00405	0.00544	0.265	0.357
Early response to dehydration protein	ERD15-like	0.00209	0.00438	0.378	0.757
Abscisic acid water deficit stress and ripening inducible protein	ABA-WDS	n.a.	0.00662	n.a.	0.048
Water deficit inducible protein	LP3-like	0.00271	0.00662	-0.085	0.713
4-coumarate: CoA ligase 1	4CL1	0.00490	0.00268	0.000	0.460
4-coumarate: CoA ligase 2	4CL2	0.00219	0.00237	0.464	1.128
Ascorbate peroxidase	APX	0.00429	0.00636	0.556	0.700

Supplementary Information 4

Additional parameter sets for the I and the PS model and ABC analysis

Additional analyses were performed for the island (I) and the population-split (PS) model to assess the influence of parameter ranges on the models, especially those ranges for which the posterior estimates were suggesting illegitimate choices. Parameter spaces were changed to $\theta \in [0,\ 4\times10^{-3}]$, $\rho \in [\ 0.5\times10^{-3},\ 3\times10^{-3}]$ and $m_{11}, m_{21} \in [50, 150]$. For the PS model, the divergence time T was taken from $[5, 11]$. Uniform priors were set for all parameters. The I model and the PS model with modified parameter ranges are referred to as model I' and PS'. Supplementary Table B.4 shows the Bayes factors from the model comparisons of the modified models with the models I and PS. Supplementary Figure B.16 shows the posterior distributions for all parameters in models I' and PS'.

Table B.4.: Bayes factors for ABC model comparison.

	I	I'	PS	PS'
I	1.00	0.09	2.94	0.09
I'	11.60	1.00	34.14	1.06
PS	0.34	0.03	1.00	0.03
PS'	10.99	0.95	32.34	1.00

Supplementary Information 5

Tajima's D values in the standard neutral model and I models with high migration rates

We compared 10,000 simulations of Tajima's D for a single locus with samples of size 142, 114 and 28 (modeling the whole, coastal and interior samples).

i) In a standard neutral coalescent model (no recombination) with a scaled mutation rate of $\theta_1 = 0.93$, $\theta_2 = 0.92$ and $\theta_3 = 0.84$ for the three samples (values are means of Watterson's estimates from the observed PUTs in the three samples).

ii) In a modified I model without recombination using scaled mutation rates $\theta_1/2$, $\theta_2/2$ and $\theta_3/2$ (halved to correct for theoretical population sizes) and scaled migration rates of either $m_{12} = 74.47$ and $m_{21} = 65.17$ or $m_{12} = 117.54$ and $m_{21} = 104.97$ (values are medians from the posterior distributions of models I and I').

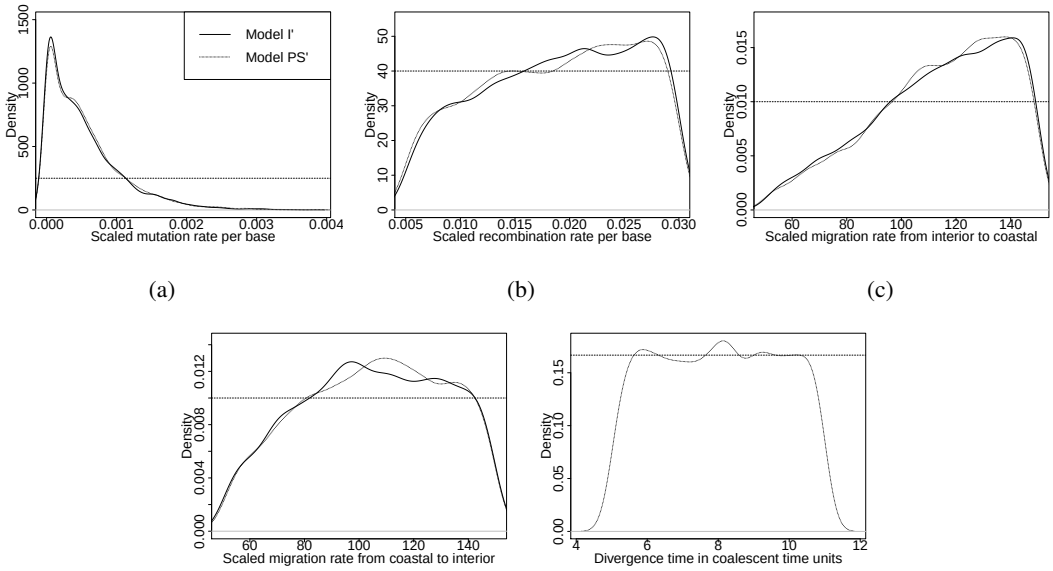


Figure B.16.: Posterior distributions of the (a) mutation rate θ, (b) recombination rate ρ, (c) migration rate from interior to coastal population, (d) migration rate from coastal to interior population, and (e) divergence time. Straight lines show posteriors for model I', dotted lines posteriors for model PS'.Prior distributions are added as dashed horizontal lines. All rates are scaled by $4N_e$ or $2N_e$ for divergence time. N_e is the population size of the underlying neutral Wright-Fisher models.

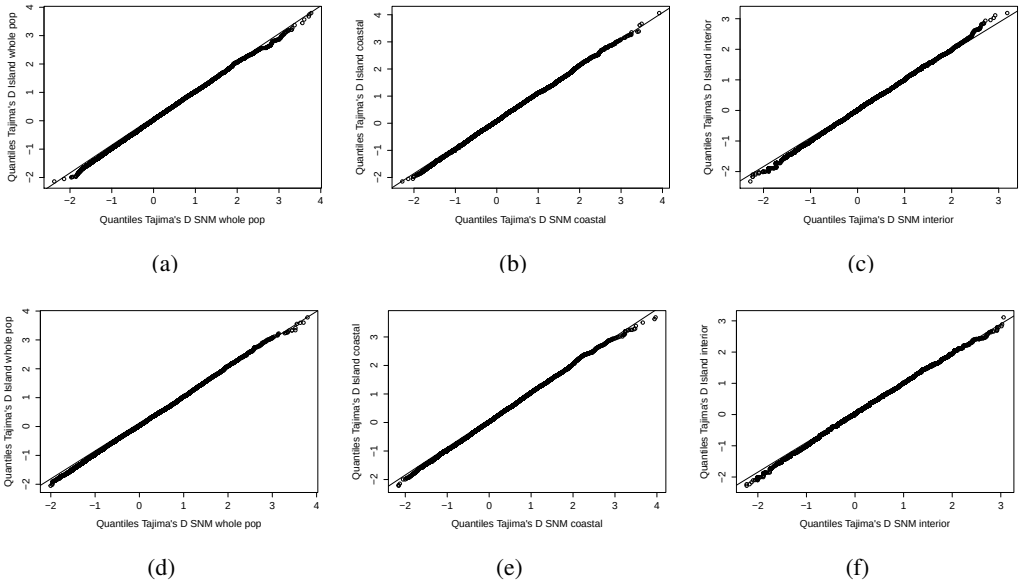
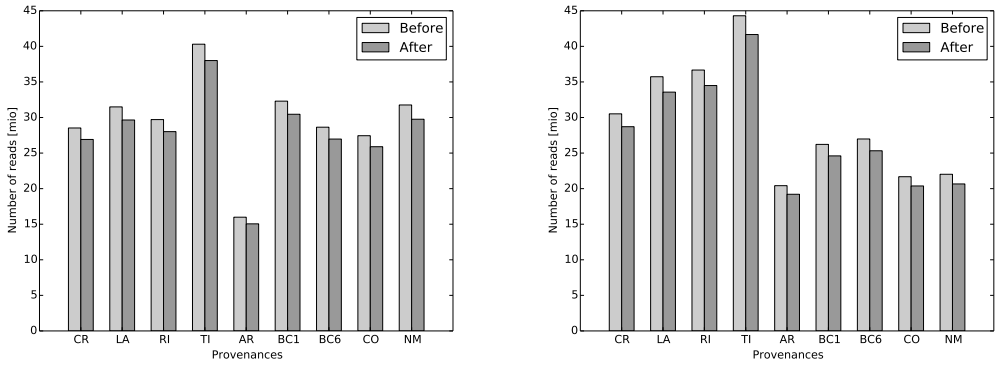


Figure B.17.: Q-Q plots for Tajima's D in the standard neutral model and in I models with high migration rates. (a)-(c) Island model with median migration rates from fitted model I, (d)-(f) Island model with median migration rates from fitted model I'.

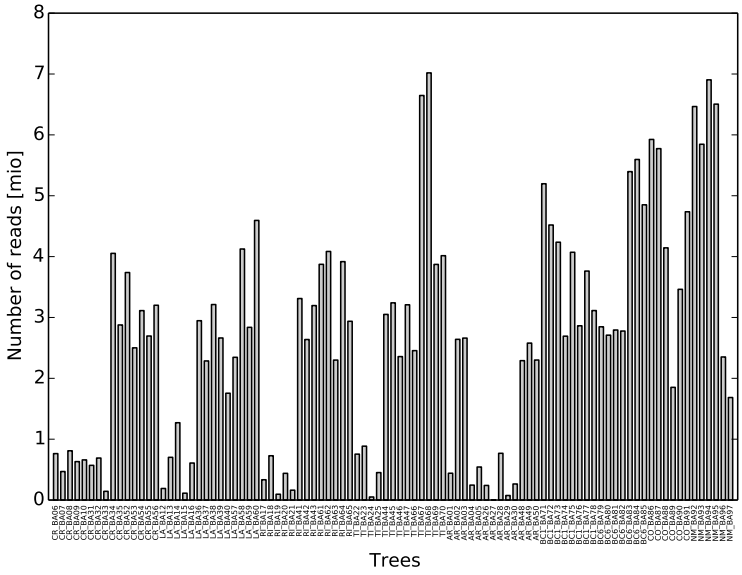
C. Comparison of genotyping-by-sequencing and sequence capture for population structure inference in Douglas-fir

Supplementary Figures

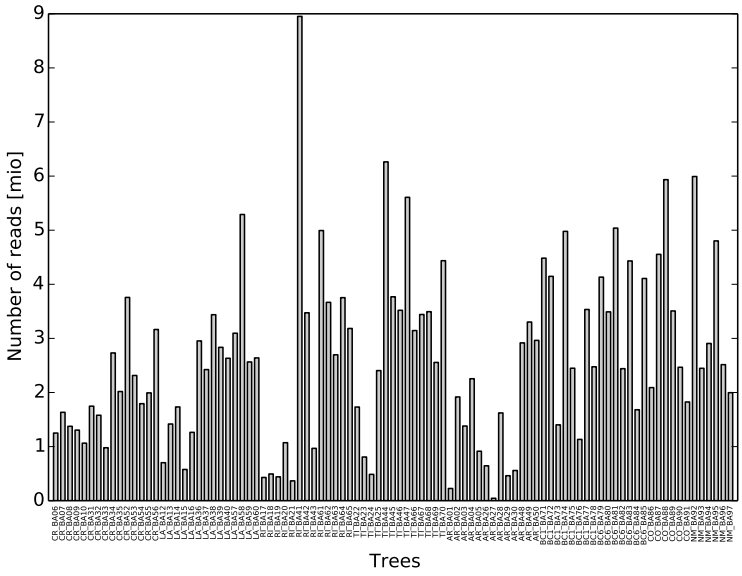


(a) (b)

Figure C.1.: Number of reads per provenance before and after preprocessing. (a) SD and (b) DD.



(a)



(b)

Figure C.2.: Number of reads per tree after preprocessing. (a) SD and (b) DD.

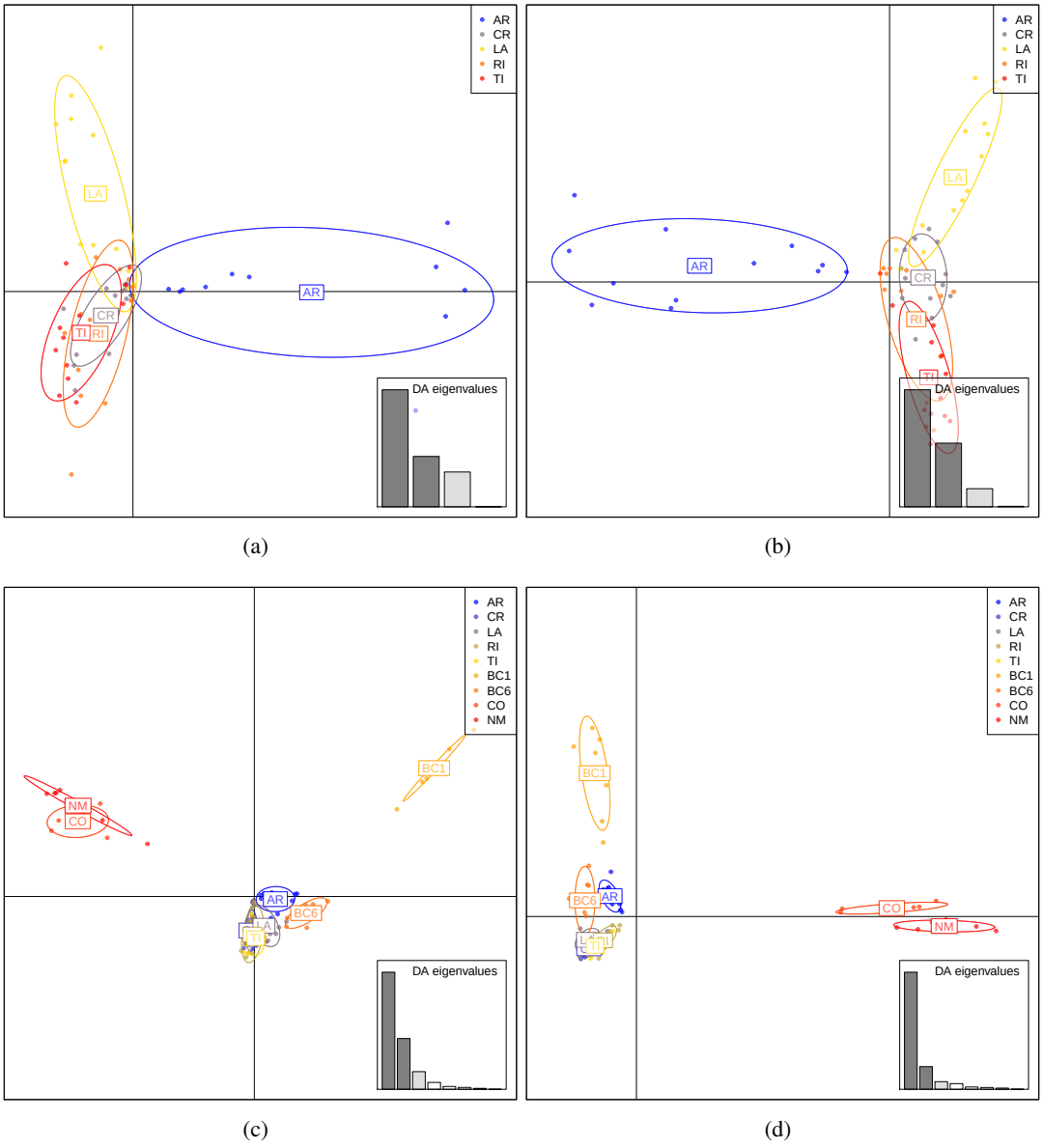
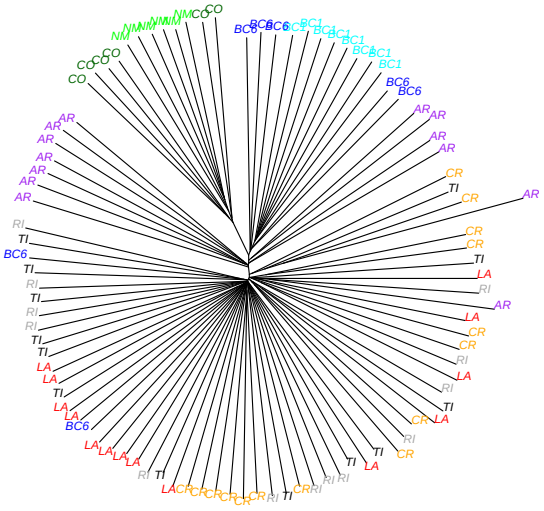
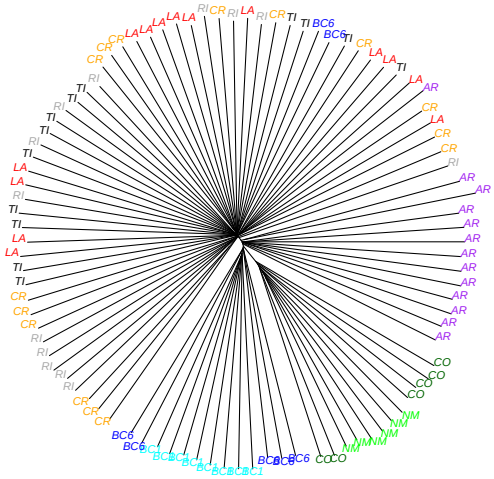


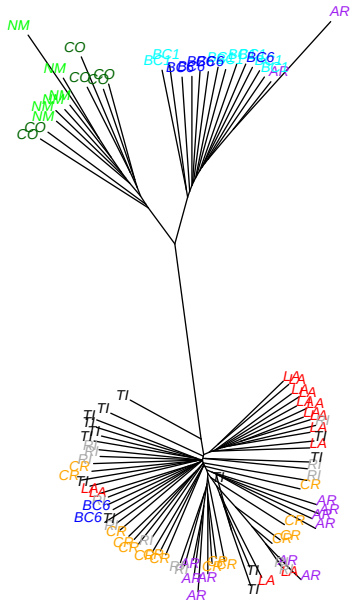
Figure C.3.: DAPC results of SDcap, DDcap, SDall, and DDall after removing libraries with less than 100,000 reads. (a) SDcap, (b) DDcap, (c) SDall, (d) DDall.



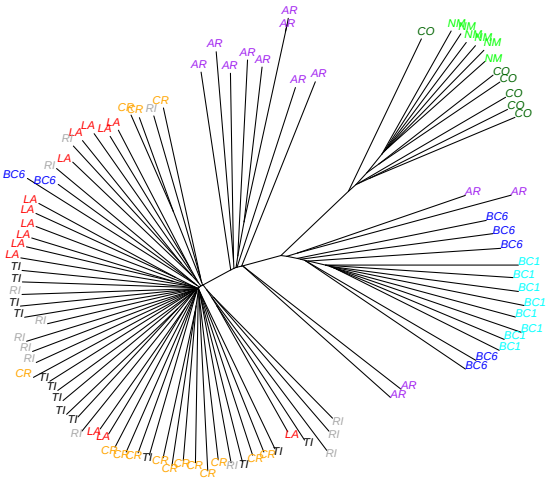
(a) SDall



(b) DDall



(c) STACKS_SD



(d) STACKS_DD

Figure C.4.: Neighbor-joining trees based on the genetic distances of the detected SNPs in (a) SDall, (b) DDall, (c) STACKS_SD and (d) STACKS_DD data.

Supplementary Tables

Table C.1.: Barcode IDs and sequences used in single and double digest GBS.

BA01	CTCC	BA26	CATCT	BA50	ACTGAA	BA74	TTCCGAA
BA02	TGCA	BA27	CCTAC	BA51	AGCCTT	BA75	AGCTTGT
BA03	ACTA	BA28	GAGGA	BA52	CTGAGA	BA76	CCACGCT
BA04	CAGA	BA29	GGAAC	BA53	GATACC	BA77	CTTGAAT
BA05	AACT	BA30	GTCAA	BA54	CGACAT	BA78	AACCACGT
BA06	GCGT	BA31	TAATA	BA55	TTCAGC	BA79	CTTGTTGA
BA07	CGAT	BA32	TACAT	BA56	AGTCGGT	BA80	AGGTCGGT
BA08	GTAA	BA33	TCGTT	BA57	CCTAAGA	BA81	TAACGAGA
BA09	AGGC	BA34	TTGTCA	BA58	TTCGTGA	BA82	GCCAACGT
BA10	GATC	BA35	AATGCT	BA59	ACGTGGT	BA83	CTGTTGGA
BA12	TGCGA	BA36	TTACGA	BA60	GGACAGT	BA84	TGAATCGT
BA13	CGCTT	BA37	GGCTAT	BA61	CACATGA	BA85	GACCATGA
BA14	TCACC	BA38	AATCGA	BA62	TGTTCTA	BA86	CGTTAGGT
BA15	CTAGC	BA39	CCGTAT	BA63	CTGAGGT	BA87	ACCATAGA
BA16	ACAAA	BA40	TTAGCC	BA64	GAAGTCA	BA88	TGTTCTGA
BA17	TTCTC	BA41	GGCATA	BA65	ACCGCAT	BA89	CTGGAGGT
BA18	AGCCC	BA42	AAGCAT	BA66	CATTGGT	BA90	ACCACGTT
BA19	GTATT	BA43	CTATGC	BA67	ACCTAGA	BA91	GAACAATA
BA20	CTGTA	BA44	TCCGCA	BA68	TGTCTCA	BA92	CTTATGAA
BA21	ACCGT	BA45	AGTATC	BA69	ATCGGTT	BA93	GCCACAAT
BA22	GCTTA	BA46	GAACCT	BA70	CCATGAA	BA94	CTGTGTTA
BA23	GGTGT	BA47	CCTTGA	BA71	GTTCCTA	BA95	TATAACGA
BA24	AGGAT	BA48	TGGACC	BA72	TGAACCA	BA96	GCACCATT
BA25	ATTGA	BA49	GAATTC	BA73	ACTGATT	BA97	CTTGGTAT

Table C.2.: Number of SNPs detected with *Stacks* using a single and a double digest GBS and different thresholds of allowed missing data points per SNP. STACKS_SD - results of *Stacks* using single digest GBS, STACKS_DD - results of *Stacks* using double digest GBS.

Missing data points per SNP	STACKS_SD	Total amount of missing data points after filtering in %	STACKS_DD	Total amount of missing data points after filtering in %
No filtering	697,616	85.81	984,472	86.71
$\leq$ 90%	341,780	75.95	421,992	75.33
$\leq$ 80%	185,999	67.74	219,462	65.58
$\leq$ 70%	88,496	58.89	109,844	55.35
$\leq$ 60%	39,446	50.33	60,315	46.80
$\leq$ 50%	14,542	40.37	31,333	38.48
$\leq$ 40%	5,621	31.59	15,543	31.19
$\leq$ 30%	1,874	22.40	5,648	23.21
$\leq$ 20%	640	15.79	1,496	15.31
$\leq$ 10%	31	8.59	194	6.20
No missing values	0	0	2	0

D. Declaration of contributions as co-author

In this thesis, I present the results of my doctoral research conducted since 2010. For all studies I developed and implemented analysis scripts. Large parts of the studies were conducted in collaboration with other scientists.

All papers of this thesis were supervised by Prof. Karl Schmid.

The data of the first study (Chapter 2, Müller et al, 2012) was analyzed by me. Greenhouse and experimental work were designed, conducted and coordinated by Ingo Ensminger; sequencing experiment was designed by Ingo Ensminger and Karl Schmid. I wrote the manuscript with contributions of Ingo Ensminger and Karl Schmid.

For the second study (Chapter 3, Müller et al, 2015a) I designed the experiment together with Karl Schmid. Furthermore, I conducted and coordinated the experimental work and analyzed the sequencing data. The ABC approach and tests against the neutral model were performed by Fabian Freund. I wrote the manuscript with contributions of Fabian Freund and Karl Schmid.

The study of the third paper (Chapter 4) was designed by Karl Schmid and me. The experimental work was conducted by Elisabeth Kokai-Kota. The data analyses were performed by me. The manuscript was written by me with contributions of Karl Schmid. The paper was submitted to Molecular Ecology Resources.

The chapters are identical to the content of the published papers except for minor formatting and typographical changes.

In addition to the work of this thesis, I contributed to further publications during my PhD, which are not part of this thesis. In a paper on barley, I assisted in SNP calling and contributed to the manuscript:

Bedada, G., Westerbergh, A., Müller, T., Galkin, E., Bdolach, E., Moshelion, M., Fridman, E. and Schmid, K.J. (2014). Transcriptome sequencing of two wild barley (*Hordeum spontaneum* L.) ecotypes differentially adapted to drought stress reveals ecotype-specific transcripts. BMC Genomics 2014, 15:995 doi:10.1186/1471-2164-15-995

In a submitted paper on cauliflower I conducted the preprocessing of raw reads and the SNP calling, assisted in the data analysis, and contributed to the manuscript:

Yousef, E.A., Müller, T., Börner, A. and Schmid K.J. (2015). Evidence for strong population

structure caused by germplasm regeneration in ex situ genebank collections of cauliflower *Brassica oleracea* var. *botrytis*. Manuscript submitted for publication.

In a study on amaranth I assisted in raw read processing, SNP calling, and data analysis:

Stetter, M., Müller, T. and Schmid, K.J. (2015). Genetic diversity and domestication of South American amaranth (*Amaranthus caudatus*) and its potential ancestors. Unpublished manuscript.

E. Eidesstattliche Versicherung

Eidesstattliche Versicherung gemäß § 7 Absatz 7 der Promotionsordnung der Universität Hohenheim zum Dr. rer. nat.

1. Bei der eingereichten Dissertation zum Thema

   Identification and analysis of a transcriptome of Douglas-fir (*Pseudotsuga menziesii*) and population structure inference using different next-generation sequencing techniques

   handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Ich habe nicht die Hilfe einer kommerziellen Promotionsvermittlung oder -beratung in Anspruch genommen.

4. Die Bedeutung der eidesstattlichen Versicherung und der strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Die Richtigkeit der vorstehenden Erklärung bestätige ich: Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Stuttgart, 6. Februar 2015 Thomas Müller

Eidesstattliche Versicherung, Belehrung

Die Universität Hohenheim verlangt eine Eidesstattliche Versicherung über die Eigenständigkeit der erbrachten wissenschaftlichen Leistungen, um sich glaubhaft zu versichern, dass die Promovendin bzw. der Promovend die wissenschaftlichen Leistungen eigenständig erbracht hat.

Weil der Gesetzgeber der Eidesstattlichen Versicherung eine besondere Bedeutung beimisst und sie erhebliche Folgen haben kann, hat der Gesetzgeber die Abgabe einer falschen eidesstattlichen Versicherung unter Strafe gestellt. Bei vorsätzlicher (also wissentlicher) Abgabe einer falschen Erklärung droht eine Freiheitsstrafe bis zu drei Jahren oder eine Geldstrafe.

Eine fahrlässige Abgabe (also Abgabe, obwohl Sie hätten erkennen müssen, dass die Erklärung nicht den Tatsachen entspricht) kann eine Freiheitsstrafe bis zu einem Jahr oder eine Geldstrafe nach sich ziehen.

Die entsprechenden Strafvorschriften sind in § 156 StGB (falsche Versicherung an Eides Statt) und in § 161 StGB (Fahrlässiger Falscheid, fahrlässige falsche Versicherung an Eides Statt) wiedergegeben.

§ 156 StGB: Falsche Versicherung an Eides Statt

Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

§ 161 StGB: Fahrlässiger Falscheid, fahrlässige falsche Versicherung an Eides Statt

Abs. 1: Wenn eine der in den §§ 154 und 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

Abs. 2: Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Absätze 2 und 3 gelten entsprechend.

Ich habe die Belehrung zur Eidesstattlichen Versicherung zur Kenntnis genommen.

Stuttgart, 6. Februar 2015 Thomas Müller

F. Curriculum vitae

Publications

Müller, T., Ensminger, I. and Schmid, K.J.: A catalogue of putative unique transcripts from Douglas-fir (*Pseudotsuga menziesii*) based on 454 transcriptome sequencing of genetically diverse, drought stressed seedlings, *BMC Genomics*, 2012, 13:673

Bedada, G., Westerbergh, A., **Müller, T.**, Galkin, E., Bdolach, E., Moshelion, M., Friedman, E. and Schmid K.J.: Identification of ecotype-specific genes in two differentially adapted and drought stressed wild barley *Hordeum spontaneum* (L.) ecotpyes by transcriptome sequencing, *BMC Genomics*, 2014, 15:995

Müller, T., Freund, F., Wildhagen, H. and Schmid, K.J.: Targeted re-sequencing of five Douglas-fir provenances reveals population structure and putative target genes of positive selection, *Tree Genetics & Genomes*, 2015, 11:816

Müller, T., Kokai-Kota, E. and Schmid, K.J.: Comparison of genotyping-by-sequencing and sequence capture for population structure inference in Douglas-fir, 2015, Manuscript submitted for publication

Yousef, E.A., **Müller, T.**, Börner, A. and Schmid K.J.: Evidence for strong population structure caused by germplasm regeneration in ex situ genebank collections of cauliflower *Brassica oleracea* var. *botrytis*, 2015, Manuscript submitted for publication

Stetter, M., **Müller, T.** and Schmid, K.J.: Genetic diversity and domestication of South American amaranth (*Amaranthus caudatus*) and its potential ancestors, 2015, Unpublished manuscript.

Stuttgart, 6. Februar 2015 Thomas Müller

G. Danksagung

Vielen Dank an alle, die auf die eine oder andere Art dazu beigetragen haben, dass diese Arbeit entstehen konnte.

Zunächst möchte ich mich bei meinem Betreuer Prof. Karl Schmid für die vergangene Zeit und die gute Zusammenarbeit bedanken. Und natürlich auch für die Finanzierung, die durch seinen erfolgreichen Antrag beim DFG für das DougAdapt Projekt sichergestellt wurde. Ein besonderer Dank auch an Prof. Andreas Schaller für die Übernahme der Zweitkorrektur und an Prof. Waltraud Schulze, die sich bereit erklärte als dritte Prüferin zur Verfügung zu stehen.

Many special thanks to all my co-authors. Thanks for ideas, suggestions, written parts, and of course proof reading.

Bei Elisabeth Kokai-Kota möchte ich mich für die Tipps und Hilfe im Labor bedanken. Bei Frau Lieb, Frau Meier und Frau Hessenauer bedanke ich mich für die Hilfe mit allen möglichen verwaltungstechnischen Sachen (Reisegenehmigung, Reisekostenerstattung, Arbeitsverträge, ...).

Allen derzeitigen und ehemaligen Mitarbeitern der Gruppe Nutzpflanzenbiodiversität und Züchtungsinformatik, sowie weiteren Institutsmitarbeitern, möchte ich für Diskussionen und Hilfestellungen, aber auch für abendliche Ablenkungen, durch Fußballspiele, Besuche des Stadion oder auch diverser anderer Lokalitäten in Stuttgart meinen Dank aussprechen: Patrick, Oli, Ivan, Torsten, Stefan, Dounia, Markus, Julie, Fabian, Christian, Anja, Tohamy, Wilmar, Anna, Wosene, Raul, Inka, Sariel, Zemach, Alex, Christian und vielen mehr.

Many thanks to Oli, Patrick, and Wosene: I had a nice time in the office with you guys with coffee-breaks and interesting and funny talks!

Ich möchte mich auch bei allen Mitgliedern des DougAdapt Projekts bedanken, besonders bei Moritz, Anita, Laura, Kirstin, Henning, Du, Muhidin und Bela. Ein Hoch auf Moskitospray und Regenklamotten!

Ann-Christin, Christian, Fabian, Patrick und Torsten, ein besonderer Dank Euch für das Korrekturlesen dieser Arbeit. Vielen Dank für Eure Tipps und Verbesserungsvorschläge.

Bei meinen Eltern und meiner Schwester möchte ich für Ihre anhaltende Unterstützung, auch während meines Studiums, bedanken.

And last but not least: Isa, vielen Dank für deine Unterstützung und einfach für alles!