Candidate genes for stress response in silver fir (*Abies alba* Mill.)

Dissertation

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

dem Fachbereich Biologie
der Philipps-Universität Marburg
vorgelegt von

David Thomas Behringer
aus Heidelberg

Marburg an der Lahn, 2017

Vom Fachbereich Biologie der Philipps-Universität Marburg
als Dissertation angenommen am 16.05.2017

Erstgutachterin: Prof. Dr. Birgit Ziegenhagen
Zweitgutachterin: Prof. Dr. Nina Farwig

Tag der mündlichen Prüfung am 01.06.2017

Contents

Summary **3**

Zusammenfassung **5**

1 General introduction **7**

2 Monitoring Plant Drought Stress Response Using Terahertz Time-Domain Spectroscopy **19**

3 A note on chlorophyll content potentially influencing *in-vivo* THz-TDS measurements **29**

4 Differential Gene Expression Reveals Candidate Genes for Drought Stress Response in *Abies alba* (Pinaceae) **37**

5 Past stress responses archived in tree-rings associate with SNP genotypes in *Abies alba* (Mill.) **57**

6 A critical note on random forest based feature selection and interaction analyses in genetic association studies **81**

7 Synthesis **95**

Bibliography **103**

Danksagung – Acknowledgments **109**

Appendix **111**

Data availability 111

Erklärung zum Eigenanteil 112

Eidesstattliche Versicherung 113

Summary

The aim of this thesis was the identification and analysis of candidate genes for stress response in silver fir (*Abies alba* Mill.). This ecologically and economically important forest tree species is native to many mountainous regions of Europe but little is known about its ecological characteristics. Silver fir populations were heavily transformed by human activity, which results in a mismatch between past and current distribution. Recent studies suggest that silver fir can occupy warmer and dryer climates than it currently does. However, the species also suffered considerably during the 1970s and 1980s, including foliar damage, radial growth depression and local diebacks in Germany. This is attributed mainly to the peak in air pollution during this period, especially sulfur dioxide (SO_2), which seems to heavily increase drought sensitivity in silver fir. The combination of both stressors, SO_2 and drought events, negatively affected silver fir even in regions where drought is usually not a problem.

In the context of anthropogenic global climate change that will very likely lead to an increase in temperature in Europe and to more extreme events such as severe drought periods, the question arises, how silver fir will cope with these environmental changes. Given the speed of the predicted changes and the increasing landscape fragmentation, silver fir might not be able to evade it via seed dispersal. As a sessile organism, the only other option is adaptation, which will likely draw from standing genetic variation.

To successfully predict the fate of silver fir, especially in the face of global climate change, and to potentially manage populations based on such predictions, the genetic architecture of silver fir in the context of such important stressors as drought and air pollution has to be understood. There exist, however, little genomic resources for silver fir and conifers in general. This is due to their large and complex genomes and the long generational cycle, which makes conifers typical non-model species. As such, methods for the identification of the genetic basis of stress response are effectively limited to a candidate gene approach.

The candidate gene approach includes the identification of functional candidate genes by measuring differential gene expression between a stressed and a control group. In the context of this thesis, the water content of silver fir seedlings was monitored in a laboratory using a novel terahertz spectroscopy setup. One group of seedlings was regularly irrigated while the other group was drought stressed. Continually measuring the water content allowed to harvest needles from both groups at a time when the water status was comparable between the individuals within each group. A differential expression analysis between the needles from both groups then revealed 296 genes that were significantly up- or down-regulated in response to drought stress. Of those genes, approximately 45% have not been previously described in any organism and are potentially unique to silver fir or conifers in general. However, since only needles of seedlings were analyzed at a specific level of drought stress, the results are limited in scope to the source material and stress intensity and cannot be directly applied to silver fir or drought stress in general. Also, this approach implies a cause-effect relationship between gene expression and a specific level of drought stress. Thus, it is very important that confounding factors are excluded from the experiment. Chlorophyll content in the needles, for example, might change over the course of the monitoring period due to the drought treatment. To test if the chlorophyll content could potentially influence the terahertz signal, chlorophyll was extracted from silver fir needles, in the course of this thesis, and different

concentrations were measured using terahertz spectroscopy, showing that chlorophyll content does not influence terahertz monitoring.

Another aspect of the candidate gene approach involves the variation within a polymorphic gene and its potential association with the variation in a phenotypic trait. Since the growth depression period of silver fir in the 1970s and 1980s was mostly influenced by the combination of air pollution and drought, in the context of this thesis, genetic variation, in the form of single nucleotide polymorphims (SNPs) in pre-selected genes, was associated with tree-ring derived phenotypes for individual trees in the Bavarian Forest National Park. These so called 'dendrophenotypes' were measures for resistance, resilience and recovery during the depression period, as well as the drought year 1976. Using general linear models and feature selection techniques based on the machine learning algorithm random forest, 15 out of 103 polymorphic candidate genes for trait variation could be identified. Since the associated dendrophenotyes are potentially adaptively relevant, the variation in this candidate genes could influence the stress coping capability of individual trees. However, this approach is of an observational nature and thus, cause-effect relationships cannot be derived from this type of experiment. The identified SNPs might be the causal variant or physically close to the true causal variant or it might just be a spurious correlation. Further, reliance on advanced statistical techniques can be troublesome, as could be demonstrated in the course of this thesis for a random forest based feature selection technique, developed for genetic association studies in conifers. Replicating this study and evaluating the algorithm, non-uniqueness of the results could be demonstrated, which not only hinders biological interpretation but can severely negatively influence downstream analyses, such as tests for interaction between SNPs.

In conclusion, this thesis presents new techniques to add to the current methodology for candidate gene selection and analysis in the stress response of the non-model organism silver fir and other conifer species. Both approaches should be combined, for example by drawing polymorphic candidate genes for trait variation from the pool of functional candidate genes to ensure the involvement of the studied genes in the variation of the trait of interest. Further, the results of this thesis add to the growing molecular resources in silver fir and thereby, hopefully, contribute to the successful prediction and management of this important forest tree species in the face of rapidly changing environmental conditions.

Zusammenfassung

Das Ziel dieser Dissertation war die Identifizierung und Analyse von Kandidatengenen für Stressreaktion in der Weißtanne (*Abies alba* Mill.). Diese ökologisch und ökonomisch wichtige Waldbaumart ist natürlich beheimatet in bergigen Regionen Europas aber es ist wenig bekannt über ihre ökologische Charakterisierung. Weißtannenpopulationen wurden durch menschlichen Einfluss deutlich transformiert, was zu einer Diskrepanz zwischen dem früheren und dem heutigen Verbreitungsgebiet geführt hat. Aktuelle Studien legen nahe, dass die Weißtanne in wärmeren und trockeneren Klimaten ansässig sein kann, als sie es derzeit ist. Die Art hat jedoch in den 1970ern und 1980ern deutlich gelitten und zeigte in Deutschland Blattschäden, eine Abnahme im Dickenwachstum und örtliches Waldsterben. Als Grund hierfür wird meist der Höchststand der Luftverschmutzung während dieses Zeitraums genannt, insbesondere von Schwefeldioxid (SO_2), das offenbar die Empfindlichkeit der Weißtanne gegenüber Trockenstress deutlich erhöht. Die Kombination dieser beiden Stressoren, SO_2 und Trockenperioden, hat die Weißtanne sogar in Regionen in denen Trockenstress üblicherweise kein Problem ist negativ beeinflusst.

Im Kontext des anthropogenen globalen Klimawandels, der sehr wahrscheinlich zu einem Ansteigen der Temperatur in Europa und zu mehr Extremereignissen, wie heftigen Trockenperioden, führen wird, stellt sich die Frage, wie die Weißtanne mit diesen Umweltveränderungen umgehen wird. Bedenkt man die Geschwindigkeit der vorhergesagten Veränderungen und die zunehmende Fragmentierung der Landschaft, besteht die Möglichkeit, dass die Weißtanne diesem nicht durch Samenausbreitung entgehen kann. Als sessiler Organismus bleibt als die einzig andere Option nur Adaptation, die sich wahrscheinlich aus der stehenden genetischen Variation speisen wird.

Um das Schicksal der Weißtanne, insbesondere im Kontext des globalen Klimawandels, erfolgreich abzuschätzen und Populationen möglicherweise aufgrund solcher Vorhersagen zu managen, bedarf es der Kenntnis der genetischen Architektur im Kontext solcher bedeutender Stressoren wie Trockenperioden und Luftverschmutzung. Es gibt jedoch sehr wenige genomische Ressourcen für die Weißtanne und Koniferen im Allgemeinen. Das liegt maßgeblich an der Größe und Komplexität der Genome und am langen Generationszyklus, was Koniferen zu typischen nicht-Modell Organismen macht. Aus diesem Grund sind Methoden zur Identifizierung der genetischen Basis von Stressantwort effektiv auf einen Kandidatengenansatz beschränkt.

Der Kandidatengenansatz beinhaltet die Identifikation von funktionellen Kandidatengenen, indem die differentielle Genexpression zwischen einer gestressten und einer Kontrollgruppe gemessen wird. Im Kontext dieser Dissertation wurde der Wassergehalt von Weißtannensämlingen mit einem neuartigen Terahertz-Spektroskopie-Aufbau in einem Labor überwacht. Eine Gruppe von Sämlingen wurde regelmäßig gegossen, während eine andere Gruppe Trockenstress ausgesetzt war. Durch die kontinuierliche Messung des Wassergehalts konnten Nadeln der Sämlingen aus beiden Gruppen zu einem Zeitpunkt geerntet werden, zu dem der Wassergehalt zwischen den Individuen einer Gruppe jeweils vergleichbar war. Eine differentielle Expressionsanalyse zwischen den Nadeln der beiden Gruppen resultierte dann in 296 Genen die als Reaktion auf Trockenstress signifikant hoch- oder herunter-reguliert waren. Ungefähr 45% dieser Gene sind zuvor noch nicht in anderen Organismen beschrieben worden und sind potentiell spezifisch für die Weißtanne oder Koniferen im Allgemeinen. Da jedoch nur Nadeln von Sämlingen bei einem bestimmten Trockenstressniveau analysiert wurden, sind die Ergebnisse in ihrem Umfang auf das Ausgangsmaterial und das spez-

ifische Stresslevel reduziert und können nicht direkt auf die Weißtanne oder Trockenstress allgemein übertragen werden. Weiterhin impliziert dieser Ansatz einen Kausalzusammenhang zwischen Genexpression und einem spezifischen Maß an Trockenstress. Daher ist es sehr wichtig störende Faktoren vom Experiment auszuschließen. So kann der Chlorophyllgehalt der Nadeln sich beispielsweise während des Messzeitraumes in Folge der Trockenstressbehandlung verändern. Um zu testen ob der Chlorophyllgehalt möglicherweise einen Einfluss auf das Terahertz-Signal hat, wurde im Rahmen dieser Dissertation Chlorophyll aus Weißtannennadeln extrahiert und unterschiedliche Konzentrationen mittels Terahertz-Spektroskopie gemessen. Dabei konnte gezeigt werden, dass der Chlorophyllgehalt keinen Einfluss auf das Terahertz-Monitoring hat.

Ein anderer Aspekt des Kandidatengenansatzes beinhaltet die Variation innerhalb eines polymorphen Gens und die mögliche Assoziation mit der Variation in einem phänotypischen Merkmal. Da die Periode der Wachstumsdepression von Weißtannen in den 1970ern und 1980ern maßgeblich durch eine Kombination von Luftverschmutzung und Trockenperioden verursacht war, wurde im Rahmen dieser Dissertation genetische Variation, in Form von Einzelnukleotid-Polymorphismen (*engl.* single nucleotide polymorphisms (SNPs)) in vorausgewählten Genen, mit Phänotypen assoziiert, die aus Jahresringen für individuelle Bäume im Nationalpark Bayerischer Wald abgeleitet wurden. Diese so genannten 'Dendrophänotypen' waren Maße für die Resistenz, Resilienz und Erholung in der Depressionsperiode und dem Trockenjahr 1976. Basierend auf allgemeinen linearen Modellen und Feature Selection Techniken, die auf dem maschinellen Lernalgorithmus Random Forest beruhen, konnten 15 aus insgesamt 103 polymorphen Kandidatengenen für Merkmalsvariationen identifiziert werden. Da die assoziierten Dendrophänotypen potentiell adaptiv relevant sind, könnte die Variation in diesen Kandidatengenen die Fähigkeit der Stressbewältigung individueller Bäume beeinflussen. Dieser Ansatz ist jedoch grundsätzlich beobachtender Natur und diese Art von Experiment erlaubt daher keine Ableitung von Kausalzusammenhängen. Die identifizierten SNPs können die ursächliche Variation sein, sie können aber auch physikalisch nah an der tatsächlich ursächlichen Variation sein oder es kann sich lediglich um einen Scheinzusammenhang handeln. Weiterhin kann das Vertrauen in fortgeschrittene statistische Verfahren problematisch sein, was im Rahmen dieser Dissertation für eine auf Random Forest basierende Feature Selection Methode gezeigt werden konnte, die für genetische Assoziationsanalysen in Koniferen entwickelt wurde. Durch die Replikation dieser Studie und die Evaluierung des Algorithmus konnte die Multiplizität der Ergebnisse demonstriert werden, die nicht nur die biologische Interpretation behindert, sondern auch nachgelagerte Analysen, wie Tests auf Interaktion zwischen SNPs, negativ beeinflusst.

Schlussfolgernd beschreibt diese Dissertation neue Techniken der Auswahl und Analyse von Kandidatengenen für die Stressreaktion im nicht-Modell-Organismus Weißtanne und anderen Koniferenarten die der gängigen Methodik hinzuzufügen sind. Beide Ansätze sollten kombiniert werden, beispielsweise indem polymorphe Kandidatengene für Merkmalsvariation aus dem Pool von funktionellen Kandidatengenen gezogen werden um die Beteiligung der untersuchten Gene an der Variation des zu untersuchenden Merkmals sicher zu stellen. Weiterhin tragen die Ergebnisse dieser Dissertation zu den wachsenden molekularen Ressourcen für die Weißtanne bei und haben dadurch, hoffentlich, einen Anteil an der erfolgreichen Vorhersage und am Management dieser wichtigen Baumart im Kontext rasanter Umweltveränderungen.

CHAPTER 1

General introduction

Forest trees – importance and threats

Forests are one of the most important ecosystem service providers on this planet. Among others, they host biological diversity, act as carbon sinks and ensure clean water supply, as well as relative climate stability (Foley et al., 2005). However, since they also provide humans with timber and often have to give way for agricultural use or human settlements, land-use has a large negative net impact on global forest change, especially in the tropics and subtropics (Hansen et al., 2013). Temperate forests, especially in Europe, on the other hand, fare generally well with an increase and stagnation in biomass and area since 1950 (Foley et al., 2005). This, however, does not necessarily imply increase or stagnation in biodiversity or species composition (Hobbs et al., 2006). For Central European forests, for example, only 0.2% are estimated to be undisturbed by human activity (Hannah et al., 1995). This includes the introduction of exotic species, overgrazing by domestic herbivores and planting of monocultures with short rotation coppice (Bengtsson et al., 2000). Thus, while maintaining biomass and area, Central European forests are often heavily transformed, which frequently is accompanied by a reduction in biodiversity. For example, the rare occurrence of deadwood and old trees in most managed forests has negative impacts on the abundance and diversity of associated organisms, such as fungi, invertebrates, bats and birds (Krajick, 2001).

While land-use is relatively unidirectional and thus can be potentially shaped by management strategies - e.g. sustainable forestry in Central Europe with the scope to increase and sustain biodiversity and ecosystem function - global climate change poses a more diffuse threat. Some aspects are relatively certain, such as an increase in global mean surface temperature and a decrease of renewable surface water and groundwater (IPCC, 2014). Due to the inertia of large systems, these effects will occur in the short term regardless of mitigation policies and efforts. Long term effects beyond the year 2100, on the other hand, can still be influenced. On a more regional scale, the picture gets more diverse. For Europe, high-temperature extremes are occurring more and low-temperature extremes less frequently since 1950 (Kovats et al., 2014). For the future this means warmer winters for Northern Europe and warmer summers for Southern Europe. Further, Europe shows a decreasing gradient in mean annual precipitation from North to South.

Climate change can have varying effects on forests and they do not necessarily have to be negative. In Northern and Central Europe, for example, warmer winters might prolong the growing season, which could further be enhanced by higher CO_2 availability (Allen et al., 2010). However, in Southern Europe and especially the Mediterranean area, tree populations are dying-off, mainly due to extreme heat waves and drought. Even if Central and Northern European populations might, in general, not be negatively affected in the near future, considering humanities sluggish actions in mitigating global climate change, these populations will likely face similar problems in a more distant future. As somewhat of a precursor, mortality and die-offs of more northern tree populations of *Quercus* and *Picea* have already been linked to a combination of summer drought and biotic stresses (Allen et al., 2010). Even in the absence of a change in precipitation, forest mortality might still be caused (Barber et al., 2000) or accelerated (Adams et al., 2009) by drought stress due to rising temperatures.

In any case, in Europe, climate change will lead to more extreme events, such as droughts, heat waves and heavy precipitation (Kovats et al., 2014). This poses a threat to all forests, since trees are sessile organisms and have to cope with any sudden change in the environment. As

such, tree populations are facing challenges as a consequence of global climate change that can be viewed according to the severity of the change over time. Since trees cannot migrate to evade threatening conditions, their only 'spatial escape' lies in seed dispersal, which could be described as migration over generations. This, however, necessitates that the change in conditions is not too sudden and severe. Any abrupt change will out-pace the generational cycle and affect the entire population. Given that the post-glacial re-colonization speed of trees, on average, was probably around 100 m per year or less (Aitken et al., 2008; Loarie et al., 2009) and that, very broadly speaking, temperature is projected to change at a velocity of around 110 m, 260 m and 350 m per year for temperate coniferous forests, the Mediterranean area and temperate broadleaf and mixed forests, respectively (Loarie et al., 2009), climate change could be too fast for some tree populations in Europe and will likely favor species with long dispersal ranges. For many species, however, seed dispersal is often limited by natural barriers, such as mountains, and habitat fragmentation.

This leaves, as a last resort, adaptation. Again, any change that is very sudden and severe might out-pace the population's ability to adapt. Just as seed dispersal, adaptation represents a migration over generations, with the difference that the former is staged on a three-dimensional spatial landscape, while the latter is staged on a multidimensional fitness landscape (Orr, 2005). The question then becomes if a population harbors enough standing variation and has enough time for an adaptive walk up the slope of a new fitness peak to cope with relatively sudden changes in the environment. Standing variation could allow for rapid adaptation since beneficial alleles might already be present in the population (Barrett and Schluter, 2008). The alternative would be newly arisen mutations but their stochastic nature makes them a gamble and they have never been 'tested' by natural selection.

Adaptation is certainly one of the key elements of future survival for many trees since their populations are often natural and harbor a large amount of genetic variation (Neale and Kremer, 2011). Given that an increase in extreme events, especially drought and heat waves, are likely to occur in Europe as a consequence of climate change and that such events can be greatly amplified in severity by other agents, such as pathogens and herbivores (Ayres and Lombardero, 2000), or air pollutants such as ozone (Karnosky et al., 2005) or sulfur dioxide (Elling et al., 2009), we need to understand the molecular basis of adaptation and the contribution of different genes and gene variants to different phenotypes. Only then will we be able to devise management strategies that include the necessary genetic component, as well as construct models to successfully predict the impact of climate change and different land-use practices.

Study species *Abies alba*

Coniferous forests make up approximately 39% of the worlds forests (Armenise et al., 2012) and conifers are a dominant component of many forests around the world (Torre et al., 2014). They are also the most important source for wood and fiber (Ahuja and Neale, 2005). Most conifers share common characteristics that differentiate them substantially from angiosperms, such as large genome sizes (1C = 6,500 Mb to 37,000 Mb) (Ahuja and Neale, 2005) and a large amount of transposable elements (Torre et al., 2014). Given the importance of conifers, there is great interest to add to the growing resources and develop new methodologies and techniques to understand

their highly complex genetic architecture.

Silver fir (*Abies alba* Mill.) is an ecologically and economically important, evergreen coniferous species belonging to the family Pinaceae. Indigenous populations are mainly distributed along mountainous regions in Eastern, Western, Southern and Central Europe. In higher altitudes, silver fir is mostly associated with spruce (*Picea abies* (L.) Karst.), otherwise with beech (*Fagus sylvatica* L.). It can grow more than 60 m in height, making it the tallest European tree species, and can live more than 500 years (Tinner et al., 2013). Silver fir is wind-pollinated, predominantly outcrossing, monoecious and diploid with 2 n = 24 chromosomes. The genome size of silver fir can be approximated by the DNA content. With a 1C content of around 16.55 pg (Roth et al., 1997) and a 2C content of around 34.58 pg (Puizina et al., 2008), this corresponds to a 1C value between 16,000 Mb and 16,900 Mb (Doležel et al., 2003). There exist very little genomic resources for silver fir and conifers in general. To this date, the only available genome assemblies available in public repositories are for members of the genus *Picea* (Birol et al., 2013; Nystedt et al., 2013) and *Pinus* (Neale et al., 2014; Zimin et al., 2014).

Drawing from palaeobotanical and genetic studies, glacial refugia and the subsequent re-colonization of Europe could be reconstructed for silver fir (Liepelt et al., 2009). However, the natural potential range of the species is largely unknown, mainly due to a mismatch of previous (over 5,000 years ago) and present distributions who are heavily influenced by human activities (Tinner et al., 2013). Current models suggest that silver fir can withstand higher summer temperatures of up to 5-7°C and dryer conditions, given that precipitation does not fall below 700-800 mm per year.

Uncertainty about the ecological characteristics of silver fir lead to the general description of the species as very fragile and vulnerable to summer drought. Some went so far as to call silver fir the mimosa of the forest (Elling et al., 2009). This drought sensitivity could indeed be shown at the southern margin of the species distribution, for example in the form of diebacks in the Mediterranean area (Nourtier et al., 2012). In Central Europe, on the other hand, silver fir only had a short period of local diebacks in the 1970s and 1980s. These diebacks were part of the so-called 'novel forest decline' ('neuartige Waldschäden') and are largely attributed to air pollution, especially sulfur dioxide (SO_2) (Elling et al., 2009). Prior to this period, silver fir was considered rather drought tolerant in Central Europe, due to its deep rooting system. While older studies refute the involvement of SO_2 in the novel forest decline (Krause et al., 1986), more recent studies strongly suggest that SO_2 is the major factor causing diebacks in silver fir during this period (Elling et al., 2009). Furthermore, SO_2 seems to increase the sensitivity to drought events in silver fir, leading to a very different reaction of silver fir during periods of water shortage, depending heavily on the presence of SO_2 and consequently explaining the alleged drought sensitivity.

Given that silver fir has the potential to cope with a warmer climate, forest management strategies could increasingly use it to stabilize local tree communities and ensure sufficient timber supply in the face of climate change. However, there is considerable need for a better understanding of the response of silver fir to different stressors and especially their interaction. As previously mentioned, silver fir shows a high susceptibility towards SO_2 pollution and the diebacks may have decreased the genetic diversity of populations (Wolf, 2003). Even if silver fir is drought tolerant, future extreme events will likely co-occur with other stressors, such as pathogens or pollutants, and could, in combination, have detrimental consequences.

The focus of this thesis will thus lie on techniques and methods to further our understanding of the genetic components involved in the response of silver fir to such important stressors as drought and air pollution.

The complex nature of stress response

Changes in the environment can create potentially unfavorable conditions for plants and as such constitute a stress (Levitt, 1980). The concept of stress is very diverse and is often used in a different context. Within the scope of this thesis I will follow the above stated, general definition by J. Levitt.

As opposed to most animals who can detect a stress and avoid it by moving, the sessile nature of plants forces them to cope with any stressor. Plants have to endure a stress and will generally go through four basic phases (Lichtenthaler, 1998). Without any stress, a plant should be in a physiological standard situation which is optimized based on the specific environmental conditions, such as water supply, light and nutrient availability. Upon onset of stress, a plant will transition from the pre-stress phase into a (1) response phase. This transition includes deviation from the physiological standard, such as changes in the metabolism which usually means a decrease in anabolism and an increase in catabolism. The net effect is reduced growth and, depending on the severity of the stress and the plant's resistance, increased senescence and acute damage. Should the stress continue, the plant will enter the (2) restitution phase. Here, the plant will activate metabolic pathways for general and specific stress response and will acclimate by shifting the physiological standard to cope with the changed conditions. Restitution is highly dependent on the plant's resistance to the specific stressor. Depending on the severity of a continuing stress, a plant with a high resistance maximum can endure and even repair damages. However, should the severity increase or the duration of the stress be too long, the plant will enter the (3) end phase which is characterized by exhaustion. Within this phase, the plant will accumulate chronic damage and eventually cell death. Given that the stress is not removed too late for the plant to recover, it will enter the (4) regeneration phase. Depending on the timing of stress release relative to the exhaustion, the plant will shift to a new post-stress physiological standard.

In temperate forests, stress is a common occurrence. Water and nutrient availability, as well as sunlight and temperature vary throughout the seasons. Within their adaptive capacity, trees should be able to cope with these episodic stress periods. Problems arise when a stress increases in severity and/or duration and thereby exceeds the coping capability of an individual. However, a stress does not necessarily have be more severe and prolonged to cause problems. While resistance to singular, and especially episodic stress events is often sufficient in tree populations, a combination of different stresses can severely threaten a population. Multiple stresses can either occur successively or at the same time. This might lead to interaction effects between the stresses that are either positive or negative with, e.g., temperate woody species showing a generally low tolerance for multifactor stresses (Niinemets, 2010).

Given that global climate change will lead to rising temperatures and consequently more severe droughts and heat waves in Europe, water shortage is certainly one of the most important stressors for silver fir. Drought stress can occur both during winter and summer but is not necessarily a

singular event. Often other stressors, such as air pollution, co-occur with drought and can enhance the sensitivity of silver fir towards drought stress (Elling et al., 2009).

How can we identify genes that are potential candidates for the involvement in the stress response of silver fir and trees in general?

Identification of stress related genes – The candidate gene approach

In order to identify genes that are involved in the stress response of any organism, the phenotypic reaction to stress has to be quantified. These phenotypic traits are often of a quantitative nature and represent the measurable effect of multiple genes or gene variants (Box 1). As such they can be utilized to identify genes that are involved in stress response.

Box 1. Discrete vs. quantitative traits

Discrete traits are phenotypes like a disease state or flower color in peas and can be inherited in a monogenic Mendelian fashion (Lander and Schork, 1994).

Quantitative traits are phenotypes that are often, but not necessarily, measured on a continuous scale, such as growth, height or gene expression, i.e. the abundance of a transcript (Rockman and Kruglyak, 2006). Quantitative traits are 'complex' traits that are influenced by multiple polymorphic genes, so called quantitative trait loci (QTLs), and do not follow monogenic Mendelian inheritance (Lander and Schork, 1994). Individual QTLs, on the other hand, do follow classical Mendelian segregation and linkage (Frankham and Weber, 2000).

The majority of the information regarding the molecular response of plants to stress, and especially drought stress, comes from experiments conducted on model organisms, such as *Arabidopsis thaliana*, rice (*Oryza sativa*) or maize (*Zea mays*) (e.g. Ingram and Bartels, 1996). Often, the goal is to identify drought tolerant variants in crop plants to potentially breed more sturdy and yielding lines. The benefit of model organisms is the availability of the entire genome, which allows for genome-wide association studies (GWAS) who scan the entire genome and do not need any pre-selection of a genomic region. Combined with the ability to cross different clonal strains and track the inheritance of multiple loci through the generations, even quantitative traits can be successively analyzed in model organisms (e.g. Mauricio, 2001; Davila Olivas et al., 2017), granted that a distinction of causative genetic variants and linked neutral markers will remain problematic, even in model organisms (Korte and Farlow, 2013).

Silver fir, on the other hand, poses a major challenge. Not only is it a non-model species with little available genomic resources (Neale and Kremer, 2011), as a conifer, it also has a very large genome size (Murray, 1998). Further, tracking loci through pedigrees and crosses is often unrealistic, due to the long generational cycle. First flowering occurs in 25-35 year old individuals, given they are isolated (Wolf, 2003). Within a forest, silver fir trees first flower between the age of 60 and 70 years.

This effectively restricts the availability of possible methods for identifying genes or gene variants that are involved in stress response to a candidate gene approach. In contrast to GWAS, the candidate gene approach focuses on a subset of predefined genes. However, the term candidate gene has a different meaning in different disciplines of biology. In physiology, candidate genes are defined as genes whose expression is linked with a specific trait (Pflieger et al., 2001). In genetics, on the other hand, candidate genes are polymorphic genes that are possibly associated with the variation in a given trait. Yet, both approaches share some common ground. While geneticists are interested in the contribution of a given gene's variation to a trait variation, the selection of these polymorphic candidate genes should be based on the biological function of the gene, i.e. the candidate gene should have a functional consequence (Tabor et al., 2002).

Functional candidate genes

Functional candidate genes can be identified by measuring gene expression in a case control setting and this has worked rather well for drought related genes (e.g. Ingram and Bartels, 1996). Not only can this approach be used in non-model organisms, given that no target genes have to be preselected, it can also reveal novel candidate genes. The method targets the transcriptome, i.e. the part of the genome that is translated into amino acids, namely messenger RNA (mRNA; at least this is mainly the focus, however, the transcriptome consists of all transcripts, which includes mRNAs, non-coding RNAs and small RNAs (Wang et al., 2009). Gene expression is measured by extracting RNA from a tissue, reverse transcribing it into complementary DNA (cDNA) and determining the amount of transcripts in the tissue by either targeting selected cDNA sequences (e.g. microarray technology), sequencing the entire cDNA (e.g. RNA-Seq technology) (Wang et al., 2009; Guo et al., 2013), or deep sequencing a reduced representation of the cDNA (e.g. Massive Analysis of cDNA Ends (MACE)) (Kahl et al., 2012).

While gene expression is a very useful trait to identify functional candidate genes, setting up an environment that excludes all but the desired factors can be challenging for trees. Since these types of studies try to unveil a cause-effect relationship, they warrant an experimental approach (Box 2). Forest trees such as silver fir can hardly be moved into a laboratory and setting up a controlled environment around a local stand is unrealistic and will always exhibit gradients in one environmental factor or another. A few year old seedlings, however, are small enough to be put in a laboratory in a sufficient enough number to ensure proper replication. This leaves the problem of controlling the level of stress in a control and a treatment group that is both comparable within groups and sufficiently different between groups. Overcoming this last challenge results in functional candidate genes for a specific stress response.

Polymorphic candidate genes for complex trait variation

Identifying polymorphic candidate genes for complex trait variation can be conducted using different methods (Box 3). Given the constraints in species like silver fir, however, only genetic association is a viable option since no pedigrees are necessary.

Genetic association involves measuring some phenotye and associating the trait variation with the variation in a set of pre-defined genes. As such, this constitutes an observational study (Box 2).

Box 2. Study designs in the field of biology

Experimental studies usually test the effect of one or multiple treatments against a control, with the treatments being randomly assigned by the researcher. In order to establish a cause-and-effect relationship, the goal is to exclude as many confounding factors as possible. For this purpose, experimental studies are often conducted in a highly controlled environment such as a laboratory. Since biology studies living organisms, genetic variation has to be controlled for as well. This is often achieved by using cloned individuals with nearly identical life histories to avoid differences in gene expression and regulation.

Observational studies are fundamentally different in that the researcher has no control over the assignment of treatments. Hence, such studies cannot reveal cause-and-effect relationships but only hint at associations because there are usually a number of confounding factors present. Observational studies are often the only realistic type of experiment available. Forest ecosystems, for example, are very complex but often have to be studied by observation. Great efforts are made to develop mathematical and statistical techniques to account for confounding factors, such as genetic population structure and environmental variability.

For a genetic association, genes with known variation in a population have to be pre-selected. The most common type of variation investigated in modern studies are single nucleotide polymorphisms (SNPs). For diploid organisms, such as silver fir, this means that within a population most individuals have a certain nucleotide (either adenine, guanine, cytosine or thymine) at a specific locus on a chromosome while some individuals have another nucleotide on the exact same position on their chromosome. Depending on the effect of the variation on the resulting protein of a SNP within the coding region of a gene, the SNP is either synonymous (both nucleotides lead to the same amino acid) or non-synonymous (one of the nucleotides changes the amino acid).

Genetic association within an observational study introduces the challenge to control as many confounding factors as possible and to select a precise and reliable phenotype. Also, genetic structure has to be accounted for, since any phenotype that occurs, due to heritage, more frequently in one of multiple groups within the sampled population will often coincide with genetic differences. A significant association between genetic and phenotypic variation could thus be solely attributable to population stratification (Lander and Schork, 1994). Given that population stratification is controlled for, the trait associated variation (e.g. a SNP in a gene) is likely to be physically close to, or is itself, the functional variant (Neale and Kremer, 2011). The reason for this is the rapid decay of linkage disequilibrium (LD) in most tree populations.

Box 3. Methods for the genetic dissection of complex traits (Lander and Schork, 1994)

Linkage analysis tracks two loci (one being associated with a phenotype - mostly a disease state - and the other being proposed to be physically linked to the former) through pedigrees and tests if both loci are inherited together more often than expected by chance. This method is mostly used for simple Mendelian traits since the Null model for complex traits can be very hard to propose.

Allele-sharing methods are based on proving that chromosomal regions do not follow random Mendelian segregation and thus are probably associated with a corresponding phenotype. This method is more robust than linkage analysis since it does not assume any model of inheritance.

Association studies do not track loci through pedigrees but simply test if an allele at a gene occurs more often with one variant of a discrete trait than expected by chance. For quantitative traits the association is tested by comparing mean phenotypic variation between alleles. This method has some limitations regarding the interpretation of the results. A significant association can either be due to a cause-effect relationship between allele and phenotype, a linkage between allele and true causal locus or, in the case of population stratification, a spurious relationship.

Experimental crosses map polygenic traits in pedigrees (including QTLs). QTL mapping is a powerful tool, especially for systems in which mutations can be specifically bred. In principle, QTL mapping is a possibility in conifers, given that progenies of crosses are available (Mauricio, 2001). However, since any QTL contains numerous loci and genes that might or might not be associated with the trait under investigation, intensive fine-mapping is necessary to identify specific genes. Thus, due to their large genome sizes, QTL mapping can currently not lead to the identification of specific genes for a trait variation in conifers (Neale and Kremer, 2011)

Goal of the thesis

The goal of this thesis was the identification of functional candidate genes for drought stress response in silver fir seedlings in a novel, highly controlled experimental setup, as well as the association of variation in pre-selected, polymorphic candidate genes with variation in novel phenotypes derived from tree-rings of adult silver fir trees that describe their individual reaction to extreme environmental stress, mainly SO_2 and drought, in an observational study.

The combination of both approaches aims at providing a framework that can contribute to the identification and analysis of candidate genes for stress response in silver fir and other non-model conifer species.

Outline of the thesis

Chapter 2 presents a novel terahertz time-domain spectrocopy (THz-TDS) setup that allows to continually measure the water status of multiple plants. This setup was used to monitor the desiccation process of silver fir seedlings during drought treatment, compared to the water status of well-watered individuals. Thus, needles from both groups could be harvested with near-identical water status within each group.

Chapter 3 is concerned with the possible influence of chlorophyll content on the terahertz signal during THz-TDS monitoring. Since decreasing water content can reportedly lead to decreasing chlorophyll content in leaves, different concentrations of extracted chlorophyll were measured using THz-TDS. Based on this simple experiment, chlorophyll concentration has no influence on the terahertz signal and can be disregarded as a potential source for error during THz-TDS monitoring.

Chapter 4 describes the differential transcriptome profiling of the harvested needles from Chapter 2. Two libraries were constructed by Massive Analysis of cDNA Ends (MACE) and subsequent analyses of differential expression between the two treatments. By comparing the expression levels between drought stressed and control group, significant up- or down-regulated transcripts could be identified and partially annotated. Thus, novel functional candidate genes for drought stress response in silver fir could be identified.

Chapter 5 introduces dendrophenotypes derived from wood cores of silver fir trees in the Bavarian Forest National Park. Silver fir shows a strong depression in tree-ring width in the period of 1974-1987, which is explained mostly by sulfur dioxide (SO_2) pollution in combination with drought stress. By associating variation in pre-selected polymorphic candidate genes (SNPs) with individual response parameters during the depression period (dendrophenotypes), genes that might potentially be of adaptational consequence, could be identified.

Chapter 6 is a critical assessment of minimal-optimal feature selection techniques based on random forest that are increasingly used in genetic association studies. Specifically in the context of biological interpretation, the potential pitfalls and caveats of such an approach are demonstrated. A proposed interaction analysis is shown as having the potential to create statistically significant, yet diametrically different results due to random sampling alone.

Chapter 7 recapitulates and synthesizes the results of the previous chapters and places the findings of this thesis in a broader context. The methods and results are critically evaluated and an outlook on necessary improvements and further studies is given.

CHAPTER 2

Monitoring Plant Drought Stress Response Using Terahertz Time-Domain Spectroscopy

Norman Born, David Behringer, Sascha Liepelt, Sarah Beyer, Michael Schwerdtfeger, Birgit Ziegenhagen, Martin Koch

Plant Physiology (2014) 164, pp. 1571-1577

Monitoring Plant Drought Stress Response Using Terahertz Time-Domain Spectroscopy[C][W]

Norman Born[1]*, David Behringer[1]*, Sascha Liepelt, Sarah Beyer, Michael Schwerdtfeger, Birgit Ziegenhagen, and Martin Koch

Faculty of Physics and Material Sciences Center, Philipps-University Marburg, D–35032 Marburg, Germany (N.B., M.S., M.K.); and Faculty of Biology, Conservation Biology, Philipps-University Marburg, D–35043 Marburg, Germany (D.B., S.L., S.B., B.Z.)

We present a novel measurement setup for monitoring changes in leaf water status using nondestructive terahertz time-domain spectroscopy (THz-TDS). Previous studies on a variety of plants showed the principal applicability of THz-TDS. In such setups, decreasing leaf water content directly correlates with increasing THz transmission. Our new system allows for continuous, nondestructive monitoring of the water status of multiple individual plants each at the same constant leaf position. It overcomes previous drawbacks, which were mainly due to the necessity of relocating the plants. Using needles of silver fir (*Abies alba*) seedlings as test subjects, we show that the transmission varies along the main axis of a single needle due to a variation in thickness. Therefore, the relocation of plants during the measuring period, which was necessary in the previous THz-TDS setups, should be avoided. Furthermore, we show a highly significant correlation between gravimetric water content and respective THz transmission. By monitoring the relative change in transmission, we were able to narrow down the permanent wilting point of the seedlings. Thus, we established groups of plants with well-defined levels of water stress that could not be detected visually. This opens up the possibility for a broad range of genetic and physiological experiments.

Climate change simulations predict an increase in the occurrence of drought events in the Mediterranean area and in central Europe due to smaller amounts of precipitation, especially during summer periods (IPCC, 2007). With the exception of the boreal zone, this leads to an increase in drought risks for every region on the European continent (Iglesias et al., 2007). Water availability is very important for a variety of plant species. Trees and crops play major roles regarding ecosystem stability and food supply. Forest trees are keystone elements in shaping long-term, regional ecosystem composition and stability and are, like most forest species, highly vulnerable to increases in drought severity (Breshears et al., 2005; Choat et al., 2012). Drought-induced forest die-offs thereby directly reduce ecosystem services such as carbon sequestration and timber supply (Allen et al., 2010). Further research is clearly necessary to elucidate the physiological traits and responses of plants regarding their water status.

European silver fir (*Abies alba*) is an important forest tree species of ecological and economic relevance. This study is embedded in the European project LinkTree, "linking genetic variability with ecological responses to environmental changes: forest trees as model systems." Our group is concerned with the identification of genes involved in the water stress response of silver fir. This species is of special interest because of its lower water-use efficiency compared with other conifer species (Guehl and Aussenac, 1987; Guehl et al., 1991).

For this purpose, monitoring plant water status without inducing other forms of stress is instrumental in order to apply well-defined levels of water stress. Obtaining information regarding the water status of a plant is highly problematic without using invasive and destructive methods that usually only allow a retrospective assessment. These include commonly established methods, such as the gravimetric water content and pressure chamber techniques, most notably Scholander's pressure bomb (Scholander et al., 1965).

Chlorophyll fluorescence, stomatal conductance, and visual assessment are examples of nondestructive and noninvasive measurement techniques. The former two only provide indirect information about the plant stress status and, therefore, the water content via photosynthetic activity (Lichtenthaler and Rinderle, 1988; Tardieu and Davies, 1993). The latter is difficult to standardize and highly dependent on the morphology of the studied plant species. Conifers especially are challenging subjects for visually assessing drought stress. Due to their needle morphology, it is nearly impossible to detect early signs of dehydration.

[1] These authors contributed equally to the article.

* Address correspondence to norman.born@physik.uni-marburg.de and david.behringer@biologie.uni-marburg.de.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Norman Born (norman.born@physik.uni-marburg.de).

D.B. and N.B. performed most of the experiments; N.B. provided technical assistance to D.B.; D.B. handled the plants, and S.L. supervised him; D.B., N.B., S.L., S.B., and M.S. designed experiments; D.B. and N.B. analyzed the data; and B.Z. and M.K. conceived of the project. D.B. and N.B. wrote the paper with input from B.Z., M.K., S.B., and S.L.

[C] Some figures in this article are displayed in color online but in black and white in the print edition.

[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.113.233601

Measurement techniques using electromagnetic radiation in the terahertz (THz) regime have shown promising results, due to the nondestructive nature and high sensitivity of THz waves to water. With THz waves, we refer to frequencies in the electromagnetic spectrum between 0.1 and 1 THz, corresponding to wavelengths between 3 and 0.3 mm, which are located between infrared light (thermal radiation) and microwave radiation (used in common wireless data communication systems). In the last decade, terahertz time-domain spectroscopy (THz-TDS) has proven to be a very strong and accurate tool for characterizing and imaging various materials (for review, see Jepsen et al., 2011). Crucial for our study is the remarkably high absorption coefficient of water in this part of the electromagnetic spectrum. Thus, it is a robust technique hardly affected by physiological concentrations of soluble substances. Using transmission geometry, the resulting absorption by plant tissues directly reflects the quantity of water molecules.

Furthermore, THz-TDS does not suffer from the disadvantages of other radiation-based techniques. These are mainly focused on the infrared or microwave spectrum but either lack the sensitivity for small changes in leaf water status or are affected by the plant's inorganic salt content, leading to significant disturbances (Ulaby and Jedlicka, 1984). Moreover, the applicability of emitting microwave radiation is limited to minimal wavelengths of approximately 2.5 mm. The Abbe diffraction limit, therefore, restricts the minimum diameter of a measurable object to approximately 1.25 mm. In order to measure small leaves, such as coniferous needles, electromagnetic radiation with shorter wavelengths is necessary.

Although presenting a useful alternative, THz-TDS was not feasible until recently, due to the difficulty of generating and detecting electromagnetic radiation with wavelengths in the THz spectrum. Despite its promising applicability in plant sciences, until now this relatively novel method relied exclusively on measurement setups that allowed only a single measurement per alternating plant (Hadjiloucas et al., 1999; Jördens et al., 2009; Breitenstein et al., 2012; Castro-Camus et al., 2013; Gente et al., 2013). For the purpose of continuously monitoring multiple plants, these setups are only of limited use, since the plants must be relocated for every measurement. This results in two problems: (1) an increase in possible disturbances (e.g. mechanical), influencing the plant's stress response, and (2) the necessity to precisely target the same measurement spot on every analyzed plant at every consecutive measurement. The latter is of crucial importance for the exact monitoring of any individual plant's water status because, as we will show in this study, the transmission varies substantially across the area of plant leaf tissue.

We present a novel measurement procedure that overcomes the drawbacks of previously proposed methods. Our approach enables us to precisely monitor changes in the water content of multiple plants simultaneously.

In the course of this study, three different experiments were performed. The profile measurement and the rehydration experiment were preliminary investigations to examine the influences of needle and tissue thickness and to define a nonlethal stress level. The main experiment established groups of plants with comparable levels of water stress.

RESULTS

The THz transmission is a measure for the proportion of radiation reaching the detector. Without any absorbing or reflecting materials, the transmission is defined as 100%. When monitoring plant leaves, the measured transmission is always a result of the volume of water at the measurement spot. A lower transmission, therefore, might be due to higher water content and/or leaf thickness. Accordingly, every individual needle produces an individual transmission baseline, which does not precisely translate into specific water content. Increase or decrease of transmission (ΔT), instead, is comparable and directly attributable to changes in the water content. The profile measurements of the needle showed a general decrease in transmission along its main axis (Fig. 1). Correspondingly, the highest transmission of 36% was measured at the tip and the lowest transmission of 22% was measured near the base of the needle. The transmission correlated with the difference in needle thickness of the tip and base, 170 and 250 μm, respectively. The repeated measurements of every position along the needle showed highly reproducible values, leading to very low SD values ranging at maximum up to 4% for five consecutive measurements.

At the time of harvesting, a statistically significant negative correlation between transmission and relative water content for all measured seedlings was evident ($r = -0.98$, $P < 0.001$; Fig. 2).

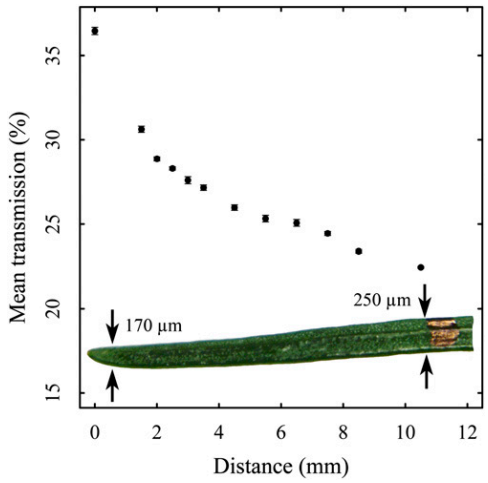

Figure 1. Transmission profile of a silver fir needle along the main axis, with the thickness of the needle at the outermost points of the profile. Each point represents the mean over five measurements. The SD is given for every point but sometimes is smaller than the pixel of the point (Supplemental Table S1). [See online article for color version of this figure.]

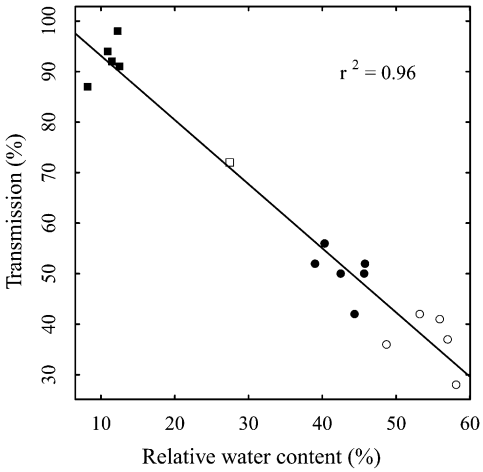

Figure 2. Correlation between THz transmission and gravimetric water content of the respective irrigated (white circles; I1–I5), stressed (black circles; S1–S6), relatively desiccated (white square; S7), and completely desiccated (black squares; D1–D5) seedlings, with the linear regression line and the corresponding coefficient of determination (r^2).

In the rehydration experiment, irrigation after drought treatment was conducted for seedlings R1 and R2 after reaching a ΔT of 58% and 43%, respectively (Fig. 3A). Correspondingly, R3 and R4 were irrigated after reaching a ΔT of 11% and 19%, respectively (Fig. 3B). Thus, the permanent wilting point (PWP) of the seedlings was narrowed down to a ΔT between 19% and 43% at 1.5 to 3 d after the onset of increased transmission. While R1 and R2 did not show any signs of recovery from drought and instead remained on high transmission levels, transmission in R3 and R4 dropped quickly after rehydration to levels close to the initial baselines.

The stressed group of seedlings in the main experiment showed an increase in transmission similar to each other, while the irrigated control seedlings remained at relatively constant levels (Fig. 4; for further information regarding the overall drought period, see Supplemental Table S2). Transmission at harvesting varied from 42% to 56% within the stressed group. ΔT, however, only showed a slight variation of 5% (Table I). Seedling S7 showed a relatively high ΔT of 24% and, accordingly, a very low gravimetric water content of 27.44%. Normalization of the data revealed that the respective pinhole was not properly adjusted, leading to a reflection of THz radiation. Hence, seedling S7 was too desiccated and, therefore, was removed from the stressed group as an artifact.

The transmission of all irrigated control seedlings exhibited a very low SD of 0.86%. The accuracy of the control measurements is illustrated by the shape of the probability-density function (Fig. 5). The differences in relative water content were highly significant between the stressed group and the irrigated control group (Student's t test, $P < 0.001$).

Room temperature fluctuated regularly during the whole measurement period around 20.3°C, with maximum values up to 21.4°C during the day and minimum values down to 19.6°C at night. Relative humidity showed strong fluctuations ranging from 44.8% to 16.8%, not following any recognizable pattern. Both temperature and relative humidity did not correlate with the fluctuations in transmission (Supplemental Figs. S1–S3). Illumination intensity during the day was 3,500 lux on average.

DISCUSSION

In this paper, we introduce a newly developed THz-TDS setup that was thoroughly validated. The main goal of our study was to provide a sensitive tool for monitoring the reaction of plants in an in vivo drought stress experiment by using ΔT. Before discussing the main results, we address the outcome of the preliminary validation steps. These were performed to demonstrate the accuracy of the experiment and to identify factors that influenced the measurements.

To better understand the results, we emphasize the fact that THz transmission is expected to be influenced not only by the water content but also by the measurement

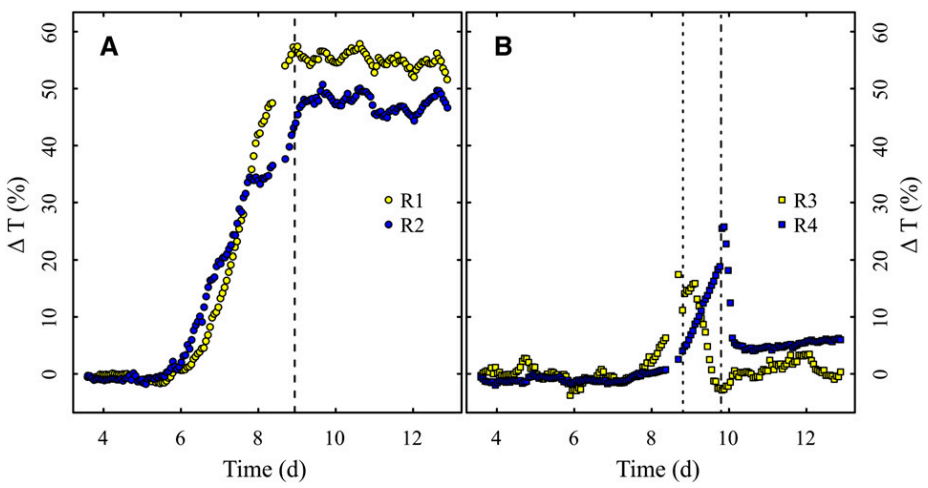

Figure 3. Monitoring ΔT over time to narrow down the PWP. A, Seedlings R1 and R2 irreversibly passed the PWP (the dashed line shows the beginning of regular irrigation). B, Seedlings R3 and R4 recovered from drought treatment after irrigation (the dotted and dotted-dashed lines show the time of needle extraction and the beginning of regular irrigation, respectively). [See online article for color version of this figure.]

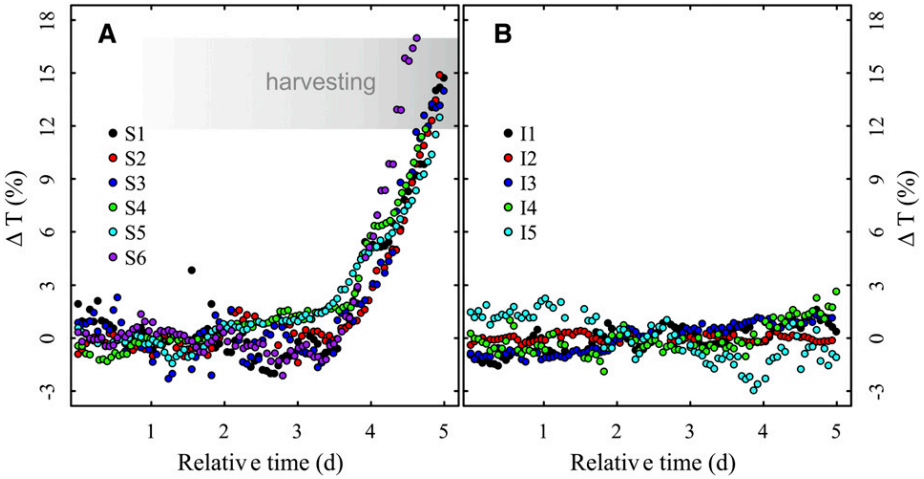

Figure 4. ΔT over a relative time for the six water-stressed seedlings S1 to S6 (A) and the five irrigated control seedlings I1 to I5 (B). The transmission curves of the stressed seedlings were aligned at their respective baselines, and the range of ΔT at harvesting is shown. While the stressed seedlings show a relatively uniform increase in transmission and therefore a decrease in water content, the irrigated seedlings remain at a relatively constant level of transmission and therefore show no signs of dehydration. [See online article for color version of this figure.]

spot (Mittleman et al., 1996). We demonstrated this by moving the measurement spot along the axis of a needle (Fig. 1). Thus, the variation of transmission along the axis can be attributed mainly to leaf thickness/composition. Using broad-leafed specimens, we observed similar variation when moving the measurement spot across the surface of a leaf (data not shown). Therefore, repeated measurements of plants at different spots introduce a variation in transmission even without a change in water content. We could demonstrate, however, that repeated measurements at the same spot are highly accurate and reproducible (see error bars in Fig. 1).

In the next step, we observed a high correlation between transmission and water content and, thus, a high sensitivity of THz radiation to the water content (Fig. 2). This correlation could be used for establishing a species- and setup-specific standard curve as a proxy for water content. The remaining variation, however, indicates that the use of absolute values is not straightforward for studying the response of plants in such an experiment, at least without knowledge of the exact thickness or composition.

This is exactly where our system was meant to provide a solution. As we discuss below, the continuous monitoring of the drought stress response clearly overcomes previous shortcomings. In advance of the main experiment, we performed a rehydration experiment, which already confirmed the advantages of our

Table I. *Fresh weight, dry weight, relative water content, and transmission at harvesting for each seedling with its respective plant and treatment group*

For each stressed seedling, the difference in transmission between the respective baseline and the point of harvesting (ΔT) is given. –, ΔT was not calculated for the irrigated and desiccated seedlings.

Plant	Treatment	Fresh Weight	Dry Weight	Relative Water Content	Transmission at Harvesting	ΔT
		mg			*%*	
S1	Stressed	117.0	67.3	42.48	50	15
S2	Stressed	102.7	61.3	40.31	56	15
S3	Stressed	129.0	70.1	45.66	50	14
S4	Stressed	171.8	104.8	39.00	52	12
S5	Stressed	140.7	76.3	45.77	52	13
S6	Stressed	126.0	70.1	44.37	42	17
S7	Stressed[a]	102.4	74.3	27.44	72	24
I1	Irrigated	261.6	122.4	53.21	42	–
I2	Irrigated	287.1	147.3	48.69	36	–
I3	Irrigated	243.3	107.3	55.90	41	–
I4	Irrigated	141.9	59.4	58.14	28	–
I5	Irrigated	142.0	61.1	56.97	37	–
D1	Desiccated	54.4	47.6	12.50	91	–
D2	Desiccated	81.8	72.4	11.49	92	–
D3	Desiccated	60.5	53.9	10.91	94	–
D4	Desiccated	66.0	60.6	8.18	87	–
D5	Desiccated	25.3	61.1	12.25	98	–

[a]Seedling S7 was considered desiccated.

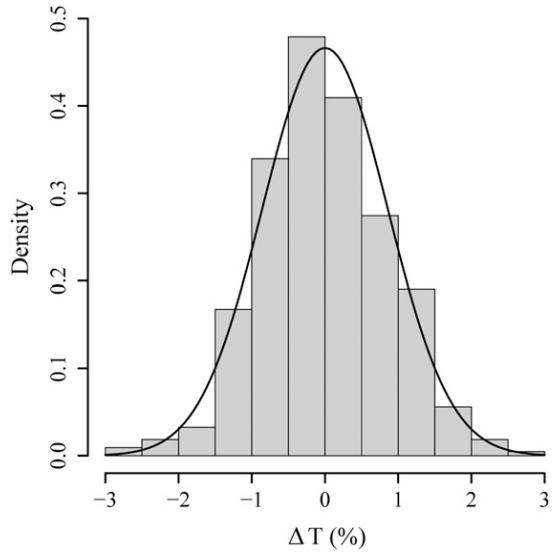

Figure 5. Histogram of ΔT for all irrigated control seedlings I1 to I5 with the corresponding probability-density function. The course of transmission for each seedling was set to a mean of 0% prior to the analysis.

setup but still served as a preliminary step for the main experiment. To be more specific, we were interested in determining the point at which drought stress affects the plants but is not yet lethal. Thus, a rough estimation of the PWP was sufficient in this case (Fig. 3). With the appropriate experimental design, our THz-TDS setup will allow for a much more precise determination of a species' or ecotype's specific PWP. Exact knowledge about the ΔT threshold, at which the PWP is reached, could provide valuable phenotypic information about a plant.

In addition, the rehydration experiment again provided evidence for variation introduced by moving the measurement spot. Prior to the irrigation, one needle was harvested from seedling R3 and one from R4, leading to an unavoidable shift in the position of the measured needle over the pinhole. Therefore, directly after remounting the needles, the transmission shifted for both samples, which corresponded precisely with the elevation of the baseline after recovery. From this evidence, we conclude that, in order to successfully monitor changes in plant water status, repeated measurements have to be performed at the exact same spot. Our novel measurement setup provides this basic ability.

In the main experiment, we observed that the stressed group of seedlings reacted in a similar way by losing water at similar rates after the onset of dehydration (Fig. 4A). This could be expected from seedlings originating from the same mother tree. Further experiments with seedlings from different genetic and geographic backgrounds are necessary to provide an assessment of the variation of the stress response within silver fir. The irrigated control group showed only slight variation (Figs. 4B and 5), demonstrating that our measurements were not significantly influenced by sources other than irrigation. This was confirmed by the recorded data for temperature and humidity, which showed no signs of correlation with the transmission of any individual seedling.

For our purpose of defining plants with nearly identical levels of water stress, the presented setup was fully sufficient. As a novelty, it enabled us to apply a level of stress that was traceable in the seedling's physiological response (i.e. water content) but not in its visual appearance (i.e. wilting).

Both the rehydration and the main experiment revealed that, instead of using absolute measures of transmission, continuously monitoring plants provides the key information, ΔT. Since this value is corrected for by the individual transmission prior to dehydration, it is possible to compare all plants in the experiment independently of tissue thickness and composition. This is applicable to both broad-leafed and coniferous plants.

Although our novel setup provides precise continuous measurements at the same spot, growing plant tissue might introduce additional variation. While this was not relevant in our specific case, this should be

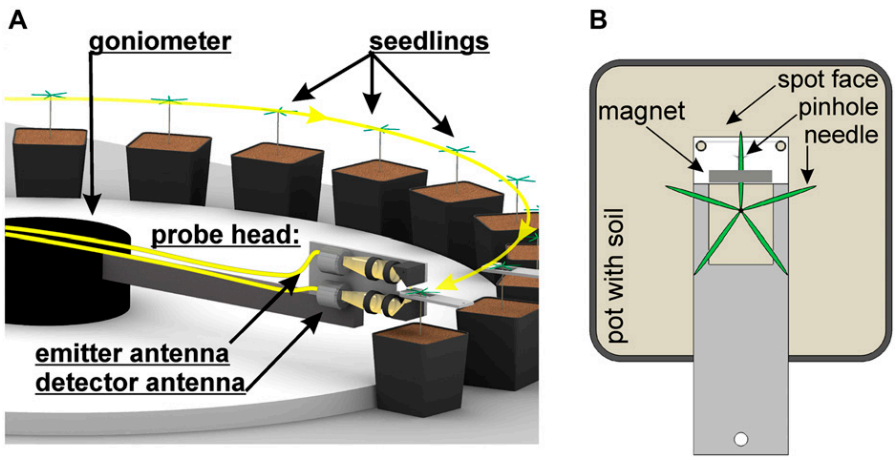

Figure 6. Measurement setup with the goniometer rotating the THz antennas in a precise angle along the positioned silver fir seedlings (A), which each have one needle clamped with a small magnet onto a holding device that positions the needle directly above a pinhole through which the THz radiation passes (B). [See online article for color version of this figure.]

considered when adapting the system to fast-growing plant species.

Previous studies using THz-TDS showed that the method works in principle and is applicable to different plant taxa. Yet, these setups are limited in experimental design (Jördens et al., 2009; Castro-Camus et al., 2013). Here, we present a setup suitable for a broad range of applications in ecophysiology and genetics. This is mainly due to three features: (1) high accuracy and reproducibility of the measurements, (2) multiple samples studied in parallel, with the ability for legitimate comparison between test subjects, and (3) the possibility to manipulate selected environmental conditions while reducing handling stress to a minimum.

PERSPECTIVE

In this study, we established an accurate tool for measuring effects directly related to water deficiency. This opens up the possibility for a broad range of experiments studying cause-effect relationships. By monitoring the reaction of an array of plants to defined levels of stress, groups exhibiting similar responses can be identified and selected for in-depth study of the underlying causes. Beyond the focus of our exemplary study, our ΔT approach allows the comparison of genotypes or accessions regarding their specific stress response. This response can be further characterized in terms of delay time until the onset of the stress reaction as well as the intensity of the response, which is defined by the variation of transmission over time ($\Delta T/\Delta t$).

MATERIALS AND METHODS

Plant Material

Silver fir (*Abies alba*) seeds were collected from the female cones of a single seed tree in a forest stand near Hagenbach, in the Black Forest region of southwestern Germany.

The seeds were cleaned, soaked in cold water for 24 h, and put into germination trays. For the stratification process, the germination trays were stored in a cooling chamber at 5°C for 6 weeks. Afterward, they were relocated to a thermal chamber that was adjusted to 28°C and moved to a greenhouse after 1 week. The germinated seedlings were individually planted into identical pots containing peat soil and kept in a greenhouse for 6 months. The positions of the single pots in the greenhouse were randomized to avoid the effects of a heterogenous environment and were irrigated three times per week. In preparation for the THz measurements, the seedlings were repotted into smaller clay pots containing slightly sandy topsoil 2 months prior to the experiments.

In the first step, we used one seedling to measure a needle profile to demonstrate the influence of varying the measurement spot. Subsequently, four seedlings (R1–R4) were chosen for a rehydration experiment. The aim was to determine the PWP. Another 12 seedlings were chosen for the main water stress experiment. The seedlings were assigned to two groups, one water-stressed group of seven seedlings, labeled S1 to S7, and one irrigated control group of five seedlings, labeled I1 to I5. Additionally, five completely desiccated seedlings (D1–D5) were chosen to be included in a correlation analysis.

THz-TDS Setup

The setup consisted of an erbium fiber laser (C-Fiber; Menlo Systems), which generated infrared light pulses at a wavelength of 1,550 nm with pulse lengths of 66 fs and a fiber length of 36 m. These laser pulses were split into two arms and guided through an optical fiber to two THz antennas (low-temperature molecular beam epitaxy-grown beryllium-doped indium-gallium-arsenide/indium-aluminium-arsenide multinanolayer; commercially available at Menlo Systems). These antennas were mounted on a probe head that was fixed on a movable goniometer arm (Fig. 6A; Supplemental Video S1). The upper antenna acted as an emitter and the lower one as a detector. The radiated THz beam was guided through the probe head via four high-density polyethylene lenses and two plane metallic mirrors, thus focusing the beam on a fir needle. Metallic holding devices were used to define a fixed measurement spot for every probed needle (Fig. 6B). These holding devices were designed to minimize any disturbances that could negatively affect plant growth. Hence, plant shading was reduced to a minimum. With a small and weak magnet, the probed needles were gently attached to a metallic spot face above a pinhole with 1.5 mm in diameter. The pinhole defined the focal spot of the THz radiation and was used to properly adjust the needle position in the optical path. In addition, the pinhole avoided spatial overexposure of the needle by blocking any radiation not guided through the pinhole, which would have led to a degradation of the measured data.

Up to 12 seedlings were measured automatically over several weeks, with each measuring cycle lasting 1.5 h. Furthermore, we applied a simulated day/night cycle with 10-h days and 14-h nights. Illumination was provided by a Philips bulb with an average of 3,500 lux at approximately 2 m distance from each holding device. During the measurement period, the facilities were monitored by an air conditioning system, which constantly adjusted the temperature within 21°C ± 0.5°C.

The data for a single measurement were acquired as a function of time by varying the spatial length of the optical paths through the detector antenna. Hence, it was possible to detect the time-resolved electrical field of a THz pulse. Fast Fourier transformation allowed calculating the frequency components comprising the THz pulse. By measuring a probe signal and comparing it with a reference signal, the frequency-resolved transmission was calculated in a frequency window ranging from 150 to 300 GHz, without any system-specific characteristics (for further information, see Koch et al., 1998).

Every third measurement was a reference M_R (i.e. one holding device without a needle). This was necessary to minimize systematic errors caused by changes in the room temperature or the humidity. This reference was used to adjust the transmission T_a of each measured needle M_N using the following equation:

$$T_a = \frac{M_R}{M_N}$$

Simultaneously, fluctuations in room temperature and humidity were recorded using a data logger (LOG 32; Dostmann Electronic), which was placed in a shady position on the measuring table adjacent to the holding devices. Illumination intensity during the daytime was measured at several spots on the height and position of the holding devices using a conventional lux meter (LM-1010; Elvos).

To exclude possible variations in transmission attributable to physical causes, two blank holding devices were placed among the others. By measuring those blank controls, variation in transmission of the probed needles could be separated between "real" biological changes in water status and "concealed" fluctuations caused by (long-term) systematic errors. Therefore, the transmission for each seedling was individually corrected by an adjustment function, based on the variation of the control measurements. To be more precise, we subtracted the transmission of the blank controls from the sample transmissions in order to exclude the concealed errors (e.g. proximity to the air conditioning system or the entrance). Finally, each curve was normalized to the determined maximum transmission of the respective holding device without the needle.

THz-TDS Measurements

Initially, one needle of a seedling was cut off and fixed to an adjustable device that allowed moving the needle in parallel above a pinhole. By measuring points along its axis, a transmission profile was obtained. Directly after the measurements, the thickness of the needle at the outermost points of the profile was measured using a digital micrometer screw (Mitutoyo).

To establish a nonlethal stress level by narrowing down the PWP, seedlings R1 and R2 were irrigated after approximately 3 d of transmission increase and seedlings R3 and R4 after 1 and 1.5 d, respectively. Directly prior to the irrigation, one needle was cut off from both R3 and R4 and stored in liquid nitrogen for future genetic analysis. Afterward, the measured needles had to be relocated above the respective pinhole.

In order to establish plants with comparable levels of water stress, seedlings I1 to I5 were irrigated every 2 d with 25 mL of tap water, while S1 to S7 were not irrigated at all. After an increase in transmission from the respective baseline

(ΔT) of approximately 14%, two needles of each seedling were cut off. Needles from the irrigated control group were cut off matching the harvesting time of the stressed seedlings. All harvested needles were stored in liquid nitrogen for future genetic analyses.

Gravimetric Water Content

To evaluate the accuracy of the THz measurements, each seedling was stored in a plastic bag directly after the measurement was finished, and the fresh weight (*FW*) was determined. After drying the seedlings at 110°C to the point of brittleness for 4 to 8 h, depending on the water content and dimensions of the individual seedling, the corresponding dry weights (*DW*) were determined, and the relative water content (*RWC*) was calculated for each seedling using the following equation:

$$RWC(\%) = \frac{FW(\mathrm{g}) - DW(\mathrm{g})}{FW(\mathrm{g})} \cdot 100$$

Statistical Analysis

To test for a significant correlation between the gravimetric water content and the respective transmission of all measured seedlings, the Pearson product-moment correlation coefficient was calculated. A significant difference in gravimetric water content between the water-stressed group (S1–S6) and the irrigated control group (I1–I5) was tested for with Student's two-sample *t* test after confirming the necessary assumptions.

For illustration purposes, the relative transmission over time of each stressed seedling was plotted in one graph. This was done by aligning the baselines of each seedling, although the individual courses had different baselines and were monitored at different times (Supplemental Figs. S1 and S2). The irrigated control seedlings were treated similarly, but in order to determine the variation of the measurement values around the mean, the respective values were set to the same mean transmission of 0%. Based on this new data set, the SD of all irrigated seedlings was calculated. All statistical analyses were carried out using version 2.13.1 of the statistical software R (R Development Core Team, 2011).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Monitoring of water-stressed seedlings with the corresponding data for temperature and humidity.

Supplemental Figure S2. Monitoring of irrigated seedlings with the corresponding data for temperature and humidity.

Supplemental Figure S3. Monitoring of the seedlings during the rehydration experiment with the corresponding data for temperature and humidity.

Supplemental Table S1. Data for the transmission profile measurements.

Supplemental Table S2. Times of final irrigation, harvesting, and total duration of the drought period for all water-stressed seedlings and the seedlings used in the rehydration experiment.

Supplemental Video S1. Video of THz measurement setup with probe head in motion and closeup of holding device with seedling.

ACKNOWLEDGMENTS

We thank Hans Lehmann (from the forestry district Oberharmersbach) for providing the seeds within the ERA-Net BiodivERsA project LinkTree, Christina Mengel (Conservation Biology group, Philipps-University Marburg) for helping with the laboratory work, Stefan Busch (Faculty of Physics and Material Sciences Center, Philipps-University Marburg) for assisting with problems regarding software, and Ben Liepelt (Port Royal Films, Großhansdorf, Germany) for editing the video. We also acknowledge the valuable comments of two anonymous reviewers.

Received December 5, 2013; accepted February 4, 2014; published February 5, 2014.

LITERATURE CITED

Allen CD, Macalady AK, Chenchouni H, Bachelet D, McDowell N, Vennetier M, Kitzberger T, Rigling A, Breshears DD, Hogg EH, et al (2010) A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. For Ecol Manage **259:** 660–684

Breitenstein B, Scheller M, Shakfa MK, Kinder T, Müller-Wirts T, Koch M, Selmar D (2012) Introducing terahertz technology into plant biology: a novel method to monitor changes in leaf water status. J Appl Bot Food Qual **84:** 158–161

Breshears DD, Cobb NS, Rich PM, Price KP, Allen CD, Balice RG, Romme WH, Kastens JH, Floyd ML, Belnap J, et al (2005) Regional vegetation die-off in response to global-change-type drought. Proc Natl Acad Sci USA **102:** 15144–15148

Castro-Camus E, Palomar M, Covarrubias AA (2013) Leaf water dynamics of Arabidopsis thaliana monitored in-vivo using terahertz time-domain spectroscopy. Scientific Reports **3:** 2910

Choat B, Jansen S, Brodribb TJ, Cochard H, Delzon S, Bhaskar R, Bucci SJ, Feild TS, Gleason SM, Hacke UG, et al (2012) Global convergence in the vulnerability of forests to drought. Nature **491:** 752–756

Gente R, Born N, Voß N, Sannemann W, Léon J, Koch M, Castro-Camus E (2013) Determination of leaf water content from terahertz time-domain spectroscopic data. J Infrared Milli Terahz Waves **34:** 316–323

Guehl JM, Aussenac G (1987) Photosynthesis decrease and stomatal control of gas exchange in *Abies alba* Mill. in response to vapor pressure difference. Plant Physiol **83:** 316–322

Guehl JM, Aussenac G, Bouachrine J, Zimmermann R, Pennes JM, Ferhi A, Grieu P (1991) Sensitivity of leaf gas exchange to atmospheric drought, soil drought, and water-use efficiency in some Mediterranean Abies species. Can J Res **21:** 1507–1515

Hadjiloucas S, Karatzas LS, Bowen JW (1999) Measurements of leaf water content using terahertz radiation. IEEE Trans Microw Theory Tech **47:** 142–149

Iglesias A, Avis K, Benzie M, Fisher P, Harley M, Hodgson N, Horrocks L, Moneo M, Webb J (2007) Adaptation to Climate Change in the Agricultural Sector. AEA Energy & Environment and Universidad de Politécnica de Madrid, Madrid

IPCC (2007) Climate Change 2007: The Physical Basis. Cambridge University Press, Cambridge, UK

Jepsen PU, Cooke DG, Koch M (2011) Terahertz spectroscopy and imaging: modern techniques and applications. Laser & Photonics Reviews **5:** 124–166

Jördens C, Scheller M, Breitenstein B, Selmar D, Koch M (2009) Evaluation of leaf water status by means of permittivity at terahertz frequencies. J Biol Phys **35:** 255–264

Koch M, Hunsche S, Schumacher P, Nuss MC, Feldmann J, Fromm J (1998) THz-imaging: a new method for density mapping of wood. Wood Sci Technol **32:** 421–427

Lichtenthaler HK, Rinderle U (1988) The role of chlorophyll fluorescence in the detection of stress conditions in plants. Crit Rev Anal Chem **19:** 29–85

Mittleman DM, Jacobsen RH, Nuss MC (1996) T-ray imaging. IEEE J Sel Top Quantum Electron **2:** 679–692

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna

Scholander PF, Bradstreet ED, Hemmingsen EA, Hammel HT (1965) Sap pressure in vascular plants: negative hydrostatic pressure can be measured in plants. Science **148:** 339–346

Tardieu F, Davies WJ (1993) Integration of hydraulic and chemical signalling in the control of stomatal conductance and water status of droughted plants. Plant Cell Environ **16:** 341–349

Ulaby FT, Jedlicka RP (1984) Microwave dielectric properties of plant materials. IEEE Trans Geosci Rem Sens **GE-22:** 406–415

CHAPTER 3

A note on chlorophyll content potentially influencing *in-vivo* THz-TDS measurements

David Behringer

Manuscript (2016)

A note on chlorophyll content potentially influencing *in-vivo* THz-TDS measurements

David Behringer

Faculty of Biology, Conservation Biology, Philipps-University Marburg, D–35043 Marburg, Germany

Abstract

Terahertz time-domain spectroscopy (THz-TDS) allows for the highly precise monitoring of the water status of a single spot on a plant leaf. The measurements are, however, dependent on the water content being the only variable influencing the terahertz signal over time. Any confounding factor would lead to the introduction of a measurement error. Since a decrease in water content is closely linked with a decrease in the chlorophyll content of a leaf, the question arises if the terahertz signal might be influenced by the chlorophyll concentration during drought stress monitoring. Extracting chlorophyll from silver fir needles and measuring different concentrations using THz-TDS, no confounding effect of chlorophyll on the terahertz signal could be observed.

Keywords: Terahertz, time domain spectroscopy, chlorophyll extraction, *Abies alba*

Introduction

Water availability is of crucial importance to trees but also very hard to assess. Especially conifers pose a challenge since their leafs frequently do not show any visible signs of dehydration. At the same time, in the context of global climate change, drought stress is increasingly threatening conifer species at their southern distribution range. As such, silver fir (*Abies alba* Mill.) already shows diebacks at its southern margin in Mediterrenean areas due to more severe periods of summer drought (Nourtier et al., 2012).

In order to assess the water status of conifer needles, Born et al. (2014) developed a novel measurement setup based on terahertz time-domain spectroscopy (THz-TDS). Using this device, the water content of one spot on a single needle of multiple silver fir seedlings could be monitored over time. Continuously measuring the same spot ensured that the leaf properties were constant over the entire measurement period, with the exception of the water content. Born et al. (2014) conceded, however, that excessive plant growth (relative to the monitoring period) could introduce measurement errors. Given that the nature of the experiment in Born et al. (2014) was to harvest needles from different seedlings with the same level of drought stress, and being aware of the documented decrease in chlorophyll content in conifer needles as a result of water shortage (Buxton et al., 1985; Wallin et al., 2002), one might argue that even without an error due to growth, the

measurements might still be influenced by a decreasing chlorophyll concentration. If the chlorophyll content has an influence on the amount of terahertz radiation passing through a needle, then continuous measurement results of the same spot would not only be a function of the water content at that spot but also of the current chlorophyll concentration.

This study, therefore, attempts to assess the influence of total chlorophyll content on the results of *in-vivo* THz-TDS measurements of silver fir needles.

Materials and Methods

Chlorophyll extraction

Two needles were harvested from an adult *Abies alba* tree in the Botanical Garden in Marburg. The needles were immediately placed in an Eppendorfer tube and cooled down in liquid nitrogen to prevent chlorophyll degradation. Upon arrival in the laboratory the weight of the individual needles was measured. All following steps were conducted on ice and in shaded areas to minimize photo-oxidation and largely follow the protocol described in Schopfer (1989). The needles were cut with a scissor and then ground down with a pestle in 0.5 ml extraction medium containing 80% Acetone, 19.5% distilled water and 0.5% concentrated NH_3 (25% by weight). The homogenate was then transferred to a test tube, covered in tin foil, and filled up to a total of 2 ml with extraction medium. After 30 minutes with intermittent shakes, the homogenate was placed in a centrifuge (Spectrafuge 24D, Labnet International, Inc., NJ, USA) for 10 min at 10,000 x g. 1750 µl of the clear supernatant was then placed in another tin covered test tube and a dilution series with three concentrations was prepared (Table 1).

Table 1. Mixture ratios for the dilution series of the chlorophyll extract of two silver fir needles.

	100% $c_{\text{Chl } a+b}$	50% $c_{\text{Chl } a+b}$	25% $c_{\text{Chl } a+b}$
Homogenate [µl]	1000	500	250
Extraction medium [µl]	-	500	750
Total volume [µl]	1000	1000	1000

The total chlorophyll content (chlorophyll a and b) of the different samples was measured with a spectrophotometer (Ultrospec 2000, Pharmacia Biotech, Uppsala, Sweden). Prior to each measurement, the machine was calibrated by measuring the absorption of the empty cuvette at 750 nm to control for turbidity. Pure extraction medium was used as a reference. Each sample was measured at 645, 652 and 663 nm, respectively. The total chlorophyll concentration in each solution was then calculated according to Bruuinsma (1963) using equations 1 and 2.

$$c_{\text{Chl } a+b}\ [\text{mg l}^{-1}] = 20.2\ E_{645} + \ 8.0\ E_{663} \tag{1}$$

$$c_{\text{Chl } a+b}\ [\text{mg l}^{-1}] = 27.8\ E_{652} \tag{2}$$

Based on the chlorophyll concentration in the liquid solution the concentration per gram was calculated using equation 3 and averaged over both values derived from equations 1 and 2.

$$c_{\text{Chl } a+b}\ [\text{mg g}^{-1}] = \frac{\text{Extraction medium [l] } c_{\text{Chl } a+b}\ [\text{mg l}^{-1}]}{\text{Sample weight [g]}} \tag{3}$$

THz-TDS measurements

All three samples, plus a reference sample consisting only of extraction medium, were covered in tin foil, put on ice and transferred to a terahertz time-domain spectrometer in an adjacent building. The spectrometer consisted of an airtight chamber that could be flooded with nitrogen gas to displace all the moisture. Samples could be placed in the center of the chamber, directly positioned between a terahertz emitter on the one side of the chamber and a detector on the other. A femtosecond laser in conjunction with a beam cutter was used to send pulses of light to both terahertz antennas to generate terahertz radiation passing through the sample. By manipulating the length of the optical path through the detector, the electrical field of a terahertz pulse could be resolved in time and then transformed using the Fast Fourier algorithm to obtain the signals in the frequency domain. The THz-TDS setup was identical to the one used and described in detail in Born et al. (2014), except for the optical path which was guided by mirrors and not by optical fibers.

For each sample, the actual measurement results were divided by the measurements of the respective empty cuvette to account for the potential absorbing and reflecting effects of the cuvette material. Further, all measurements for the three concentrations were divided by the results from the reference sample (0% chlorophyll) to calculate the part of the transmission (i.e. terahertz radiation reaching the detector) that was only attributable to chlorophyll content.

Results

Chlorophyll extraction

The fresh weight of the two needles was 0.0213 g and 0.0257 g respectively. The total sample weight was 0.047 g. The three samples (100%, 50% and 25% $c_{\text{Chl } a+b}$) showed different transmissions at the three measured wavelengths, roughly corresponding to the respective dilution (Table 2). Accordingly, the chlorophyll content correlated quite closely with the dilution factor (Table 3).

Table 2. Absorption of the three different chlorophyll concentrations at different wavelengths λ, measured with a spectrophotometer.

λ	100% $c_{\text{Chl } a+b}$	50% $c_{\text{Chl } a+b}$	25% $c_{\text{Chl } a+b}$
645	0.760	0.376	0.184
652	1.833	0.947	0.459
663	1.098	0.545	0.263

Table 3. Chlorophyll content of the three solutions based on the absorption at different wavelengths (645, 652 and 663 nm). TCC = $c_{\mathrm{Chl}\,a+b}$ [mg l^{-1}].

$c_{\mathrm{Chl}\,a+b}$	TCC ($\lambda_{645+663}$)	TCC ($\lambda_{645+663}$)	TCC (λ_{652})	TCC (λ_{652})	Average TCC
100%	30.02	1.28	30.52	1.30	1.29
50%	15.17	0.65	15.15	0.64	0.65
25%	7.39	0.31	7.31	0.31	0.31

THz-TDS measurements

All samples were measured over a frequency range of zero to ten terahertz (Fig. 1) with a usable section from slightly above zero to one terahertz (Fig. 2). Within this section, the samples showed, on average, slight differences in signal strength with the 25% and 50% solution having about the same transmission. The solution with the highest chlorophyll concentration (100%) on the other hand had a higher signal strength than the other solutions. The reference samples containing only extraction medium transmitted less terahertz radiation than the samples for each concentration (all lines in Fig. 2 are above one, i.e. the samples have a stronger signal and therefore more transmission than the respective references).

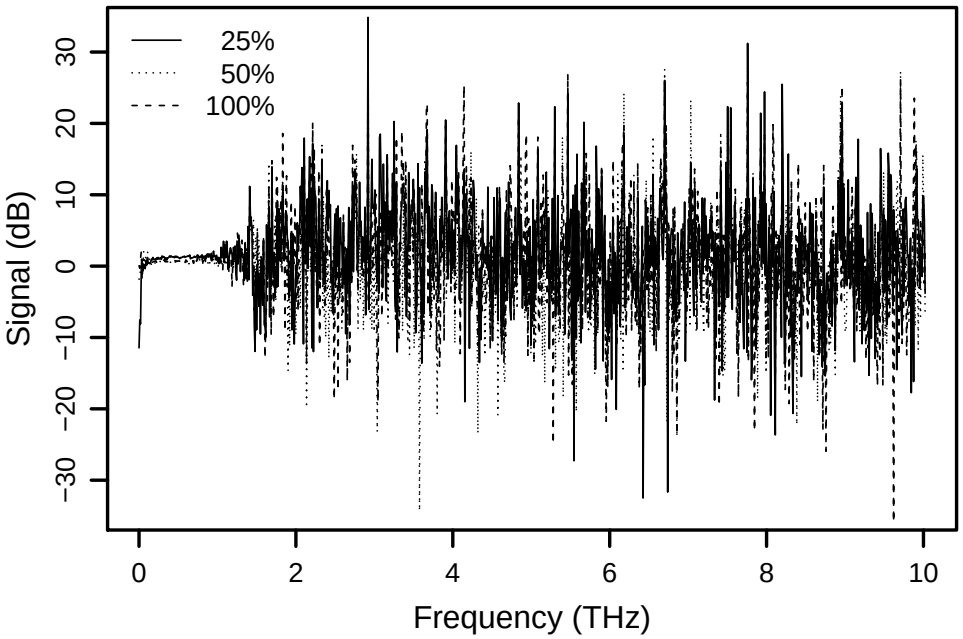

Figure 1. Signal strength over the entire measured frequency range for the three chlorophyll concentrations. Shown are the processed data that already account for possible effects of the cuvettes, as well as the extraction medium. The y-axis shows the signal strength in decibel (dB).

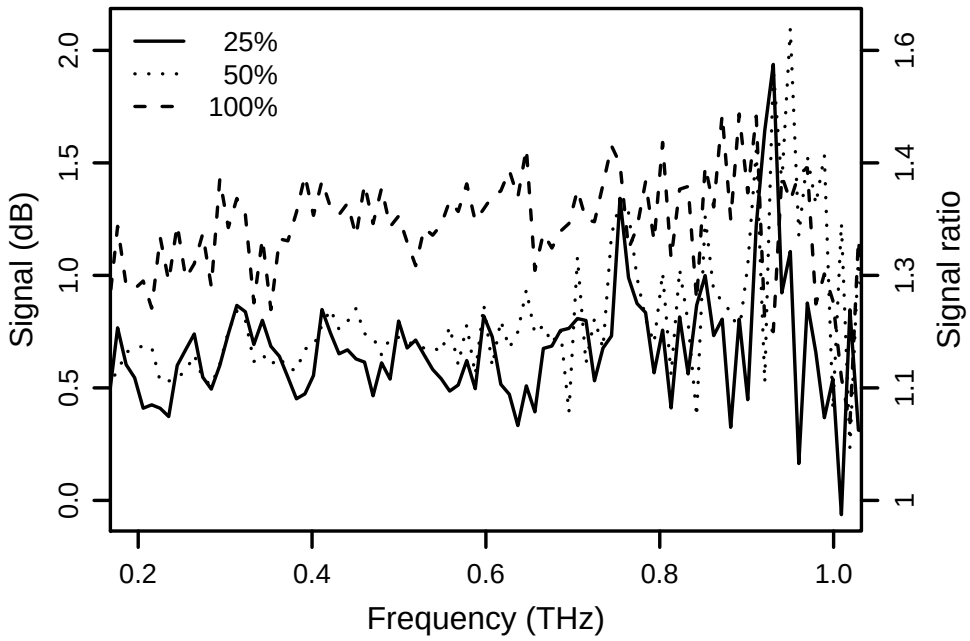

Figure 2. Section of the usable frequency range for all chlorophyll concentrations. Shown are the processed data that already account for possible effects of the cuvettes, as well as the extraction medium. The left y-axis shows the signal strength in decibel (dB). For the right y-axis the data was converted back to a linear representation with a value of one representing a 1:1 signal ratio between sample and reference (only extraction medium).[1]

Discussion

Regarding the potential influence of the chlorophyll content on THz-TDS measurements of silver fir needles, the results indicate no confounding effect.

Since more terahertz radiation passes through the samples containing chlorophyll than the respective reference samples and the solution with proportionally the least extraction medium (100% chlorophyll content) has the highest transmission, the extraction medium presumably has a bigger effect on the terahertz signal than the chlorophyll content. The effect of the latter is, at least based on the results of this study, negligible. It should further be mentioned that the differences in concentration of the measured chlorophyll samples are rather excessive. One would not expect a 50% drop in chlorophyll in a living plant as a consequence of drought stress. For example, after 20 days of drought stress (60% field capacity), five different varieties of *Helianthus annuus* plants showed an average decrease in total chlorophyll content of 0.13 mg g^{-1} fresh weight, compared to plants with 100% field capacity (Manivannan et al., 2007). Accordingly, after 40 days the decrease was, on average, 0.2 mg g^{-1} fresh weight. This corresponds to an average decrease in total chlorophyll content of 8.65% and 10.48% after 20 and 40 days of drought stress, respectively. In comparison, the drought stress period in Born et al. (2014) did not exceed five days.

Even if in the context of this study a confounding effect of chlorophyll content could not be shown, further research is clearly necessary. On one hand, the methods employed could surely be improved, both regarding the chlorophyll extraction as well as the THz-TDS measurements. Multiple extraction methods and solutions should be compared and the individual terahertz transmission

[1]The signal ratio (SR) in dB (SRdB) is $10 \log_{10}(\mathrm{SR})$. So to convert back from SRdB to SR is $10^{\mathrm{SRdB}\ 10^{-1}}$.

profile of every used substance should be accounted for. On the other hand, the measurements lack proper replication, both in sample size and, to widen the scope beyond silver fir, across different taxa. Further, chlorophyll might show different transmission profiles *in-vivo* than in an extracted solution based on e.g. electrical or magnetic properties. Next steps should also include juvenile leaf material and other plant substances, such as secondary metabolites, that could potentially influence THz-TDS measurements.

Acknowledgments

Special thanks go to Heike Zimmermann for help in sifting through, at times astonishingly erroneous, old literature, Michael Schwerdtfeger for help in conducting the THz-TDS measurements and Ralf Gente for fruitful discussions.

References

Born, N., Behringer, D., Liepelt, S., Beyer, S., Schwerdtfeger, M., Ziegenhagen, B. and Koch, M. (2014), 'Monitoring Plant Drought Stress Response Using Terahertz Time-Domain Spectroscopy', *Plant Physiol.* **164**(4), 1571–1577.

Bruuinsma, J. (1963), 'THE QUANTITATIVE ANALYSIS OF CHLOROPHYLLS a AND b IN PLANT EXTRACTS', *Photochemistry and Photobiology* **2**(2), 241–249.

Buxton, G. F., Cyr, D. R., Dumbroff, E. B. and Webb, D. P. (1985), 'Physiological responses of three northern conifers to rapid and slow induction of moisture stress', *Can. J. Bot.* **63**(7), 1171–1176.

Manivannan, P., Jaleel, C. A., Sankar, B., Kishorekumar, A., Somasundaram, R., Lakshmanan, G. and Panneerselvam, R. (2007), 'Growth, biochemical modifications and proline metabolism in Helianthus annuus L. as induced by drought stress', *Colloids and Surfaces B: Biointerfaces* **59**(2), 141–149.

Nourtier, M., Chanzy, A., Cailleret, M., Yingge, X., Huc, R. and Davi, H. (2012), 'Transpiration of silver Fir (Abies alba mill.) during and after drought in relation to soil properties in a Mediterranean mountain area', *Annals of Forest Science* pp. 1–13.

Schopfer, P. (1989), Qualitative und quantitative Analyse von Pflanzenmaterial, *in* 'Experimentelle Pflanzenphysiologie', Springer Berlin Heidelberg, pp. 1–51. DOI: 10.1007/978-3-642-61336-4_1.

Wallin, G., Karlsson, P. E., Selldén, G., Ottosson, S., Medin, E.-L., Pleijel, H. and Skärby, L. (2002), 'Impact of four years exposure to different levels of ozone, phosphorus and drought on chlorophyll, mineral nutrients, and stem volume of Norway spruce, Picea abies', *Physiologia Plantarum* **114**(2), 192–206.

CHAPTER 4

Differential Gene Expression Reveals Candidate Genes for Drought Stress Response in *Abies alba* (Pinaceae)

David Behringer, Heike Zimmermann, Birgit Ziegenhagen, Sascha Liepelt

PLoS ONE (2015), 10, e0124564

PLOS | ONE

RESEARCH ARTICLE

Differential Gene Expression Reveals Candidate Genes for Drought Stress Response in *Abies alba* (Pinaceae)

David Behringer*☯, Heike Zimmermann¤☯, Birgit Ziegenhagen, Sascha Liepelt

Conservation Biology Group, Philipps-University of Marburg, Germany

☯ These authors contributed equally to this work.
¤ Current address: Alfred-Wegener-Institute for Polar and Marine Research, Research Unit Potsdam, Department of Periglacial Research, Potsdam, Germany
* david.behringer@biologie.uni-marburg.de

Citation: Behringer D, Zimmermann H, Ziegenhagen B, Liepelt S (2015) Differential Gene Expression Reveals Candidate Genes for Drought Stress Response in *Abies alba* (Pinaceae). PLoS ONE 10 (4): e0124564. doi:10.1371/journal.pone.0124564

Academic Editor: Manoj Prasad, National Institute of Plant Genome Research, INDIA

Received: November 13, 2014

Accepted: March 5, 2015

Published: April 29, 2015

Data Availability Statement: FASTQ files from the two Illumina-based MACE libraries were deposited in the SRA (Short Read Archive, NCBI) with the following accession: PRJNA266095. FASTA files of all candidate genes selected in this paper have been uploaded as Supporting Information files.

Funding: Support was provided by LinkTree: Linking genetic variability with ecological responses to environmental changes: forest trees as model systems (Grant number: 01LC0822A). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Increasing drought periods as a result of global climate change pose a threat to many tree species by possibly outpacing their adaptive capabilities. Revealing the genetic basis of drought stress response is therefore implemental for future conservation strategies and risk assessment. Access to informative genomic regions is however challenging, especially for conifers, partially due to their large genomes, which puts constraints on the feasibility of whole genome scans. Candidate genes offer a valuable tool to reduce the complexity of the analysis and the amount of sequencing work and costs. For this study we combined an improved drought stress phenotyping of needles via a novel terahertz water monitoring technique with Massive Analysis of cDNA Ends to identify candidate genes for drought stress response in European silver fir (*Abies alba* Mill.). A pooled cDNA library was constructed from the cotyledons of six drought stressed and six well-watered silver fir seedlings, respectively. Differential expression analyses of these libraries revealed 296 candidate genes for drought stress response in silver fir (247 up- and 49 down-regulated) of which a subset was validated by RT-qPCR of the twelve individual cotyledons. A majority of these genes code for currently uncharacterized proteins and hint on new genomic resources to be explored in conifers. Furthermore, we could show that some traditional reference genes from model plant species (*GAPDH* and *eIF4A2*) are not suitable for differential analysis and we propose a new reference gene, *TPC1*, for drought stress expression profiling in needles of conifer seedlings.

Introduction

Smaller amounts of precipitation and an increase in the occurrence of drought events are being predicted for the Mediterranean region and parts of Central Europe, especially during summer periods [1]. Drought stress poses a major threat to trees by possibly causing hydraulic failure. Facing low water availability, trees react with stomatal closure and reduced photosynthesis

Competing Interests: The authors have declared that no competing interests exist.

while maintaining metabolism. Thus, severe drought periods can ultimately cause forest diebacks by xylem cavitation, carbon starvation and damage by pathogens and insects [2]. While single trees might reveal sufficient plasticity, tree populations can cope with changing climatic conditions either by migration and/or adaptation [3]. Migration via seed dispersal might be increasingly constrained by fragmented landscapes and is generally limited by natural barriers, especially for mountainous species [4]. Adaptation only works via selection on standing genetic variation or newly arisen mutations [5]. Standing variation offers the best chance for rapid adaptation since potentially beneficial alleles might already be numerously present in the population [5].

To assess the adaptive potential of tree populations it is therefore necessary to identify genes that are, among other stress-related traits, involved in the drought stress response and to study their variation in natural populations. Many studies on tree populations today use a candidate gene approach for the following reasons: While genomic resources are readily available for model species such as *Arabidopsis*, rice or maize, for most tree species, and especially conifers, such resources are scarce [6]. Furthermore, the identification of drought stress related genes is a big challenge when working with conifers, due to their large genome sizes [7] which make genome-wide association studies very resource-intensive [8]. Candidate genes are thus used as an alternative for detecting selective signals [9,10]. This approach traditionally involves F_{st} outlier analysis of single-nucleotide polymorphisms (SNPs) within those candidate genes ("bottom-up" approach) [11]. Alternatively, candidate genes allow for a reasonable selection of target genes for association studies ("top-down" approach), especially when handling large genomes. Moreover, drought stress is hard to assess in conifers and the response is a highly quantitative trait.

However, using a novel terahertz spectroscopy setup, it is now possible to continuously measure the water content of multiple plants and thereby precisely monitor the drought stress response [12]. This allows sampling leaves from multiple plants with identical drought stress response and analyzing them with next generation sequencing methods. Thus it is possible to identify those specific genes that are underlying an accurately assessed drought stress phenotype. This approach provides a unique opportunity for detecting and exploring novel candidate genes in non-model species which may not be found annotated from traditional model species. We chose European silver fir (*Abies alba* Mill.) as the model species for our study. *A. alba* is an ecologically and economically valuable coniferous tree, which has its main area of distribution in mountainous regions of Central and Southern Europe [13]. Effects of drought were shown to manifest in a reduced growth rate [14–16], reduced photosynthetic activity and stomatal conductance [17,18], crown-damage [19,20] and an increasing susceptibility to damage caused by pathogens or insects [21,22]. A dieback as a response to frequent and severe water shortages can already be observed, e.g. at Mont Ventoux in Southern France [23].

The major goals of our study were (I) the identification of candidate genes for drought stress response in *A. alba*, (II) the comparison of drought stress related genes between *A. alba* and model organisms to identify conifer-specific genes, (III) the validation of the expression profiles by reverse-transcription quantitative real-time PCR (RT-qPCR) and (IV) the identification of reference genes for RT-qPCR data normalization.

Materials and Methods

Plant material and drought stress monitoring

Silver fir seedlings were propagated from seeds of female cones of a single tree in a forest stand near Hagenbach, a Black Forest region of South-Western Germany (the seeds were provided with permission by Hans Lehman from the forestry office Oberharmersbach). Thus, all

seedlings used in the experiment were either half-siblings or full-siblings. To establish groups of plants with highly controlled levels of drought stress, a novel terahertz time-domain spectroscopy setup was used in a preliminary study conducted by Born et al. [12]. This allowed the manipulation and monitoring of the individual water status of multiple seedlings by continuously measuring the cotyledons, without inducing other forms of stress. Twelve seedlings were measured this way. While six of them were well-watered, the other six seedlings were not watered until they reached comparable levels of considerable drought stress (for a more detailed account of the drought stress monitoring and its results see Born et al. [12]). At this point, two cotyledons were cut off from each seedling for RNA extraction and immediately stored in liquid nitrogen. Cotyledons were also harvested from the control group of well-watered seedlings at corresponding times of the day.

RNA extraction

For sequencing, total RNA from every individual needle was extracted using the InviTrap Spin Plant RNA Mini Kit (STRATEC Molecular GmbH, Berlin, Germany). The cotyledons were ground in liquid nitrogen with mortar and pestle in lysis buffer RP and β-Mercaptoethanol. Half of each lysate was used for RNA extraction by GenXPro GmbH (Frankfurt am Main, Germany) while the rest was stored at -80°C for RT-qPCR validation. To remove genomic DNA contaminants the samples were treated "off-column" with Baseline-ZeroTM DNase (Epicentre/Biozym, Hessisch Oldendorf Germany) and subsequently purified using RNA Clean & ConcentratorTM-5 Kit (Zymo Research Europe, Freiburg Germany). RNA samples for RT-qPCR validation were immediately stored in a deep freezer at -80°C. RNA concentration and purity were measured via ratios of optical density ($OD_{260/280}$, $OD_{260/230}$) using NanoDrop 1000 spectrophotometer (PEQLAB Biotechnologie GmbH, Erlangen Germany). The absence of DNA contamination was confirmed after performing a PCR using a primer pair which targets the nuclear microsatellite marker NFH15 (GenBank Accession Number: AY966492, [24]) at an annealing temperature of 57°C. Integrity was assessed using gel-electrophoresis. Complementary DNA (cDNA) was synthesized using the Maxima First Strand cDNA Synthesis Kit for RT-qPCR (ThermoScientific, Schwerte Germany). The cDNA samples were immediately stored in aliquots at -80°C. All kits were applied according to the manufacturer's protocol. Any modifications are explicitly described.

Transcriptome sequencing

Prior to synthesizing cDNA, the extracted mRNA from the drought stressed and the well-watered seedlings was pooled, respectively. From each pool a cDNA library was constructed targeting sequences near the cDNA 3'-ends. This Massive Analysis of cDNA Ends (MACE) was conducted by GenXPro GmbH as described in Kahl *et al.* [25]. The 5'-ends of 50–500 bp long fragments were sequenced (single-read) using the Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA, USA), generating 100 bp long tags. Illumina's HiSeq Control Software v. 2.0.5 was used for sequencing, RTA v. 1.17.20.0 for real time analysis and CASAVA v. 1.8.2 (Consensus Assessment of Sequence and Variation) for base calling and demultiplexing. To prevent PCR-biased quantification, GenXPro's "TrueQuant" method was applied, thereby eliminating PCR-based copies from the dataset. For this purpose, unique oligonucleotides were ligated to each tag prior to PCR, making it possible to identify and eliminate PCR copies with identical barcode-tag-combinations [26,27].

Assembly, annotation and gene expression profiling

After sequencing, the tags were assembled and annotated using the TIGR Plant Transcript Assemblies database (http://plantta.jcvi.org/). Not annotated tags were assembled and subsequently blasted (BLASTx) against the Swiss-Prot and TrEMBLE databases (http://www.uniprot.org/). A differential expression analysis was conducted using the MA-plot based method with random sampling model (MARS) of the DEGseq R package [28]. Prior to this analysis the libraries were normalized according to their respective size by dividing each tag frequency through the sum of the total tags and multiplied by 10^6 (tags per million). For multiple testing corrections a *p*-value-threshold of 1e-10 for significantly differentially expressed (DE) transcripts was set. Following, the enrichment of each gene ontology (GO) term was tested using Fisher's exact test (two-tailed) [29]. Additionally, we analyzed the MACE results using the R packages DESeq [30] with a single estimated dispersion condition, a size factor normalization and an FDR (false discovery rate) threshold of $q < 0.1$ as well as NOISeq [31] with simulated technical replicates (NOISeq-sim), a trimmed mean of M-values normalization and a threshold of $q = 0.9$.

RT-qPCR validation

The gene expression of a small number of genes was assessed in each individual seedling by RT-qPCR using relative quantification according to the MIQE criteria (Minimum Information for the Publication of Quantitative Real-Time PCR Experiments) [32].

The MACE dataset was first searched for DE transcripts with $\log_2$ fold changes higher than 3 (for up-regulated transcripts) or lower than -3 (for down-regulated transcripts) since they were most likely responsive to dehydration. To minimize the rate of false positives introduced by rare transcripts, a threshold of at least 50 different tags with match in sense orientation (5'-3') to a database entry was set. Genes for validation were selected from this filtered subset of DE transcripts that were significantly assigned (enrichment-*p*-value < 1e-10) to the GO terms response to water stimulus (GO:0009415), response to water deprivation (GO:0009414) and response to osmotic stress (GO:0006970). Furthermore, genes were selected from the subset of filtered DE transcripts with the ten highest and ten lowest fold changes that were significantly assigned to the GO domain biological process (GO:0008150). Primer pairs (Metabion, Martinsried, Germany) were designed based on the assembled MACE tag sequences for each selected gene using Primer3 v. 4.0.0 (http://bioinfo.ut.ee/primer3/) with default parameters for a product size of 60 bp to 150 bp and an optimum annealing temperature of 60°C. Primer pairs were considered specific when (1) there was no amplicon present in genomic DNA samples, (2) the first derivative of the corresponding melting curves resulted in a single peak, (3) gel-electrophoresis showed one product with the expected size and (4) the amplicon-sequence was identical with the target sequence which was verified by re-sequencing of the PCR-products using the Macrogen Europe Laboratory sequencing service (Amsterdam, The Netherlands).

To select adequate reference genes for the normalization of the RT-qPCR data two different approaches were used. First, by searching the literature for conifer gene expression studies, traditionally used reference genes were identified. Second, the MACE-dataset was searched for sequence tags which were neither up- nor down-regulated (*p*-value > 0.99, $\log_2$ fold change: -0.005 to 0.005), and had a minimum amount of ten sequence tags. The potential reference genes were tested for their expression stability among the drought stressed and well-watered seedlings using geNorm [33] and Normfinder [34]. Both algorithms were implemented in GenEx v. 5.4.4.119 (MultiD Analyses AB, Göteborg Sweden) which also provided the expected accumulated standard deviation to assess the optimal number of reference genes to be included for the most precise data normalization [35]. Real-time PCR was performed on the Roche LightCycler 480 II System (Roche Diagnostics, Mannheim, Germany) using the sample

maximization method with samples in triplicates at an optimized and standardized temperature and cycle program (Table S1 in S1 File) in which only the annealing temperatures were varied according to the optimum of the primer pairs (Table S2 in S1 File). Each PCR reaction was performed with KAPA SYBR Fast Universal Master Mix (Peqlab, Erlangen, Germany).

After the RT-qPCR the quantification cycles (Cq) were determined using the second derivative maximum method implemented in the Roche LightCycler 480 Instrument Software v. 1.5.0. The gene expression ratio was calculated using the Pair Wise Fixed Reallocation Randomization Test implemented in the Relative Expression Software Tool-384 v. 1 (REST) using 5000 iterations [36]. The ratio was corrected for the amplification efficiencies which were calculated according to Liu & Saint [37]. The intra- and inter-assay variations were assessed by calculating the coefficient of variance as the standard deviation relative to the mean of the Cq-values. Therefore, thirty replicates of the same cDNA sample were used to amplify *GAPDH* in three separate qPCR runs (each with ten of the replicates) on three different days. The coefficient of variance was not supposed to exceed four percent on the Cq basis [38].

Results

MACE libraries

After sequencing, the MACE method yielded two libraries containing, in total, 15.4 million tags with 6.2 million tags for the drought stressed pool and 9.2 million tags for the well-watered pool (Table 1).

Annotation of the tags resulted in a total of 65,535 transcripts, which were assigned to the three main gene ontology (GO) domains: molecular function (GO:0003674) contained 38,745 transcripts, cellular component (GO:0005575) 39,776 and biological process (GO:0008150) 37,140. Since this analysis aimed to find candidate genes associated with drought stress response, transcripts assigned to the GO domain biological process were most interesting. Within this domain, GO terms associated with metabolic processes were most enriched. In response to drought stress these GO terms were mostly down-regulated, as was methylation (GO:0032259) and photosynthesis (GO:0015979) (Fig 1 and Table S3 & S4 in S1 File). In contrast, GO terms associated with stimuli and stress were generally up-regulated, especially terms most obviously linked to drought stress, namely response to water stimulus, response to water deprivation and response to osmotic stress.

Differential gene expression analyses

The DEGseq analysis resulted in a total of 3,407 significantly DE transcripts ($p < 1e\text{-}10$) between the drought stressed and the well-watered pool (Fig 2A). The NOISeq ($q = 0.9$, Fig 2B) and DESeq ($q < 0.1$, Fig 2C) analyses yielded 2,694 and 342 DE transcripts, respectively. DEGseq uniquely identified 1,726 transcript and NOISeq 1009 transcripts, while DESeq shared all identified transcripts with either NOISeq or both DEGseq and NOISeq (Fig 3).

Table 1. Characteristics of the MACE libraries constructed from the drought stressed and the well-watered seedlings.

	Tags (total)	Tags (unique)	Drought stressed	Well-watered
Hits (S+AS)	14,162,592	6,435,157	5,664,178	8,498,414
No hit	1,275,004	1,029,367	542,833	732,171
Total	15,437,596	7,464,524	6,207,011	9,230,585

Shown is the amount of total tags analyzed for the library, the amount of unique tags and the amount of tags for each treatment pool (drought stressed and well-watered).

S: sense direction; AS: antisense direction.

doi:10.1371/journal.pone.0124564.t001

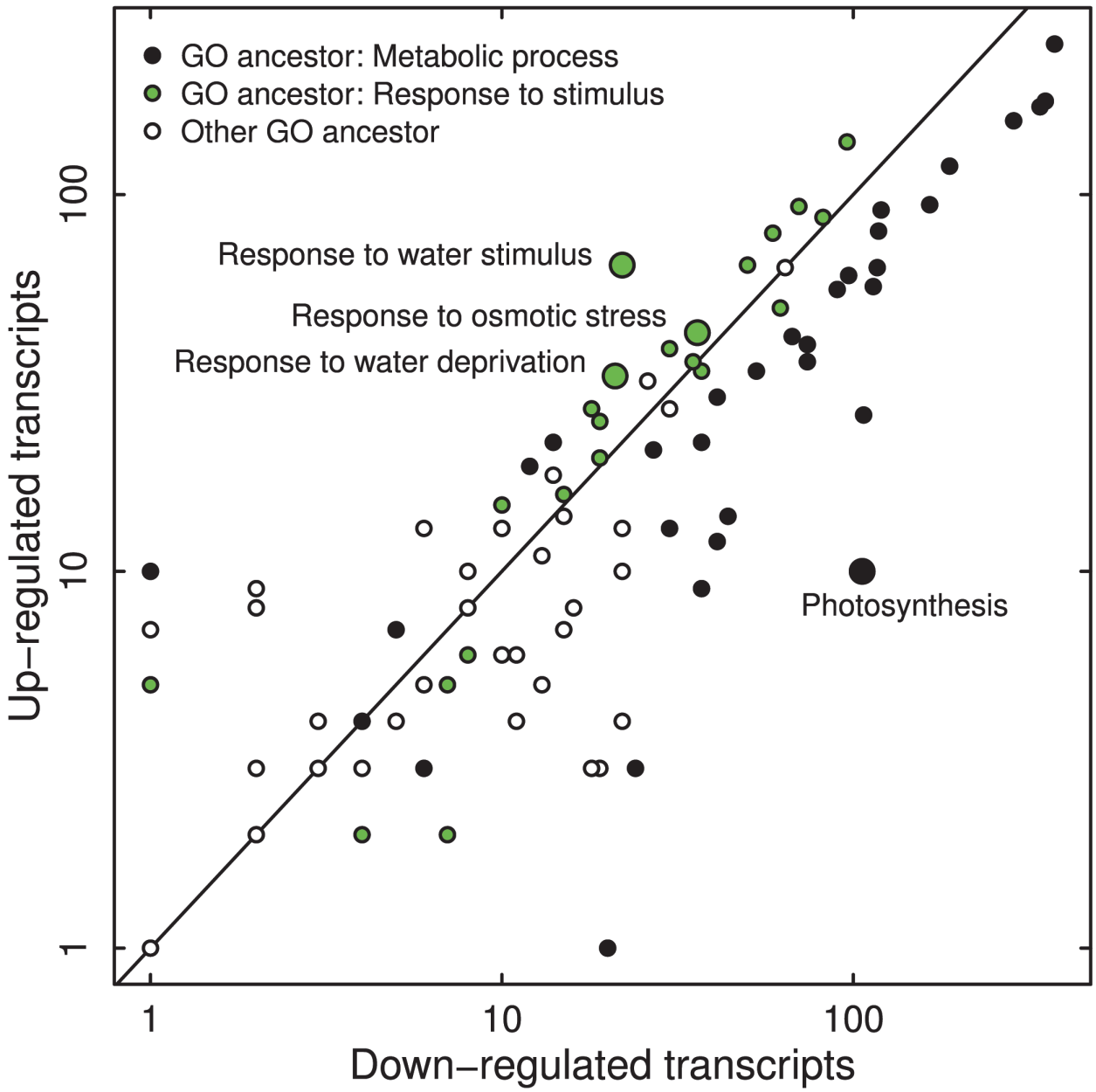

Fig 1. Log-log plot of up- and down-regulated transcripts in response to drought stress on GO-level 4 in silver fir seedlings. Transcripts are differentiated by their GO ancestor: metabolic process (GO:0008152), response to stimulus (GO:0050896) or other ancestor. Most obvious GO terms associated with drought stress, as well as photosynthesis, are highlighted and labeled specifically.

doi:10.1371/journal.pone.0124564.g001

RT-qPCR validation and candidate gene selection

After filtering by fold change and the minimum number of different sense tags ($\geq$ 50), out of the 3,407 DE transcripts identified by DEGseq, 832 transcripts could be listed (Table 2, FASTA files of all 832 transcripts available in S2 File). From the subset of filtered DE transcripts with a significant assignment to the GO terms response to water stimulus, response to water

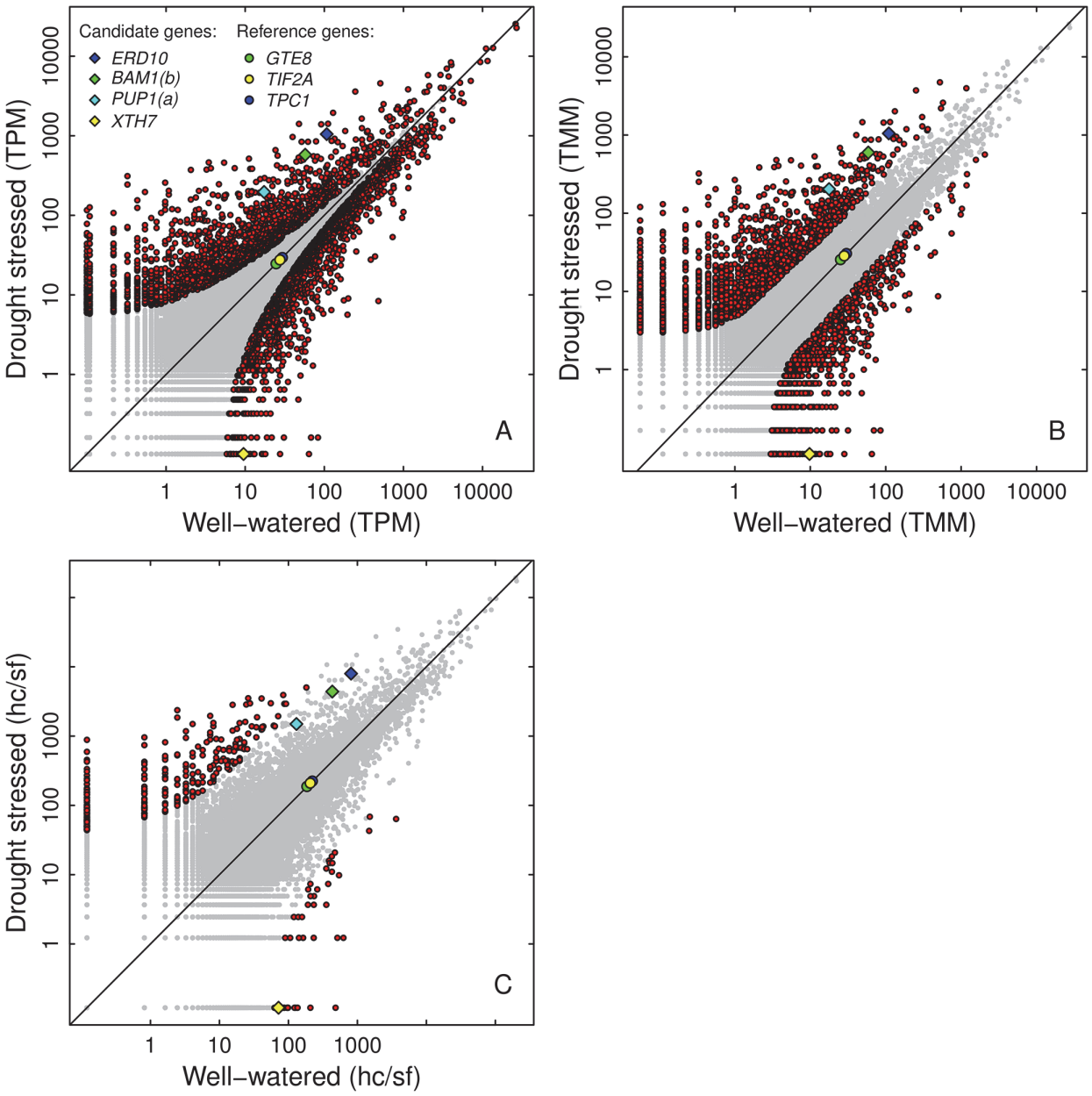

Fig 2. Scatter plots of the MACE results and subsequent analyses of differential expression for DEGseq (A), NOISeq (B) and DESeq (C). Each plot contains all identified transcripts (grey dots), as well as the analysis-specific DE transcripts (red dots). Further, the candidate genes validated via RT-qPCR are shown, as well as the corresponding stably expressed reference genes. The x- and y-axis give the transcript count in the well-watered and the drought stressed pool, respectively. Counts are normalized differently for the three analyses: tags per million (TPM) for DEGseq, trimmed mean of M-values (TMM) for NOISeq and hitcount/size factor (hc/fc) for DESeq. Transcripts falling on the straight line (90° bisecting line) are equally expressed in both pools. Transcripts above the line are up-regulated in response to drought stress, while those below the line are down-regulated.

doi:10.1371/journal.pone.0124564.g002

deprivation and response to osmotic stress and the top ten up- and down-regulated transcripts significantly assigned to biological process, 29 different up-regulated (Table S5, S6 & S8 in S1 File) and 14 different down-regulated (Table S7 & S8 in S1 File) genes for validation could be

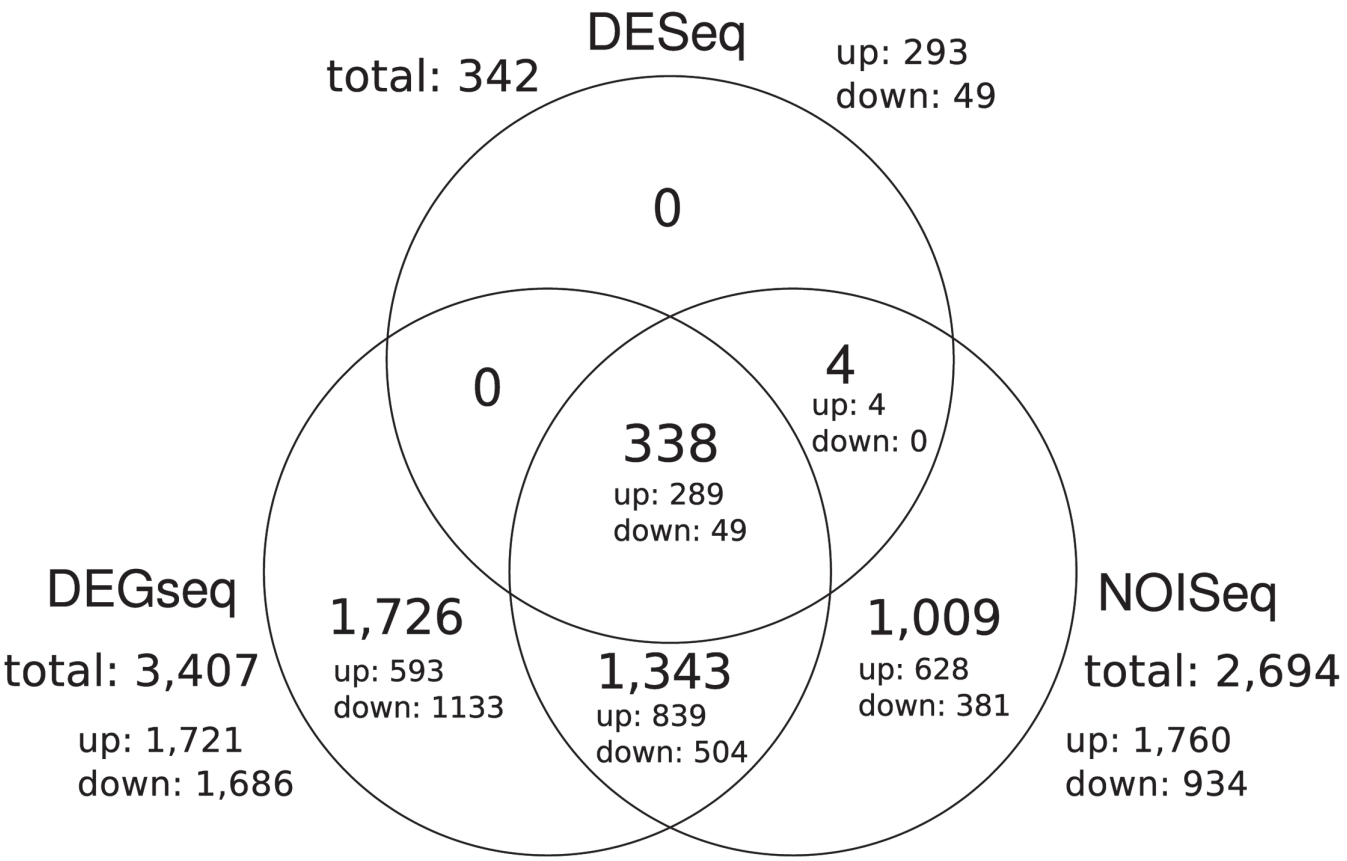

Fig 3. Venn diagram of the overlapping DE transcripts from the DEGseq, DESeq and NOISeq analyses. The amount of up- and down-regulated transcripts is given for each segment.

doi:10.1371/journal.pone.0124564.g003

listed (Table 3). All these selected genes were represented in the NOISeq results, while DESeq did not identify eight up-regulated and seven down-regulated genes as DE transcripts.

The literature search revealed twelve reference genes from published conifer studies (Table S9 in S1 File). Profiling the MACE dataset revealed three potential reference genes: the

Table 2. Differentially expressed transcripts resulting from the MACE and DEGseq analyses, filtered by log_2 fold change and minimum different sense tags ($\geq$ 50).

	Up-regulated	Down-regulated	Total
With database hit[a]	330	125	455
Assigned to biological process	217	77	294
Not assigned to biological process	113	48	161
Without database hit	253	124	377
Total	583	249	832

Shown is the amount of transcripts with and without database hit and significant assignment to the GO domain biological process (enrichment-*p*-value < 1e-10).

[a]'Database hit' refers to transcripts with database accession number or similarity to a UniProt Reference Cluster (UniRef) sequence.

doi:10.1371/journal.pone.0124564.t002

Table 3. Genes for validation derived from the DEGseq analysis; sorted by log_2 fold change in descending order.

Gene name	Abbreviation	Protein name	Possible isoform	Accession number	Source organism	Fold change
-	-	Polyphenol oxidase A1	-	Q06215	*Vicia faba*	10.61
Os01g0656200	-	Probable protein phosphatase 2C 8	-	Q5SN75	*Oryza sativa subsp. japonica*	9.55
-	-	Glucan endo-1,3-beta-glucosidase, acidic isoform	-	P49237	*Zea mays*	9.24
At4g33300	-	Probable disease resistance protein At4g33300	-	Q9SZA7	*Arabidopsis thaliana*	8.63
dhn2	-	Dehydrin 2 (fragment)	-	E1A556	*Pinus pinaster*	8.56
-	*PUP2*	Putative uncharacterized protein 2	-	A9NPH4	*Picea sitchensis*	8.50
CXE15 or CXE2	-	Probable carboxylesterase 15 or 2	-	Q9FG13 or Q9SX78	*A. thaliana*	8.42
Cht8	-	Chitinase 8	-	Q7XCK6	*O. sativa subsp. japonica*	8.37
GOLS2	-	Galactinol synthase 2	*b*	Q9FXB2	*A. thaliana*	8.27
GSTU19	-	Glutathione S-transferase U19	-	Q9ZRW8	*A. thaliana*	6.82
BAM1	-	Beta amylase 1, chloroplastic	*a*	Q9LIR6	*A. thaliana*	6.74
TIP1-1	-	Aquaporin TIP1-1	-	P25818	*A. thaliana*	5.75
LTI6B	-	Hydrophobic protein LTI6B	*a*	Q0DKW8	*O. sativa subsp. japonica*	5.63
STP13	-	Sugar transport protein 13	-	Q94AZ2	*A. thaliana*	5.52
RHA2A	-	E3 ubiquitin-protein ligase RHA2A	*a*	Q9ZT50	*A. thaliana*	5.36
GOLS1	-	Galactinol synthase 1	-	Q947G8	*Solanum lycopersicum*	4.94
LEA14-A[a]	-	LEA protein Lea14-A	-	P46518	*Gossypium hirsutum*	4.88
GOLS2	-	Galactinol synthase 2	*a*	C7G304	*S. lycopersicum*	4.60
LTI6B	-	Hydrophobic protein LTI6B	*b*	Q0DKW8	*O. sativa subsp. japonica*	4.52
RCI2A[a]	-	Hydrophobic protein RCI2A	-	Q9ZNQ7	*A. thaliana*	4.02
CIPK17[a]	-	CBL-interacting protein kinase 17	-	Q75L42	*O. sativa subsp. japonica*	3.73
-	*PUP1*	Putative uncharacterized protein 1	*b*	A9NLY4	*P. sitchensis*	3.55
Zlp[a]	-	Zeamatin	-	P33679	*Z. mays*	3.51
-	*PUP1**	Putative uncharacterized protein 1	*a*	A9NLY4	*P. sitchensis*	3.50
BAM[a]*	-	Beta amylase 1, chloroplastic	*b*	Q9LIR6	*A. thaliana*	3.33
ERD10[a]*	-	Dehydrin ERD10	-	P42759	*A. thaliana*	3.29
KCS11	-	3-ketoacyl-CoA synthase 11	*a*	O48780	*A. thaliana*	3.20
-	-	LEA protein[a]	-	P21298	*Raphanus sativus*	3.16
MGL[a]	-	Methionine gamma-lyase	-	Q9SGU9	*A. thaliana*	3.00
TUBB8[a]	-	Tubulin beta-8 chain	-	P29516	*A. thaliana*	-3.04
UGT74E2[a]	-	UDP-glycosyltransferase 74E2	-	Q9SYK9	*A. thaliana*	-3.16
XTH6[a]	-	Probable xyloglucan endotransglucosylase/ hydrolase protein 6	-	Q8LF99	*A. thaliana*	-4.25
KCS11[a]	-	3-ketoacyl-CoA synthase 11	*b*	O48780	*A. thaliana*	-5.69
GPPS3[a]	-	Geranyl diphosphate synthase	-	Q8LKJ1	*Abies grandis*	-5.92
-	*PUP5*[a]	Putative uncharacterized protein 5	-	B8LN73	*P. sitchensis*	-5.94
RHA2A	-	E3 ubiquitin-protein ligase RHA2A	*b*	Q9ZT50	*A. thaliana*	-6.02
-	*PUP4*[a]	Putative uncharacterized protein 4	-	C0PT89	*P. sitchensis*	-7.27
*XTH7**	-	Probable xyloglucan endotransglucosylase/ hydrolase protein 7	-	Q8LER3	*A. thaliana*	-7.57

(*Continued*)

Table 3. (*Continued*)

Gene name	Abbreviation	Protein name	Possible isoform	Accession number	Source organism	Fold change
COL6	-	Zinc finger protein CONSTANS-LIKE 6	-	Q8LG76	*A. thaliana*	-7.80
-	-	Patatin-like protein 3	-	B6TPQ5	*Z. mays*	-8.02
VTE4	-	Tocopherol O-methyltransferase	-	Q9ZSK1	*A. thaliana*	-8.72
INR1	-	Inducible nitrate reductase [NADH] 1	-	P54233	*Glycine max*	-9.01
-	*PUP3*	Putative uncharacterized protein 3	-	F6HZZ7	*Vitis vinifera*	-10.32

Gene names according to UniProt Protein Knowledgebase (http://www.uniprot.org/) with the corresponding database accession number.

- No gene name available or abbreviation assigned;

* Validated via RT-qPCR;

[a] Not identified as DE by DESeq.

doi:10.1371/journal.pone.0124564.t003

global transcription factor group E8 (*GTE8*), the transcription initiation factor 2A (*TIF2A*) and the two pore calcium channel protein 1 (*TPC1*). All three exhibited medium abundance levels and were therefore optimal candidates as reference genes. Out of the 15 potential reference genes, four (*GAPDH*, *TPC1*, 18S rRNA and *elF4A2*) specifically amplified their target in the RT-qPCR. The expression stability of these genes was tested using Normfinder [34] and geNorm [33]. Both algorithms identified *TPC1* and 18S rRNA as the most stable reference genes (Table S10 in S1 File). Furthermore, their combination exhibited the lowest accumulated standard deviation. Thus, *TPC1* and 18S rRNA were both applied for normalization in REST.

Out of the 50 tested primer pairs for validation (Table S2 in S1 File), four could meet our conservative criteria and were specifically amplifying their targets during the qPCR: *ERD10*, *BAM1(b)*, *XTH7* and *PUP1(a)* (Table S11 in S1 File). The calculated gene expression ratios were similar to those obtained by the MACE method (Table 4). The assay precision was assessed by calculating the intra-assay variation (repeatability) and inter-assay variation (reproducibility) based on Cq-values. The intra-assay coefficient of variance ranged between 0.8% and 1.08% while the inter-assay coefficient of variance was 0.92%.

From the 338 transcripts identified unanimously by all Seq-analyses (Fig 1), 296 remained after filtering by minimum different sense tags (Table 5, FASTA files of all 296 transcripts available in S3 File). Almost half of these transcripts (~45%) were unknown and many of the transcripts with a database hit (~38% of the remaining ~55%) were not yet properly assigned.

Discussion

Analyzing the adaptive potential of plant species to drought stress is of crucial importance in the context of rapid climate change. For species with large genomes, such as conifers, where

Table 4. Log_2 fold changes of the four specifically amplified genes for validation resulting from the MACE analysis and RT-qPCR (calculated using REST with correction of amplification efficiencies).

Gene name	log_2 fold change MACE	log_2 fold change RT-qPCR ± SE	Pair-wise fixed reallocation randomization test *p*-value
BAM1(b)	3.33	2.50 ± 0.97	0.0052
ERD10	3.29	2.91 ± 1.36	0.0016
PUP1(a)	3.50	3.95 ± 2.60	0.0016
XTH7	-7.57	-3.83 ± 4.958	0.001

SE = standard error.

doi:10.1371/journal.pone.0124564.t004

Table 5. Properties of the sense-tag-filtered consensus transcripts identified by the three different Seq-analyses.

	Up-regulated	Down-regulated	Total
With database hit[a]	140	22	162
PUPs (*Picea sitchensis*)	43	5	48
PUPs (other species)	2	1	3
UPs	7	1	8
Similarity to UniRef sequence	3	0	3
Without database hit	107	27	134
Total	247	49	296

PUP: Putative uncharacterized protein; UP: Uncharacterized protein.

[a]'Database hit' refers to transcripts with database accession number or similarity to a UniProt Reference Cluster (UniRef) sequence.

doi:10.1371/journal.pone.0124564.t005

genomic resources are scarce, it is often necessary to reduce the pool of target genes to an affordable size. Combining a new water monitoring setup with the MACE technique we were able to link a standardized phenotypic response to specific genes and thereby identify novel candidate genes for drought stress response in *A. alba*.

For our approach, we pooled RNA from individual seedlings for each treatment group and subsequently analyzed both pools via transcriptome sequencing. To exclude bias by individual expression patterns we applied RT-qPCR to validate the differential expression for a subset of genes known to be involved in the drought stress response of model plant species. Among these genes, only *UGT74E2* displayed a gene-regulation pattern that clearly differed from expectation according to the respective literature. In *Arabidopsis*, an ectopic over-expression of *UGT74E2* increased the tolerance to salinity and drought stress and reduced the plants' water loss [39]. Therefore, one would assume an up-regulation in response to drought stress. However, in silver fir the opposite was the case. Hence, *UGT74E2* might play a different physiological role in the phylogenetically distant silver fir, compared to *Arabidopsis*. Further research regarding the function of *UGT74E2* in other conifer taxa are necessary to offer an explanation for the different expression in response to drought stress.

The up-regulation of *PUP1(a)*, *BAM1(b)* and *ERD10* based on the MACE technique and subsequent DEGseq analysis could be verified by the RT-qPCR. Individual differences among the seedlings were expected but proved to be low for *BAM1(b)* and *ERD10* and moderate in the case of *PUP1(a)*. Though varying to a larger extent, the overall down-regulation of *XTH7* could be affirmed. However, DEGseq predicted a much higher gene expression ratio than observed with RT-qPCR. Individual differences were more pronounced for *XTH7* and may partially be attributed to PCR inhibition and/or stochastic cDNA template variation.

In order to select adequate reference genes for the RT-qPCR, we tested traditional reference genes which were previously used in other conifer gene expression studies. Depending on the treatment, ontogenetic stage or the tissue under investigation some of them showed expression stability [40–42], even though in other studies [33,40,42,43] these genes showed significant variability in expression patterns. Here we tested *GAPDH*, *elF4A2* and 18S rRNA, with only 18S rRNA showing expression stability across all individual seedlings. Since the expression of *GAPDH* and *eIF4A2* varied among the drought stressed and well-watered seedlings, these genes were not suitable as internal controls. However, 18S rRNA should not be used as the only reference gene for normalization due to the possible mismatch of rRNA and mRNA abundance [44]. Fortunately, the selection of reference genes based on the MACE results proved to

be successful. We could verify the expression stability of *TPC1* for all individual seedlings using RT-qPCR. In *Arabidopsis* and rice *TPC1* is known to be ubiquitously expressed across several tissues [45,46]. However, Wang *et al.* [47] identified a *TPC1* homologous gene which was induced in *Triticum aestivum* as a response to high salinity, polyethylene glycol (PEG), low temperature and abscisic acid treatment. They suggested an important role of *TPC1* in the stomatal closure and abiotic stress response of *T. aestivum*. On the one hand this implies the necessity for further investigation on the expression stability of *TPC1* as a potential reference gene when investigating drought stress response. On the other hand the *TPC1* homologue might be involved in the osmotic rather than the drought stress response. Since the expression stability of *TPC1* was only tested in cotyledons, future studies need to address other tissues such as roots, as well as different age stages of needles.

Our study design lacked biological replicates, which is a serious limitation but was dampened by pooling the samples for the two libraries. Pooling samples for RNA-Seq analysis has proven to be a reliable method for estimating gene expression, especially for genes exhibiting high expression levels [48]. Furthermore, the THz measurements were highly precise which ensured that the pools had very homogenous stress levels [12]. It is also notable that the THz approach enabled us to measure a stress response solely induced by water deprivation. Other approaches, such as PEG treatment, might largely lead to the differential expression of genes involved in osmotic stress response rather than in the response to water shortage. Since the individuals in our study were pooled in two treatment groups we could not estimate the biological variation within those groups. As stated in other studies facing the same problem [49–51], the results must be taken cautiously and the candidates should be further examined. Nonetheless, in order to analyze our data, we chose a conservative approach. Therefore, apart from DEGseq, we additionally employed DEseq and NOISeq. All three analyses showed different results as was expected according to comparative studies of the used methods [52,53]. DESeq is generally more conservative, while DEGseq and NOISeq are more aggressive but prone to false positives. However, DESeq did not identify three of the four genes verified by RT-qPCR as DE transcripts (Fig 2). Since the genes for validation were selected by very strict criteria and the RT-qPCR was conducted on the individual seedlings and not on pooled plant material, we conclude that the DEGseq and NOISeq results likely include a relatively high amount of false positives, while the 43 genes for validation (Table 3) should be correctly defined as DE transcripts. Since these genes were representatives of the group of 832 filtered DE transcripts, we define the whole set as potential candidate genes for drought stress response for further studies. However, the most conservative selection would only include the 296 filtered consensus transcripts.

Some of these genes or close variants were previously identified or used as candidates in other studies regarding drought stress response in conifers. For example, xyloglucan endotransglycosylase/hydrolase (*AoXET1*) was down-regulated in needles and stems of *Pinus pinaster* seedlings in response to drought stress [54]. A glutathione S-transferase and chitinases (*cht1* and *cht2*) were up-regulated in response to drought- and pathogen-related-stress in roots and shoots of six-week-old seedlings of *Picea abies* [55]. Velasco-Conde et al. [56] measured the expression pattern of several dehydrins (*dhn1*, *dhn2*, *dhn3*, *dhn7*, *dhn9* and a *dhn*-like protein) in needles of 3-year-old cuttings of drought-sensitive and drought-resistant genotypes of *P. pinaster*. Only *dhn3* and *dhn4* showed an involvement in drought resistance between genotypes, while *dhn2* was consistently down-regulated in response to drought stress. Our results suggest that *dhn2* might play a different role in drought stress response in *A. alba*, since it was significantly up-regulated. Dehydrins (*dhn1* and *dhn2*), aquaporin (*aquaMIP*) and early responsive to dehydration 3 (*erd3*) were used as candidate genes for drought stress response in megagametophytes of *Pinus taeda* [57]. Similarly, a putative glucan-endo-1,3-beta-glucosidase precursor, dehydrins (*dhn1* and *dhn2*) and *erd3* were used as candidates for outlier analyses in

megagametophytes of *P. pinaster* [58]. Chitinase 4 and a putative LEA protein were identified as drought stress responsive in needles, stems and roots of *P. pinaster* [59] and a glutathione S-transferase in needles of adult *Pinus halepensis* trees [50].

To our knowledge, no studies exist, which aimed to identify adaptive genes for drought stress response in any member of the genus *Abies*. However, such an analysis would surely benefit from genus-specific candidate genes. Here we present a selection of 296 genes that contains previously identified candidates but predominantly adds to the possible selection of candidates for future studies. Many of these genes are linked by their biological function to a specific response to water deprivation. For example, 3-ketoacyl-CoA synthase 11 belongs to the group of "very long chain fatty acids", which are required for wax synthesis [60]. The leaf cuticle is protected by the wax layer against non-stomatal water loss, which could explain its up-regulation. Also involved in water management are aquaporins, which are expressed very variably in response to drought stress [61]. Tonoplast intrinsic proteins (TIP1s) are usually found in the lytic vacuole membrane [62]. Hence, *TIP1-1* is probably up-regulated during drought periods to allow better access to the water stored in the vacuole. Chloroplastic beta-amylase 1 is involved in the breakdown of leaf starch [63]. During daytime, plants store glucose as starch in chloroplasts and access this energy during nighttimes via starch breakdown. Up-regulation of *BAM1* is most likely a response to the down-regulation of metabolic processes and especially photosynthesis during drought periods (Fig 1). Analogous to "regular" nighttimes, metabolism with reduced photosynthesis can only be maintained by breaking down the energy storage, e.g. starch. Accordingly, sugar transport proteins, which transport hexoses through cellular membranes, are necessary for metabolism but may also play a role in distributing osmolytes throughout the plant. The down-regulation of *XTH6* and *XTH7* indicates limited cell growth during drought periods, since xyloglucan endotransglucosylase/hydrolases are involved in cell enlargement and restructuring [64]. Protein kinases add phosphate groups to a substrate, while protein phosphatases remove them, thus either activating or deactivating enzymes [65]. Both protein kinases and phosphatases are key factors in signal transduction as response to drought stress, by regulating enzyme activity. E3 ubiquitin-protein ligase RHA2A functions in an ABA-mediated signaling pathway during early seedling development, positively regulating the plants response to osmotic stress [66]. The fact that *RHA2A* is part of the subset of both up- and down-regulated candidate genes highlights the necessity for a distinction between possibly different protein isoforms for all genes. Late embryogenesis abundant (LEA) proteins play a protective role against desiccation-damage during drought periods, presumably by suppressing protein aggregation [67]. Dehydrins were initially categorized as "Group II LEA proteins" and indeed protect plant cells from desiccation-damage but are also involved in pathogen resistance [68]. Galactinol synthases are induced by drought, cold and ABA [69] and are interesting from the perspective of adaptation. Taji et al. [70] found that genes encoding galactinol synthase in transgenic *Arabidopsis* improved drought stress tolerance, which might be attributable to the role of galactinol synthase in the biosynthesis of "raffinose family oligosaccharides". The resulting accumulation of galactinol and raffinose may enhance drought stress tolerance via osmoprotection. Chitinase 8, the acidic isoform of glucan endo-1,3-beta-glucosidase and zeamatin are all involved in pathogen defense response, mainly against fungi [71–73] and in the case of probable disease resistance protein At4g33300 against bacteria [74]. Reason for the up-regulation of pathogen-resistance genes in silver fir in response to drought might be that forest trees are especially prone to drought-disease interactions, mainly involving fungi [21]. Glutathione S-transferases are most notably detoxification enzymes with many other, still unknown, involvements hypothesized in plant stress response [75]. As such, they are classified as early responsive to dehydration (ERD), i.e. genes that are activated swiftly in response to drought stress, a group that also contains dehydrins [76].

A group of gene products that stands out are the putative uncharacterized proteins (*PUPs*) inferred from the transcriptome sequencing of *Picea sitchensis* [77]. *PUP1* belongs to the dehydrin family, *PUP2* to the NAC domain and *PUP4* as well as *PUP5* to the family of UDP glycosyltransferases (UGT), according to the *UniProt Protein Knowledgebase* (http://www.uniprot.org/). All *PUPs* identified in this study seem to play an important role in the drought stress response of silver fir and should be focused on in further studies. Correspondingly, further research should focus on the DE transcripts that did not have a database hit, since these genes are not yet described for any plant species and are most likely involved in drought stress response. As such, they might possibly be specific to the Pinaceae family or conifers in general.

In conclusion our study provides first insights into the drought stress response of *A. alba* at the transcriptome level and offers a set of candidate genes for use in future studies. The majority of these candidates are yet unknown or lack a proper assignment and add to the growing genomic resources available for non-model conifer species. Such resources will be increasingly important for investigating the adaptive potential of long-lived organisms such as trees in the face of rapid climate change.

Supporting Information

S1 File. File containing all supporting tables.
(DOCX)

S2 File. FASTA files of all 832 DE transcripts identified by DEGseq and filtered by different sense tags ($\geq$ 50) and $\log_2$ fold change ($<$ -3 or $>$ 3).
(GZ)

S3 File. FASTA files of all 296 DE transcripts identified unanimously by DEGseq, DESeq and NOISeq and filtered by different sense tags ($\geq$ 50).
(GZ)

Acknowledgments

We would like to thank Björn Rotter and Ralf Horres from GenXPro GmbH for advice and help with the MACE data, as well as Norman Born and Martin Koch from the Faculty of Physics and Material Sciences Center at the University of Marburg for providing access to and advising on the THz setup. Furthermore we would like to express our thanks to Christina Mengel from the Conservation Biology group at the University of Marburg for assisting in the lab work.

Author Contributions

Conceived and designed the experiments: DB HZ BZ SL. Performed the experiments: DB HZ. Analyzed the data: DB HZ. Wrote the paper: DB HZ BZ SL.

References

1. IPCC. Climate Change 2007: Working Group I: The Physical Basis. Contribution to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Chapter 11: Regional Climate Projections. 2007.
2. Allen CD, Macalady AK, Chenchouni H, Bachelet D, McDowell N, Vennetier M, et al. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. Forest Ecology and Management. 2010; 259:660–684. doi: 10.1016/j.foreco.2009.09.001
3. Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S. Adaptation, migration or extirpation: climate change outcomes for tree populations. Evolutionary Applications. 2008; 1:95–111. doi: 10.1111/j.1752-4571.2007.00013.x PMID: 25567494

4. Pearson RG, Dawson TP. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography. 2003; 12:361–371. doi: 10.1046/j.1466-822X.2003.00042.x
5. Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends in Ecology & Evolution. 2008; 23:38–44. doi: 10.1016/j.tree.2007.09.008
6. Neale DB, Kremer A. Forest tree genomics: growing resources and applications. Nat Rev Genet. 2011; 12:111–122. doi: 10.1038/nrg2931 PMID: 21245829
7. Murray BG. Nuclear DNA Amounts in Gymnosperms. Ann Bot. 1998; 82:3–15.
8. Neale DB, Savolainen O. Association genetics of complex traits in conifers. Trends in Plant Science. 2004; 9:325–330. doi: 10.1016/j.tplants.2004.05.006 PMID: 15231277
9. Eveno E, Collada C, Guevara MA, Léger V, Soto A, Díaz L, et al. "Contrasting Patterns of Selection at Pinus pinaster Ait. Drought Stress Candidate Genes as Revealed by Genetic Differentiation Analyses." Mol Biol Evol. 2008; 25:417–437. doi: 10.1093/molbev/msm272 PMID: 18065486
10. Müller T, Freund F, Wildhagen H, Schmid KJ. Targeted re-sequencing of five Douglas-fir provenances reveals population structure and putative target genes of positive selection. Tree Genetics & Genomes. 2015; 11:1–17. doi: 10.1007/s11295-014-0816-z
11. Sork VL, Aitken SN, Dyer RJ, Eckert AJ, Legendre P, Neale DB. Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. Tree Genetics & Genomes. 2013; 9:901–911. doi: 10.1007/s11295-013-0596-x
12. Born N, Behringer D, Liepelt S, Beyer S, Schwerdtfeger M, Ziegenhagen B, et al. Monitoring Plant Drought Stress Response Using Terahertz Time-Domain Spectroscopy. Plant Physiol. 2014; 164:1571–1577. doi: 10.1104/pp.113.233601 PMID: 24501000
13. Konnert M, Bergmann F. The geographical distribution of genetic variation of silver fir (Abies alba, Pinaceae) in relation to its migration history. Pl Syst Evol. 1995; 196:19–30. doi: 10.1007/BF00985333
14. Macias M, Andreu L, Bosch O, Camarero JJ, Gutiérrez E. Increasing Aridity is Enhancing Silver Fir Abies Alba Mill.) Water Stress in its South-Western Distribution Limit. Climatic Change. 2006; 79:289–313. doi: 10.1007/s10584-006-9071-0
15. Lebourgeois F. Climatic signal in annual growth variation of silver fir (Abies alba Mill.) and spruce (Picea abies Karst.) from the French Permanent Plot Network (RENECOFOR). Ann For Sci. 2007; 64:333–343. doi: 10.1051/forest:20070101
16. Toromani E, Sanxhaku M, Pasho E. Growth responses to climate and drought in silver fir (Abies alba) along an altitudinal gradient in southern Kosovo. Can J For Res. 2011; 41:1795–1807. doi: 10.1139/x11-096
17. Peguero-Pina JJ, Camarero JJ, Abadía A, Martín E, González-Cascón R, Morales F, et al. Physiological performance of silver-fir (Abies alba Mill.) populations under contrasting climates near the southwestern distribution limit of the species. Flora—Morphology, Distribution, Functional Ecology of Plants. 2007; 202:226–236. doi: 10.1016/j.flora.2006.06.004
18. Piovani P, Leonardi S, Magnani F, Menozzi P. Variability of stomatal conductance in a small and isolated population of silver fir (Abies alba Mill.). Tree Physiol. 2011; 31:500–507. doi: 10.1093/treephys/tpr029 PMID: 21636691
19. Linares JC, Camarero JJ. Silver Fir Defoliation Likelihood Is Related to Negative Growth Trends and High Warming Sensitivity at Their Southernmost Distribution Limit. ISRN Forestry. 2012; 2012: e437690. doi: 10.5402/2012/437690
20. Cailleret M, Nourtier M, Amm A, Durand-Gillmann M, Davi H. Drought-induced decline and mortality of silver fir differ among three sites in Southern France. Annals of Forest Science. 2013;1–15. doi: 10.1007/s13595-013-0265-0
21. Desprez-Loustau M-L, Marçais B, Nageleisen L-M, Piou D, Vannini A. Interactive effects of drought and pathogens in forest trees. Annals of Forest Science. 2006; 63:597–612. doi: 10.1051/forest:2006040
22. Durand-Gillmann M, Cailleret M, Boivin T, Nageleisen L-M, Davi H. Individual vulnerability factors of Silver fir (Abies alba Mill.) to parasitism by two contrasting biotic agents: mistletoe (Viscum album L. ssp. abietis) and bark beetles (Coleoptera: Curculionidae: Scolytinae) during a decline process. Annals of Forest Science. 2012;1–15. doi: 10.1007/s13595-012-0251-y
23. Nourtier M, Chanzy A, Cailleret M, Yingge X, Huc R, Davi H. Transpiration of silver Fir (Abies alba mill.) during and after drought in relation to soil properties in a Mediterranean mountain area. Annals of Forest Science. 2012;1–13. doi: 10.1007/s13595-012-0229-9
24. Hansen OK, Vendramin GG, Sebastiani F, Edwards KJ. Development of microsatellite markers in Abies nordmanniana (Stev.) Spach and cross-species amplification in the Abies genus. Molecular Ecology Notes. 2005; 5:784–787. doi: 10.1111/j.1471-8286.2005.01062.x

25. Kahl G, Molina C, Rotter B, Jüngling R, Frank A, Krezdorn N, et al. Reduced representation sequencing of plant stress transcriptomes. J Plant Biochem Biotechnol. 2012; 21:119–127. doi: 10.1007/s13562-012-0129-y
26. Yakovlev I, Fossdal CG, Skrøppa T, Olsen JE, Jahren AH, Johnsen Ø. An adaptive epigenetic memory in conifers with important implications for seed production. Seed Science Research. 2012; 22:63–76. doi: 10.1017/S0960258511000535
27. Lenz TL, Eizaguirre C, Rotter B, Kalbe M, Milinski M. Exploring local immunological adaptation of two stickleback ecotypes by experimental infection and transcriptome-wide digital gene expression analysis. Molecular Ecology. 2013; 22:774–786. doi: 10.1111/j.1365-294X.2012.05756.x PMID: 22971109
28. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010; 26:136–138. doi: 10.1093/bioinformatics/btp612 PMID: 19855105
29. Rivals I, Personnaz L, Taing L, Potier M-C. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics. 2007; 23:401–407. doi: 10.1093/bioinformatics/btl633 PMID: 17182697
30. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11:R106. doi: 10.1186/gb-2010-11-10-r106 PMID: 20979621
31. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Res. 2011; 21:2213–2223. doi: 10.1101/gr.124321.111 PMID: 21903743
32. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry. 2009; 55:611–622. doi: 10.1373/clinchem.2008.112797 PMID: 19246619
33. Vandesompele J, Preter KD, Pattyn F, Poppe B, Roy NV, Paepe AD, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biology. 2002; 3:research0034. doi: 10.1186/gb-2002-3-7-research0034
34. Andersen CL, Jensen JL, Ørntoft TF. Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets. Cancer Res. 2004; 64:5245–5250. doi: 10.1158/0008-5472.CAN-04-0496 PMID: 15289330
35. Kubista M, Rusnakova V, Sec D, Sjögreen B, Tichopad A. GenEx: Data Analysis Software. Quantitative Real-time PCR in Applied (ed Filion M) Microbiology. Horizon Scientific Press; 2012. pp. 63–84.
36. Pfaffl MW, Horgan GW, Dempfle L. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucl Acids Res. 2002; 30: e36–e36. doi: 10.1093/nar/30.9.e36 PMID: 11972351
37. Liu W, Saint DA. A New Quantitative Method of Real Time Reverse Transcription Polymerase Chain Reaction Assay Based on Simulation of Polymerase Chain Reaction Kinetics. Analytical Biochemistry. 2002; 302:52–59. doi: 10.1006/abio.2001.5530 PMID: 11846375
38. Pfaffl MW. 3.2. Markers of a Successful Real-Time RT-PCR Assay. A-Z of quantitative PCR (ed SA Bustin). La Jolla, CA: International University Line; 2004. pp. 90–106.
39. Tognetti VB, Aken OV, Morreel K, Vandenbroucke K, Cotte B van de, Clercq ID, et al. Perturbation of Indole-3-Butyric Acid Homeostasis by the UDP-Glucosyltransferase UGT74E2 Modulates Arabidopsis Architecture and Water Stress Tolerance. Plant Cell. 2010; 22:2660–2679. doi: 10.1105/tpc.109.071316 PMID: 20798329
40. Brunner AM, Yakovlev IA, Strauss SH. Validating internal controls for quantitative plant gene expression studies. BMC Plant Biology. 2004; 4:14. doi: 10.1186/1471-2229-4-14 PMID: 15317655
41. Nicot N, Hausman J-F, Hoffmann L, Evers D. Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. J Exp Bot. 2005; 56:2907–2914. doi: 10.1093/jxb/eri285 PMID: 16188960
42. Palovaara J, Hakman I. Conifer WOX-related homeodomain transcription factors, developmental consideration and expression dynamic of WOX2 during Picea abies somatic embryogenesis. Plant Mol Biol. 2008; 66:533–549. doi: 10.1007/s11103-008-9289-5 PMID: 18209956
43. Gonçalves S, Cairney J, Maroco J, Oliveira MM, Miguel C. Evaluation of control transcripts in real-time RT-PCR expression analysis during maritime pine embryogenesis. Planta. 2005; 222:556–563. doi: 10.1007/s00425-005-1562-0 PMID: 16034587
44. Tenea GN, Bota AP, Raposo FC, Maquet A. Reference genes for gene expression studies in wheat flag leaves grown under different farming conditions. BMC Research Notes. 2011; 4:373. doi: 10.1186/1756-0500-4-373 PMID: 21951810

45. Furuichi T, Cunningham KW, Muto S. A Putative Two Pore Channel AtTPC1 Mediates Ca2+ Flux in Arabidopsis Leaf Cells. Plant Cell Physiol. 2001; 42:900–905. doi: 10.1093/pcp/pce145 PMID: 11577183

46. Kurusu T, Yagala T, Miyao A, Hirochika H, Kuchitsu K. Identification of a putative voltage-gated Ca2+ channel as a key regulator of elicitor-induced hypersensitive cell death and mitogen-activated protein kinase activation in rice. The Plant Journal. 2005; 42:798–809. doi: 10.1111/j.1365-313X.2005.02415.x PMID: 15941394

47. Wang Y-J, Yu J-N, Chen T, Zhang Z-G, Hao Y-J, Zhang J-S, et al. Functional analysis of a putative Ca2 + channel gene TaTPC1 from wheat. J Exp Bot. 2005; 56:3051–3060. doi: 10.1093/jxb/eri302 PMID: 16275671

48. Konczal M, Koteja P, Stuglik MT, Radwan J, Babik W. Accuracy of allele frequency estimation using pooled RNA-Seq. Mol Ecol Resour. 2014; 14:381–392. doi: 10.1111/1755-0998.12186 PMID: 24119300

49. Pauletto M, Milan M, Moreira R, Novoa B, Figueras A, Babbucci M, et al. Deep transcriptome sequencing of Pecten maximus hemocytes: A genomic resource for bivalve immunology. Fish & Shellfish Immunology. 2014; 37: 154–165. doi: 10.1016/j.fsi.2014.01.017

50. Pinosio S, González-Martínez SC, Bagnoli F, Cattonaro F, Grivet D, Marroni F, et al. First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, Pinus halepensis Mill. Molecular Ecology Resources. 2014;n/a–n/a. doi: 10.1111/1755-0998.12232

51. Zhao X, Yu H, Kong L, Liu S, Li Q. Comparative Transcriptome Analysis of Two Oysters, Crassostrea gigas and Crassostrea hongkongensis Provides Insights into Adaptation to Hypo-Osmotic Conditions. PLoS ONE. 2014; 9: e111915. doi: 10.1371/journal.pone.0111915 PMID: 25369077

52. Zheng X, Moriyama EN. Comparative studies of differential gene calling using RNA-Seq data. BMC Bioinformatics. 2013; 14:S7. doi: 10.1186/1471-2105-14-S13-S7 PMID: 24564719

53. Guo Y, Li C-I, Ye F, Shyr Y. Evaluation of read count based RNAseq analysis methods. BMC Genomics. 2013; 14:S2. doi: 10.1186/1471-2164-14-S8-S2 PMID: 24341380

54. Dubos C, Plomion C. Identification of water-deficit responsive genes in maritime pine (Pinus pinaster Ait.) roots. Plant Mol Biol. 2003; 51:249–262. doi: 10.1023/A:1021168811590 PMID: 12602883

55. Fossdal CG, Nagy NE, Johnsen Ø, Dalen LS. Local and systemic stress responses in Norway spruce: Similarities in gene expression between a compatible pathogen interaction and drought stress. Physiological and Molecular Plant Pathology. 2007; 70:161–173. doi: 10.1016/j.pmpp.2007.09.002

56. Velasco-Conde T, Yakovlev I, Majada JP, Aranda I, Johnsen Ø. Dehydrins in maritime pine (Pinus pinaster) and their expression related to drought stress response. Tree Genetics & Genomes. 2012; 8:957–973. doi: 10.1007/s11295-012-0476-9

57. González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB. DNA Sequence Variation and Selection of Tag Single-Nucleotide Polymorphisms at Candidate Genes for Drought-Stress Response in Pinus taeda L. Genetics. 2006; 172:1915–1926. doi: 10.1534/genetics.105.047126 PMID: 16387885

58. Eveno E, Collada C, Guevara MA, Léger V, Soto A, Díaz L, et al. "Contrasting Patterns of Selection at Pinus pinaster Ait. Drought Stress Candidate Genes as Revealed by Genetic Differentiation Analyses." Mol Biol Evol. 2008; 25:417–437. doi: 10.1093/molbev/msm272 PMID: 18065486

59. Perdiguero P, Collada C, Barbero M del C, García Casado G, Cervera MT, Soto Á. Identification of water stress genes in Pinus pinaster Ait. by controlled progressive stress and suppression-subtractive hybridization. Plant Physiology and Biochemistry. 2012; 50:44–53. doi: 10.1016/j.plaphy.2011.09.022 PMID: 22099518

60. Todd J, Post-Beittenmiller D, Jaworski JG. KCS1encodes a fatty acid elongase 3-ketoacyl-CoA synthase affecting wax biosynthesis inArabidopsis thaliana. The Plant Journal. 1999; 17:119–130. doi: 10.1046/j.1365-313X.1999.00352.x59 PMID: 10074711

61. Hamanishi ET, Campbell MM. Genome-wide responses to drought in forest trees. Forestry. 2011; doi: 10.1093/forestry/cpr012

62. Maurel C, Verdoucq L, Luu D-T, Santoni V. Plant Aquaporins: Membrane Channels with Multiple Integrated Functions. Annual Review of Plant Biology. 2008; 59:595–624. doi: 10.1146/annurev.arplant.59.032607.092734 PMID: 18444909

63. Fulton DC, Stettler M, Mettler T, Vaughan CK, Li J, Francisco P, et al. β-AMYLASE4, a Noncatalytic Protein Required for Starch Breakdown, Acts Upstream of Three Active β-Amylases in Arabidopsis Chloroplasts. Plant Cell. 2008; 20:1040–1058. doi: 10.1105/tpc.107.056507 PMID: 18390594

64. Cosgrove DJ. Growth of the plant cell wall. Nature Reviews Molecular Cell Biology. 2005; 6:850–861. doi: 10.1038/nrm1746 PMID: 16261190

65. Singh A, Giri J, Kapoor S, Tyagi AK, Pandey GK. Protein phosphatase complement in rice: genome-wide identification and transcriptional analysis under abiotic stress conditions and reproductive development. BMC Genomics. 2010; 11:435. doi: 10.1186/1471-2164-11-435 PMID: 20637108

66. Bu Q, Li H, Zhao Q, Jiang H, Zhai Q, Zhang J, et al. The Arabidopsis RING Finger E3 Ligase RHA2a Is a Novel Positive Regulator of Abscisic Acid Signaling during Seed Germination and Early Seedling Development. Plant Physiol. 2009; 150:463–481. doi: 10.1104/pp.109.135269 PMID: 19286935

67. Goyal K, Walton LJ, Tunnacliffe A. LEA proteins prevent protein aggregation due to water stress. Biochemical Journal. 2005; 388:151. doi: 10.1042/BJ20041931 PMID: 15631617

68. Yang Y, He M, Zhu Z, Li S, Xu Y, Zhang C, et al. Identification of the dehydrin gene family from grapevine species and analysis of their responsiveness to various forms of abiotic and biotic stress. BMC Plant Biology. 2012; 12:140. doi: 10.1186/1471-2229-12-140 PMID: 22882870

69. Shinozaki K, Yamaguchi-Shinozaki K. Gene networks involved in drought stress response and tolerance. J Exp Bot. 2007; 58:221–227. doi: 10.1093/jxb/erl164 PMID: 17075077

70. Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, et al. Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in Arabidopsis thaliana. Plant J. 2002; 29:417–426. PMID: 11846875

71. Malehorn DE, Borgmeyer JR, Smith CE, Shah DM. Characterization and Expression of an Antifungal Zeamatin-like Protein (Zlp) Gene from Zea mays. Plant Physiol. 1994; 106:1471–1481. doi: 10.1104/pp.106.4.1471 PMID: 7846159

72. Wu S, Kriz AL, Widholm JM. Nucleotide Sequence of a Maize cDNA for a Class II, Acidic [beta]-1,3-Glucanase. Plant Physiol. 1994; 106:1709–1710. doi: 10.1104/pp.106.4.1709 PMID: 7846180

73. Witmer X, Nonogaki H, Beers EP, Bradford KJ, Welbaum GE. Characterization of chitinase activity and gene expression in muskmelon seeds. Seed Science Research. 2003; 13:167–178. doi: 10.1079/SSR2003134

74. Bonardi V, Tang S, Stallmann A, Roberts M, Cherkis K, Dangl JL. Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. PNAS. 2011; 108:16463–16468. doi: 10.1073/pnas.1113726108 PMID: 21911370

75. Sheehan D, Meade G, Foley VM, Dowd CA. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. Biochem J. 2001; 360:1–16. PMID: 11695986

76. Siqueira M, Gomes L. Functional Diversity of Early Responsive to Dehydration (ERD) Genes in Soybean. In: Board J, editor. A Comprehensive Survey of International Soybean Research—Genetics, Physiology, Agronomy and Nitrogen Relationships. InTech; 2013. Available: http://www.intechopen.com/books/a-comprehensive-survey-of-international-soybean-research-genetics-physiology-agronomy-and-nitrogen-relationships/functional-diversity-of-early-responsive-to-dehydration-erd-genes-in-soybean doi: 10.1016/j.yrtph.2010.06.017 PMID: 20615445

77. Ralph SG, Chun H, Kolosova N, Cooper D, Oddy C, Ritland CE, et al. A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis). BMC Genomics. 2008; 9:484. doi: 10.1186/1471-2164-9-484 PMID: 18854048

CHAPTER 5

Past stress responses archived in tree-rings associate with SNP genotypes in *Abies alba* (Mill.)

Katrin Heer, David Behringer, Alma Piermattei, Claus Bässler, Bruno Fady, Hans Jehl, Sascha Liepelt, Sven Lorch, Giovanni Guiseppe Vendramin, Max Weller, Birgit Ziegenhagen, Ulf Büntgen, Lars Opgenoorth

Manuscript (2017)

Past stress responses archived in tree-rings associate with SNP genotypes in *Abies alba* (Mill.)

Heer, Katrin*[1], Behringer, David*[1], Piermattei, Alma[2,3], Bässler, Claus[4], Fady, Bruno[5], Jehl, Hans[4], Liepelt, Sascha[1], Lorch, Sven[6], Vendramin, Giovanni Guiseppe[7], Weller, Max[6], Ziegenhagen, Birgit[1], Büntgen, Ulf[3,8,9], Opgenoorth, Lars[6]

* both authors contributed equally to this manuscript

[1] Philipps-University Marburg, Faculty of Biology, Conservation Biology, Marburg, Germany
[2] Marche Polytechnic University, Ancona, Italy
[3] DendroScience, Swiss Federal Research Institute WSL, Zürcherstrasse 111, 8903
[4] Bavarian Forest National Park, Freyunger Str. 2, Grafenau, Germany
[5] INRA, UR Ecologie des Forêts Méditerranéennes, Avignon, France
[6] Philipps-University Marburg, Faculty of Biology, Department of Ecology, Marburg, Germany
[7] Consiglio Nazionale delle Ricerche, Institute of Biosciences and Bioresources, Sesto Fiorentino, Firenze, Italy
[8] Department of Geography, University of Cambridge, Downing Place, CB2 3EN Cambridge, UK
[9] CzechGlobe, Global Change Research Institute CAS and Masaryk University, Kotlářská 2, 61137 Brno, Czech Republic

Abstract

Air pollution, especially containing sulfur dioxide (SO_2), is suspected to be the main contributor to the foliar damage and dieback of silver fir (*Abies alba* Mill.) populations in the 1970s and 1980s in Germany. In combination with an increased sensitivity to drought, caused by SO_2, this led to a marked decrease in radial growth in many silver fir trees. This growth depression period is archived in the annual tree-ring data, which is usually studied on the population level. We derived 'dendrophenotypes' that characterize resistance, resilience and recovery during the depression period based on individual tree-ring widths of silver fir trees from stands at two elevations in the Bavarian Forest National Park in Germany. Our goal was to associate genetic variation, in the form of SNPs, with variation in the dendrophenotypes to identify candidate genes for potentially adaptively relevant stress responsive traits in silver fir. Using feature selection techniques based on the machine learning algorithm random forest, we could identify 15 candidate genes whose products are mostly involved in photosynthetic or chloroplast development and some in drought response. This shows that individual-level dendrophenotypes are a valuable measure for genetic association studies in forest trees and can strongly increase our understanding of the genetic basis of environmental stress response, specifically to extreme episodic events.

Keywords: Air pollution, silver fir, candidate genes, random forest, SO_2

Introduction

In plants, genetic association studies often focus on reactions to environmental stress events. Extreme episodic stress events, such as drought, are of particular interest as they are expected to increase significantly during the 21st century due to human-induced global climate change (IPCC, 2014). In Germany, past environmental stress events have caused growth reduction and severe forest diebacks in many forest tree species (McLaughlin, 1985; Krause et al., 1986). One remarkable example were the diebacks in the 1970s and 1980s, where many forest stands, in particular silver fir (*Abies alba* Mill.) and Norway spruce (*Picea abies* (L.) Karst.), showed severe foliar damages and dieback. Initially this was attributed to drought alone but the rapid spread of the disease led to the consideration of an interactive effect of multiple agents, including air pollution and particularly sulfur dioxide (SO_2) and ozone (O_3) emissions (McLaughlin, 1985; Krause et al., 1986). This was further substantiated by the fact that the increase in SO_2 emissions in Europe correlated quite well with an observed decrease in radial growth in silver fir and Norway spruce for more than 20 years prior to the dieback events (McLaughlin, 1985).

For many surviving silver fir trees, the growth depression is well documented in the tree-ring width (TRW) during this episode (Büntgen et al., 2014). Further, there are indications that the physiological reactions to air pollution (especially SO_2) led to an increased sensitivity to drought stress in silver fir during this period (Elling et al., 2009). Thus, a number of particularly dry years in the 1970s and 1980s affected silver fir even in areas where drought is usually not a problem. Since then, silver fir has recovered and shows a strong growth increase in many areas. While the ultimate cause for the depression, as well as the recent increase in growth remain unclear, the latter has been attributed to the reduction in pollutants, the less dense forest structure after the dieback, as well as the elevated nitrogen availability and increasing temperature (Wilson and Elling, 2004; Büntgen et al., 2014).

A study from the Carpathian mountains found that two silver fir lineages differed in their reaction during the depression period and in subsequent growth increase (Bosela et al., 2016), and it seems reasonable to assume a genetic background for varying predispositions to air pollution. In addition, a number of progeny tests and common garden experiments documented that reactions to climatic extremes are influenced by the trees' genetic background at the level of provenances. King et al. (2013) accounted for that by adding a genetic component to Cook's (1985) linear aggregate model of TRW, and argued that the genetic component that influences the response to climatic conditions is of particular interest for local adaptation. However, most dendrochronological studies focus on mean signals at the level of stands or – in the case of common garden trials – of provenances while individual fluctuations are disregarded as noise (Carrer, 2011). We suggest that in genetic association studies this 'noise' can be seen as relevant signals of individual growth reactions.

A few studies have already jointly analyzed genetic and dendroecological data using molecular markers, and related basic genetic parameters such as heterozygosity (Babushkina et al., 2016), pairwise relatedness (King et al., 2013) or single amplified fragment length polymorphisms (AFLPs) (Pluess and Weber, 2012) with growth parameters. None of the studies found a strong genetic signal in their growth related parameters, which could either be attributed to stronger effects of the environmental signals compared to the genetic influence on growth processes, or to a lack of

adequate genetic data (e.g. loci that are relevant for the considered phenotypic traits).

In this study we utilize the fact that trees as long-lived organisms, archive their own physiological history in their annual growth rings (Cook and Kairiukstis, 2013). In the context of extreme environmental stress events, we focus on the individual reactions of trees, characterized by a number of measures derived from TRW data. We define all direct and derived measures that are based on TRW data as 'dendrophenotypes' in this study. In particular, we are interested in identifying whether differences in individual reactions are associated with genetic variation among individuals.

For this purpose, we derived dendrophenotypes for 193 silver fir trees and associated them with 130 SNPs in candidate genes mainly related to stress reactions and particularly to drought (Roschanski et al., 2015). Specifically, we focused on data of annual increment, characterizing the depression period in the 1970s and 1980s. For this purpose, we referred to the resilience concept applied to tree growth by Lloret et al. (2011). As growth is a quantitative trait, which is likely influenced by many genes, we not only used a single-locus approach for the genetic association, but supplemented them with random forest analyses to capture both the marginal effect of a SNP on a phenotype, as well as effects of multiple SNPs on a phenotype. Thereby we seek to determine the genetic basis of the observed dendrochronological patterns. This would support the utility of dendrophenotypes as an entry point to further explore the genetic basis of growth decline and resilience in stress scenarios in the context of climate change.

Material and Methods

Study site

Silver fir trees were sampled and monitored at two sampling sites in the Bavarian Forest National Park, Germany. Commercial forest management has been abandoned there since the foundation of the national park in 1970. The park covers 24,217 ha and its elevation ranges from 650 to 1,450 m a.s.l. Mean annual temperature varies between 3.8°C and 5.8°C with an annual precipitation total of around 1,200 mm to 1,800 mm. Our sampling sites were located at 770 m a.s.l. (Filzwald, 48.929°N, 13.406°E) and 1,120 m a.s.l (Rachelsee, 48.975°N, 13.400°E) on the Southern slope of Mt. Rachel (Fig. 1). Both sampling sites were chosen because of the relatively high abundance of mature silver fir trees within a mixed mountain forest (*Fagus sylvatica*, *Abies alba* and *Picea abies*), which is the typical forest composition for elevations below 1,150 m in the region. As silver fir has rarely been planted by forestry, and given that some of the trees have a confirmed age of more than 300 years, we are confident that the sample population is autochthonous.

At both sampling sites, 100 adult silver fir trees were geo-referenced and permanently marked with numbered tags. Temperature and humidity were recorded at both sampling sites with data loggers (DK320 DM HumiLog, Drießen & Kern, Bad Bramstedt, Germany) starting in spring 2014. The low elevation sampling site at Filzwald (hereafter referred to as 'low site') is characterized by flat terrain and subjected to accumulating cold air from higher elevations, which leads to frequent late and early frost events in spring and autumn, respectively. In contrast, the sampling site at Rachelsee (hereafter referred to as 'high site') is located on a steep slope surrounding the Rachel lake. The lake influences the local climatic conditions by buffering cold temperatures. Therefore,

early fall and late spring frost events are less frequent and maximum temperatures are lower at the high site compared to the low site. Mean temperature, however, did not significantly differ between sites during our study period (Table S1).

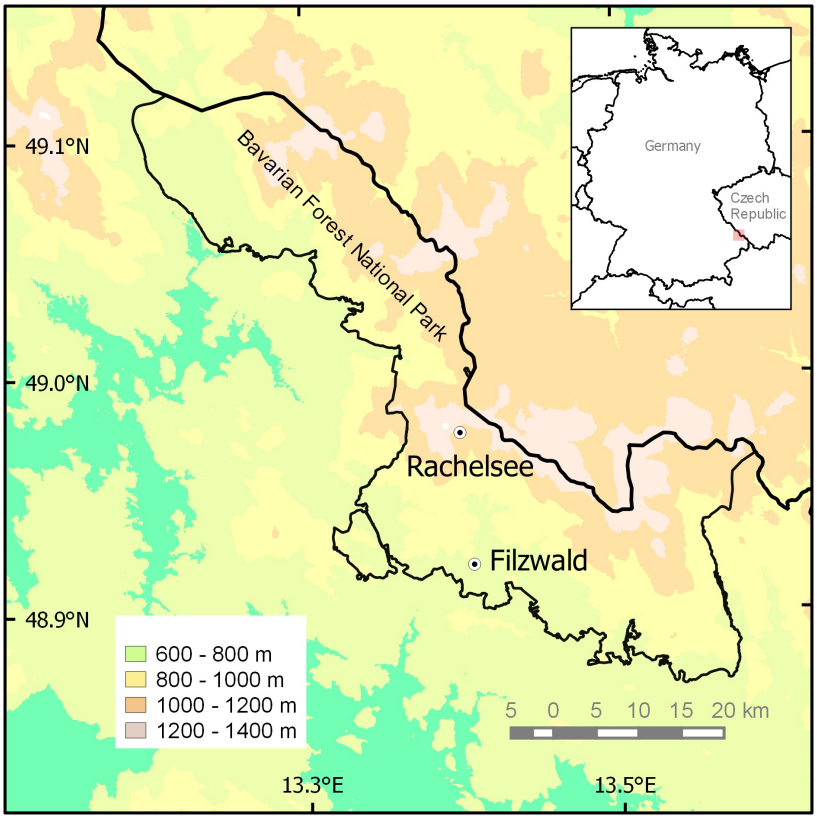

Figure 1. Study area in the Bavarian Forest National Park with the sampling sites Rachelsee (1,120 m a.s.l.) and Filzwald (770 m a.s.l.)

Genotyping

For genotyping, we used 267 polymorphic and functionally annotated SNPs (Roschanski et al., 2015). Out of these, 241 could be successfully genotyped in our samples using KASP assays (LGC Genomics, Middlesex, United Kingdom). The dataset was first roughly filtered by removing SNPs with more than 80% missing data. In a second step, individuals and SNPs with more than 10% missing data, as well as monomorphic SNPs (i.e. SNPs with only a single genotype) were removed. We selected only SNPs with a minor allele frequency $> 3\%$. All SNPs were tested for pairwise linkage disequilibrium (LD) using the Genome Variation Server 147 v. 12.00 (National Heart, Lung, and Blood Institute, http://gvs.gs.washington.edu/GVS147/index.jsp) and for SNP pairs with a coefficient of determination $r^2=1$ (i.e. SNPs that carried identical information), one of the SNPs, which were always located within the same contig, was removed. With this set of SNPs, we imputed missing genotypes using Beagle 4.1 (Browning and Browning, 2016) without using a reference sequence. After imputation, the dataset was cleaned again using the same filtering steps as described above. Finally, all SNPs with less than five individuals per SNP genotype were removed from the dataset to ensure sufficient replication, resulting in 130 SNPs within 103 genes from 193 individuals.

Phenotypes from wood cores

To obtain data on tree-ring width (TRW), we extracted two wood cores per tree at breast height with an increment borer. If trees grew on slopes, they were cored at a 90° angle to the slope to avoid compression wood. After drying, intact or slightly fractured wood cores were cut with a microtome (WSL, Birmensdorf, Switzerland) to obtain a smooth surface. The contrast between earlywood and latewood was enhanced with chalk. Cores that had several fractures (21 out of 275 cores) were mounted on wooden holders and smoothed with grid paper. TRW was measured with a precision of 0.01 mm using a LINTAB digital positioning table whose movements were transmitted to the TSAP-Win Scientific Software (Rinntech, Heidelberg). We constructed a master series for each plot using COFECHA (Grissino-Mayer, 2001) and each series was cross-dated against this master series to avoid mis-dating due to missing rings. We obtained reliable data from a total of 275 cores from 193 trees.

All TRW data was then detrended (standardized) to a dimensionless tree-ring index (TRI) with a mean value of one (Fritts, 2001) using the *detrend* function with `method="Mean"` in the **dplR** package (Bunn, 2008) in the statistical software R (R Core Team, 2016). Detrending is necessary since TRW is not only a function of past climate but also of 'noisy' biological effects, e.g. decreasing growth with increasing age, that have to be removed from the data prior to the analyses of climatic effects on growth (Cook and Peters, 1981).

The previously described growth decline became visible in our tree-ring chronology as early as the 1880s and peaked around the late 1970s and early 1980s (Fig. S1). The strong decline in TRI from the year 1973 to 1974 was clearly visible in the large majority of the trees, and also in the stand chronology (Fig. 2 and 3) and coincides with the peak in SO_2 levels (compare Elling et al. (2009)). Throughout this text, we will refer to this period as 'depression period' and based on the existing data assume that the main driver of the growth decline was air pollution, in particular SO_2. On average, the depression period lasted until the mid-1980s, after which many trees showed a strong increase in TRI.

To characterize this depression period for each single tree and to obtain numerical measures (dendrophenotypes) that characterize the reaction of individual trees, we followed the definitions of Lloret et al. (2011) to determine the resistance, recovery and resilience of the trees to air pollution (Fig. 2). Within this framework, resistance describes the relation of TRI during vs. before the extreme event, recovery describes the ratio of TRI after vs. during the event, and resilience describes the ratio of TRI after vs. before the event.

We then calculated the following dendrophenotypes: (1) the steepness of the start of the depression period in 1974 as the slope of the TRI between the years 1973 and 1974, (2) the resistance towards air pollution as the ratio between the average of ten TRIs from 1964 to 1973 and the average of the TRI during the depression period (1974-1983), (3) the recovery after the depression period as the ratio between the average TRI in the ten years after 1983 and the average TRI during the depression period (1974-1983), and (4) the end of the depression period (i.e. beginning of the resilience phase) as the year when TRI surpassed the values prior to the growth depression. To calculate the latter, we compared mean TRI after 1973 in a moving window of three years to the mean growth in the years 1964-1973 as a reference period. We defined the end of the depression as the year when the three-year mean first surpassed growth of the reference period.

In addition, we focused on the year 1976, for which most trees showed a further decrease in growth during the depression period (Fig. 2 and 3) and which has been identified as one of the driest years for south-east England (Wigley and Atkinson, 1977) and Europe in general (Briffa et al., 2009). Although we could not define this year as a so-called pointer year (Cropper, 1979), likely because growth was already dramatically dropping during prior years, most trees showed a marked growth decline in 1976. We used the *res.comp* function from the R package **pointRes** (van der Maaten-Theunissen et al., 2015) to calculate (5) the recovery, (6) resilience and (7) resistance for each tree towards the conditions in 1976. For the calculation, we considered a two-year window to take into account the reduced growth in the two prior years that already fall within the period of growth depression (Lloret et al., 2011).

For all dendrophenotypes the mean values and corresponding standard deviations (SD) for each site were calculated. Significant differences between the means from the high and low site for each dendrophenotype were tested using Welch's unequal variances *t*-tests. For the genetic association, all dendrophenotypes were centered and scaled within the two sites to exclude confounding effects due to environmental and site conditions using the *scale* function with default parameters in the R package **base** (R Core Team, 2016).

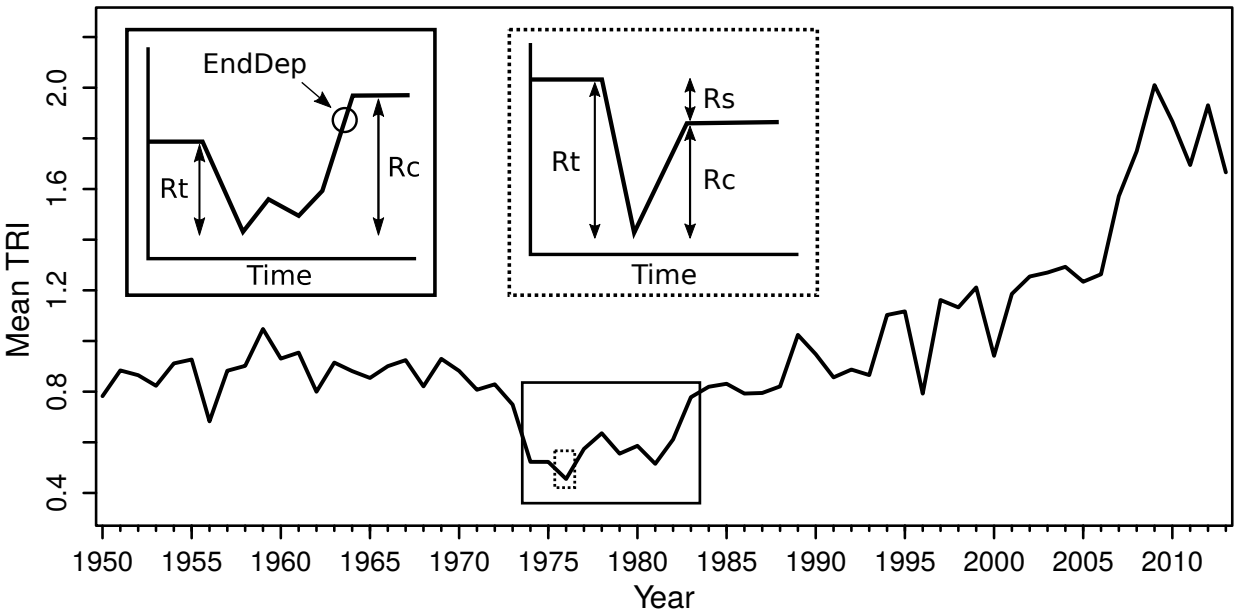

Figure 2. Mean tree-ring index (TRI) of all individuals across both sites for each year in the period from 1950 to 2013. The insets (modified after Lloret et al. (2011)) graphically represent the dendrophenotypes resistance (Rt), recovery (Rc), resilience (Rs) and end of depression period (EndDep) for the depression period (solid box) and for the drought year 1976 (dashed box).

Population structure

Population stratification can lead to spurious associations (Lander and Schork, 1994) and the methods we used for genetic association are sensitive to population structure (Zhang et al., 2010; Zhao et al., 2012). Thus, we applied two approaches to determine whether we could detect a genetic structure within or between sampling sites. First, we used the Bayesian clustering algorithm implemented in STRUCTURE 2.3.4 (Pritchard et al., 2000). We used the admixture model with

correlated allele frequencies, set the burn-in to 10^5 iterations, followed by 5 x 10^5 Markov chain Monte Carlo (MCMC) repetitions, and conducted 10 runs for each number of populations K from 1-6. Second, we performed a principal component analysis (PCA) with the SNP data. For this, we first converted the SNP data into a binary format (0, 1 and 2, corresponding to the occurrence of the respective minor allele of a SNP) using a custom R-script and then used the *glPCA* function of the R package **adegenet** (Jombart and Ahmed, 2011).

Genetic association

For the association analysis of SNPs and dendrophenotypes, we used two approaches. First, we applied a frequently used univariate approach, namely general linear models (GLMs) as implemented in TASSEL v. 5.0 (Bradbury et al., 2007), with each SNP as the independent variable and each dendrophenotype as the response variable. For each dendrophenotype we ran GLMs with 10,000 permutations to obtain *p*-values independent of the data distribution and Bonferroni-corrected the permutation *p*-values by only considering SNPs as significant with a permutation *p*-value $< 0.05/7 = 0.007$ (permutation accounts only for multiple testing within each dendrophenotype and since we tested seven dendrophenotypes, we divided the significance-threshold by this number of tests).

Apart from this single locus approach, we also applied the machine learning algorithm random forest, which captures both marginal and interaction effects among SNPs. Random forest is a nonparametric decision tree algorithm that assigns an importance value to each variable in a dataset based on a vote over all the trees in the forest (Breiman, 2001). Each tree is grown using a different bootstrap sampled subset of the original data. This process, called 'bagging', results in a fraction of the data that is not sampled and thus called 'out-of-bag' (OOB) sample. Using the bagged samples to predict the OOB samples gives a measure of prediction accuracy, the OOB error. The average increase over all trees in the OOB error of a variable, compared to its randomly permuted counterpart, provides a measure of variable importance for this predictor (i.e. SNP). This variable importance gives a measure for the comparison between SNPs within a given set but does not provide a threshold to select truly relevant SNPs. For this purpose we used the feature selection procedures as implemented in the R packages **Boruta** (Kursa and Rudnicki, 2010) and **VSURF** (Genuer et al., 2015).

Boruta is an all-relevant feature selection algorithm and attempts to identify all significantly important predictors (in our case SNPs). As a wrapper algorithm for the R package **randomForest** (Liaw and Wiener, 2002), Boruta creates shadow attributes for each predictor by randomly shuffling the original values. A random forest model is applied to the extended dataset and the importance of each predictor is assessed in the form of Z scores. These are calculated based on the mean accuracy loss upon permutation of the predictor over all trees in the forest, divided by the corresponding standard deviation. The maximum Z score among the shadow attributes is then compared with the Z score for each predictor. Predictors with significantly ($\alpha = 0.05$) higher Z scores are deemed 'important' while such with significantly lower Z scores are deemed 'unimportant'. This process is repeated until all predictors are evaluated or the user-defined limit of runs is reached, at which point all remaining predictors are deemed 'tentative'. To account for possibly high fluctuations in the Z scores when there are many predictors, Boruta first runs three rounds

with more relaxed criteria in which only rejection but no confirmation of predictors takes place. Boruta was executed with 2000 trees in each forest and a maximum of 5000 runs to ensure that all SNPs were classified as either important or unimportant.

VSURF can conduct both an all-relevant, as well as a minimal-optimal feature selection (Genuer et al., 2015). While the former has the same objective as Boruta - namely the identification of all relevant predictors, including redundancies, that are associated with a variable (dendrophenotype) - the latter is aimed at selecting the smallest set of predictors that explains most of the variation in a given variable. Since we were interested in identifying all relevant SNPs for biological interpretation, as opposed to the most accurate prediction of a dendrophenotype based on some set of SNPs, we only employed the first approach. For this, VSURF removes unimportant predictors from the dataset based on the standard deviations of their variable importance averaged over 50 random forests. The remaining predictors are used for variable selection to identify all relevant predictors. OOB error rates from nested random forests are calculated starting with only the most important predictor and stepwise including the next best predictor until all predictors that were selected in the previous step are included. The predictors from the model with the smallest OOB error are then chosen as the set for interpretation. VSURF was implemented with default settings which includes 2000 trees in each forest.

Although SNPs were already annotated for a previous publication (Roschanski et al., 2015), we repeated this step, as the information in the NCBI database in constantly growing. All SNPs that could be associated with the dendrophenotyopes with more than one of the above mentioned methods were compared to known sequences from NCBI's GenBank non-redundant protein database (NR) using the translated BLAST algorithm (blastx v. 2.6.1+, Altschul et al., 1997). The best hit, based on the Expect value that provided a functional annotation was selected for each gene and the corresponding biological process keywords were retrieved from the Gene Ontology (GO) database (UniProtKB; The UniProt Consortium, 2015).

Results

Dendrophenotypes

We obtained data on dendrophenotypes for 98 and 95 individuals from the high and low site, respectively. Overall, trees from high elevations showed a more pronounced and uniform reaction during the depression period and to the conditions in 1976 than trees from the low site (Fig. 3). At high elevations, 96.8% of the trees showed a decline in growth as expressed in the resistance to air pollution, while only 79.6% of the trees at the low site showed this reaction.

On average, the dendrophenotypes differed between sites (Fig. 4 and Table 1). While the individuals on the high site showed a steeper drop into the depression period and had, on average, a lower resistance to air pollution, the recovery from the pollution was stronger when compared to the low site. All mentioned differences were significant (Table 1). Regarding the drought year 1976, the individuals from the low site consistently showed higher mean values, which were significantly different from the high site for resistance and resilience but not for recovery. The end of depression only differed by one year and was not significantly different between sites.

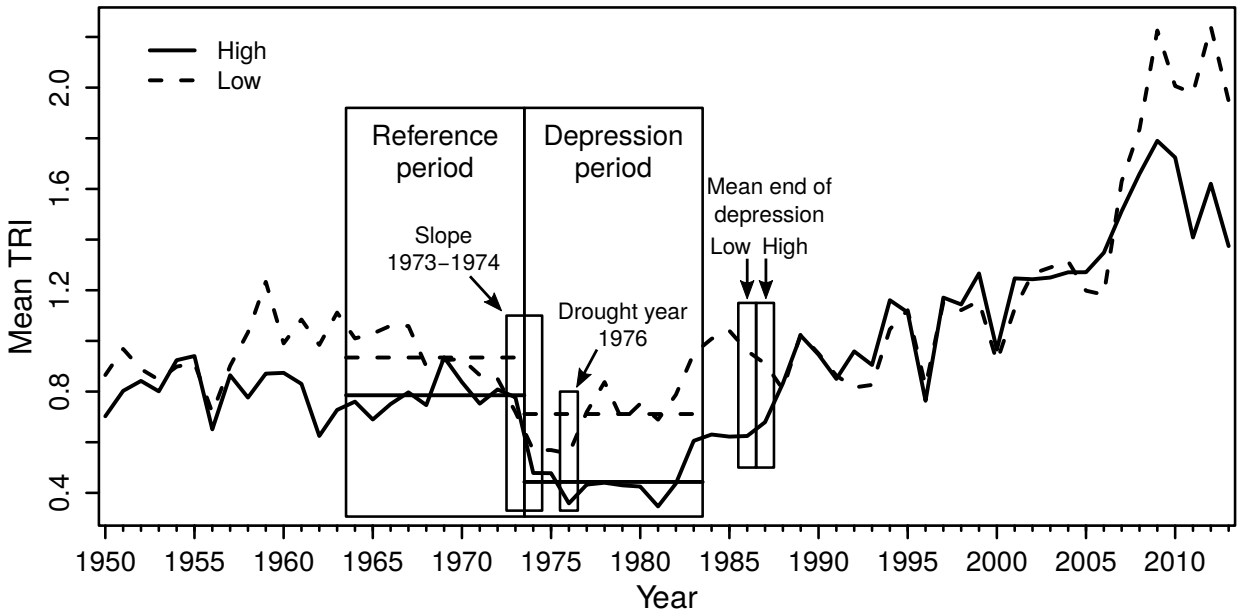

Figure 3. Mean tree-ring index (TRI) of all individuals across both sites (High and Low) for each year in the period from 1950 to 2013. The reference and depression period are the basis for the calculation of the resistance to air pollution. The horizontal lines within the boxes represent the mean TRI for the corresponding period and site. Smaller boxes from left to right mark the onset of the depression period for which we calculated the slope, the drought year 1976, and the mean end of the growth depression for low and high elevations, respectively.

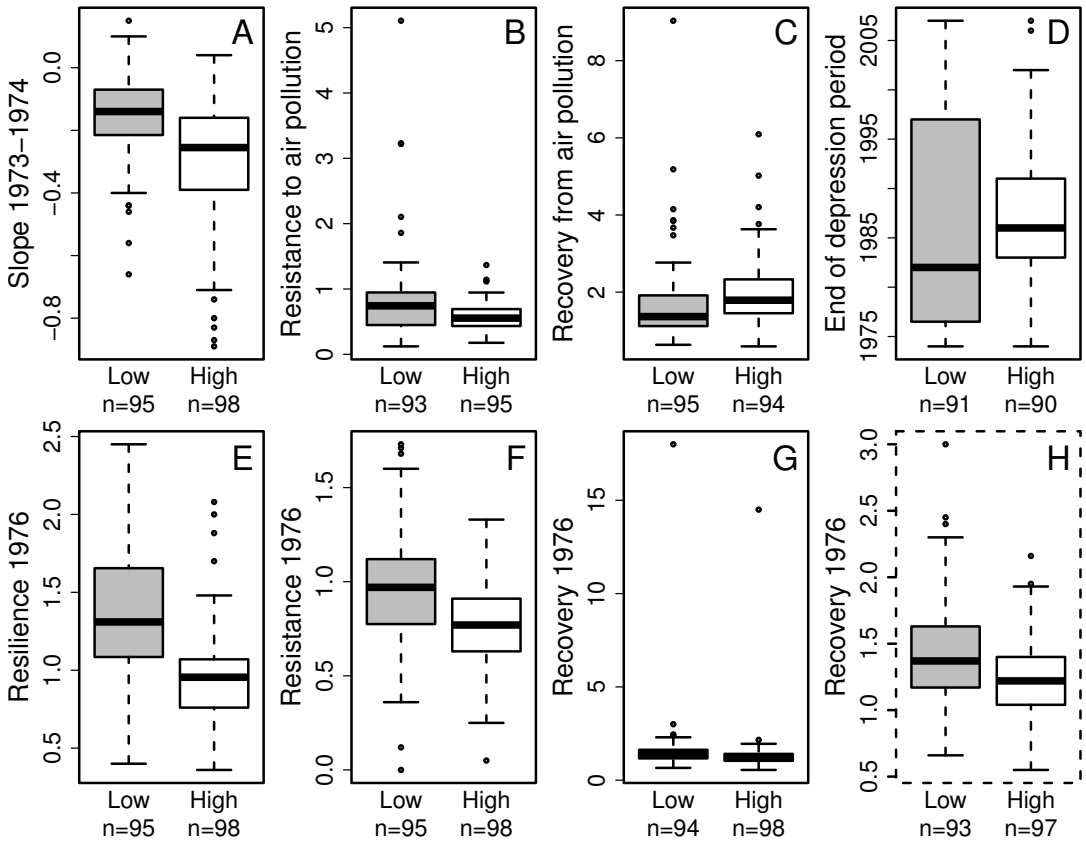

Figure 4. Dendrophenotypes for trees from both sites (Low and High). Phenotypes that describe the start (A) and end of the depression period (D), and the trees' resistance (B) and recovery (C) are depicted. Resilience (E) and resistance (F) to and recovery (G) from the drought year 1976 are shown. For a better visualization of the data distributions, the extreme values for Recovery 1976 were removed (H).

Table 1. Summary statistics for the dendrophenotypes from both sites (Low and High) and results of the Welch's *t*-tests between the two sites for each dendrophenotype.

	Slope 1973-74	Resistance to air pollution	Recovery from air pollution	End of depression period	Resilience 1976	Resistance 1976	Recovery 1976
Low (mean ± SD)	-0.15 ± 0.13	0.82 ± 0.67	1.66 ± 1.11	1986.49 ± 11.21	1.37 ± 0.41	0.97 ± 0.30	1.59 ± 1.76
High (mean ± SD)	-0.30 ± 0.20	0.58 ± 0.21	1.96 ± 0.89	1987.48 ± 6.57	0.95 ± 0.30	0.76 ± 0.22	1.41 ± 1.37
t	-5.98	-3.33	2.01	0.72	-7.92	-5.44	-0.81
df	169	109	179	146	171	173	176
p-value	<0.001	0.0012	0.04616	0.4721	<0.001	<0.001	0.4813

SD: Standard deviation, *t*: *t*-statistic, df: degrees of freedom

Population structure

Neither the STRUCTURE analysis, nor the PCA provided any indication for population substructure. The visual inspection of the bar plots in STRUCTURE clearly showed that almost all individuals were assigned to both clusters in a scenario with K = 2 without any apparent pattern (Fig. S2), and Ln P(D) declined steadily with increasing K (Fig. S3). In accordance, point clouds resulting from the PCA showed no apparent difference between sampling sites (Fig. S4).

Genetic association

The different association methods resulted in largely different numbers of SNPs associated to the dendrophenotypes (Table 2). VSURF identified 10 to 22 SNPs for every dendrophenotype with the exception of Recovery 1976, for which VSURF only found one SNP. Boruta detected a lower number of SNPs, ranging from zero to six, which were also detected by VSURF in most cases. The GLM in TASSEL yielded no significant results (permutation p-value < 0.007). In total, 15 SNPs were jointly identified by at least two approaches. Most of these SNPs are located in genes that code for membrane proteins related to transport and stress reactions (Table 3).

Table 2. Overview of the results of the different association methods (TASSEL GLM, VSURF and Boruta) for each dendrophenotype. Values in cells indicate the number of SNPs identified by each method for a given dendrophenotype.

Method	Slope 1973-74	Resistance to air pollution	Recovery from air pollution	End of depression period	Resilience 1976	Resistance 1976	Recovery 1976
GLM	0	0	0	0	0	0	0
VSURF	22	10	18	11	17	13	1
Boruta	1	4	6	1	3	6	0
VSURF + Boruta	1	3	5	0	3	6	0

Table 3. SNPs associated with scaled dendrophenotypes using three different association methods (TASSEL GLM, Boruta, VSURF). Only SNPs that were associated with a given dendrophenotype by more than one method are shown. A full table with all detected associations is provided in the supplementary material (Table S2).

SNP	Protein	Biological Process (starting with capital letter: GO keyword)	Syn-state	AAC	Slope 1973-74	Resistance to air pollution	Recovery from air pollution	Resilience 1976	Resis-tance 1976
628.35	mitochondrial arginine transporter BAC2	Stress response, Transport	?	-	-	Boruta, VSURF	-	-	-
716.144	heat shock 70 kDa protein 7, chloroplastic	Protein transport, Stress response, Transport	?	-	-	Boruta, VSURF	-	-	-
2190.265	ATP-dependent Clp protease proteolytic subunit 4, chloroplastic-like	chloroplast organization, regulation of timing of transition from vegetative to reproductive phase	?	-	Boruta, VSURF	-	-	-	-
4538.47	60S ribosomal protein L7-2-like	cytoplasmic translation, maturation of LSU-rRNA from tricistroic rRNA transcript	?	-	-	-	Boruta, VSURF	-	-
8092.366	T-complex protein 1 subunit epsilon	protein folding	syn	-	-	-	Boruta, VSURF	-	-
8855.137	Lhca4 protein, Type 4 protein of light-harvesting complex of photosystem I	Photosynthesis	?	-	-	-	Boruta, VSURF	-	-
9197.63	aquaporin TIP2-1	Transport	syn	-	-	-	-	-	Boruta, VSURF
10568.484	GDP-mannose pyrophosphorylase	cellulose biosynthetic process, defense response to bacterium, GDP-mannose biosynthetic process, L-ascorbic acid biosynthetic process, response to ammonium ion, response to heat, response to jasmonic acid, response to ozone, response to salt stress	syn	-	-	-	Boruta, VSURF	-	-
12178.301	glucan endo-1,3-beta-glucosidase	Plant defense	syn	-	-	-	-	Boruta, VSURF	-
14580.627	ATP-dependent Clp protease proteolytic subunit-related protein 4, chloroplastic	regulation of gene expression, response to reactive oxygen species	syn	-	-	-	Boruta, VSURF	-	-
15256.604	Ferredoxin-NADP reductase, leaf-type isozyme, chloroplastic, partial	Electron transport, Photosynthesis, Transport	?	-	-	Boruta, VSURF	-	Boruta, VSURF	Boruta, VSURF
16332.419	proteasome subunit beta type-6	proteolysis involved in cellular protein catabolic process	non-syn	D→E	-	-	-	Boruta, VSURF	-
16411.197	mitochondrial carnitine/acylcarnitine carrier-like protein	Transport	?	-	-	-	-	-	Boruta, VSURF
16430.504	NADH dehydrogenase (ubiquinone) iron-sulfur protein 7, mitochondrial	Electron transport, Respiratory chain, Transport	?	-	-	-	Boruta, VSURF	-	Boruta, VSURF
24318.117	LIM domain-containing protein WLIM2b isoform X1	-	?	-	-	-	-	-	Boruta, VSURF

Syn synonymous, **AAC** Amino Acid Conversion, **?** no information available, **-** no result, **D** Aspartic acid, **E** Glutamic acid

Discussion

We present the first association study that links the reaction of individual trees to environmental stresses that are archived in tree-rings with SNPs in candidate genes, putatively related to such stresses. Specifically, we characterized the strong growth decline of silver fir during the 1970s and early 1980s, which can be related to the peak in SO_2 emissions and a series of dry years. Starting in the mid-1980s, silver fir growth increased and surpassed pre-depression growth levels, which has been related to the decrease in air pollution, competition release caused by forest dieback, as well as an increase in air temperatures and atmospheric fertilization (Büntgen et al., 2014). More importantly, besides the population level signal, we derived measures for resilience and resistance of individual trees towards these stresses. When we tested for associations between these individual dendrophenotypes and the SNP data from 103 candidate genes, we could identify 15 genes related to drought and general stress pathways, suggesting a genetic background for individual differences in the response to the conditions during the depression period. While the historic levels of air pollution probably remain a unique stress episode, there are indications that growth decline was enhanced by an increased drought sensitivity due to a physiological response to SO_2 (Elling et al., 2009). Thus, the observed individual differences in resilience potentially play a role in adaptation to future drought events. The major advancement of our study thereby is the use of dendrophenotypes to derive information on individual tree reaction to episodic and complex stresses. Furthermore, dendrophenotypes have the major advantage that they can be collected for large numbers of trees which are required for genotype-phenotype association studies in natural populations. This discussion deals with the ecological, genomic and methodological implications.

Dendrophenotypes

As expected, we found pronounced population level growth declines in the 1970s and 1980s, and population level recovery thereafter as described earlier for silver fir in the whole of Southern Germany (Elling et al., 2009). While our data is limited to surviving trees, inventory data from the area showed that silver fir dieback was substantial in the 1970s and 1980s. For example, some forest stands lost more than 75% of their 80–120 year old silver firs (unpublished inventory data, draft of the National Park Plan 1992).

Further, we found that the drought year of 1976 had a strong negative effect on the growth of the silver fir stands. In contrast, dry years that did not coincide with the depression period (e.g. 1959, 1972, 1982 and 2003) did not have a strong effect on silver fir growth (Fig. 3). This is in line with Elling et al. (2009), who argued that SO_2 pollution not only causes direct harm to silver fir trees by impeding photosynthesis and leading to the shedding of needles but that it also increases sensitivity to drought, which might be attributable to damages of the fine-root system.

In general, we observed that trees at high elevations were affected more severely, as reflected by the generally lower resistance and resilience. However, since there are no on-site measurements of SO_2 concentrations, we can only speculate that the high site might have been more severely affected by SO_2 emissions, or alternatively, that the effect was aggravated by site-specific conditions around lake Rachel.

As was shown before, the growth after the depression period exceeds prior levels, which is usu-

ally attributed to a combination of less dense forest structure and elevated nitrogen supply (Pinto et al., 2007; Elling et al., 2009; Büntgen et al., 2014). It has also been speculated that tropospheric ozone (O_3) might be a major contributor to forest decline (Krause et al., 1986; Schmieden and Wild, 1995). However, since tropospheric O_3 concentrations increased well into the 1980s, our data does not directly indicate a major influence on growth in silver fir stands in Southern Germany.

Dendrophenotype-genotype association

In total, 15 out of 130 SNPs were associated by two methods with one of the dendrophenotypes, and are thus supported in their status as 'candidate' genes for the reaction towards extreme environmental stress. The investigated candidate genes were selected as they either had drought-related Gene Ontology terms, or were previously detected to be associated with adaptive processes and/or stress response (Roschanski et al., 2013, 2015). Interestingly, the vast majority of genes we found in association with dendrophenotypes were membrane proteins of the chloroplast, mitochondria or tonoplast, and thus, tightly linked to photosynthesis or chloroplast development. For example, SNPs in contigs 716 and 14580, which are associated to the resistance and recovery during the depression period, respectively, encode for a heat shock 70 kDa protein and a proteolytic subunit of the ATP-dependent Clp protease which are both involved in protein folding with effects on chloroplast development and function (Sjögren et al., 2006; Latijnhouwers et al., 2010). Since SO_2 pollution likely has negative effects on photosynthesis (Silvius et al., 1975), genes involved in these pathways could potentially determine how individual trees cope with these extreme conditions.

Two of the genes that were exclusively associated with resistance and resilience in the drought year 1976 can be directly related to drought stress response: aquaporin TIP2-1 and glucan-endo-1,3-beta-glucosidase. Aquaporins are regularly involved in drought response (Maurel et al., 2008; Hamanishi and Campbell, 2011) and a similar aquaporin (TIP1-1) has already been identified as differentially expressed in response to drought stress in silver fir seedlings (Behringer et al., 2015). Glucan-endo-1,3-beta-glucosidase was previously used as a drought stress candidate gene in *Pinus pinaster* (Eveno et al., 2008) and was also differentially expressed in response to drought stress in silver fir seedlings (Behringer et al., 2015).

Although, in general, biological functions have been described for *Arabidopsis thaliana* or other flowering plants, there is evidence that many genes maintained their functions across the plant kingdom (Groover, 2005). Still, the physiological function of these genes, and even more specifically of the investigated SNPs, remains to be determined. In many cases, we found associations with SNPs that are synonymous. Nevertheless, these mutations might impact gene expression, and it has been shown that there is a codon bias in conifers which affects translational efficiency (Torre et al., 2015).

For the genetic association analysis, we not solely relied on commonly used single locus approaches, but also applied a random forest based feature selection to identify SNPs that are likely associated with certain dendrophenotypes. The Boruta algorithm provides significant results by testing if the importance of a SNP for explaining a dendrophenotype is significantly ($\alpha = 5\%$) higher than the importance of the most important shadow attribute, which, under the null hypothesis is only associated by chance (Kursa and Rudnicki, 2010). In contrast, VSURF does not

incorporate any formal statistical hypothesis-test, but selects the most important SNPs regarding the association with a specific dendrophenotype (Genuer et al., 2015). However, this does not imply a statistically significant association.

Both feature selection techniques are wrappers for the random forest algorithm and, as such, the importance value is a measure for the marginal effect of a SNP, as well as the interaction effect of all SNPs under consideration. It should be mentioned, however, that the relative contribution of marginal and interaction effect cannot be directly determined (Boulesteix et al., 2012). Thus, SNPs identified by random forest procedures do not represent a network and have to be viewed independently. The benefit of such analyses is, however, that the influence of all other SNPs are incorporated in the importance of any given SNP, which provides a much better representation, given that in association studies in conifers a single SNP did not explain more than 5% in the variation of a trait (e.g. Eckert et al., 2009; González-Martínez et al., 2006, 2008).

Outlook: The future of dendrophenotypes in association studies

Episodic environmental extremes like droughts, storms, or calamities are predicted to increase in both intensity and frequency due to global climate change (IPCC, 2014). While a better understanding of the response of trees to such extreme events is urgently needed, this is challenged by the unforeseeable timing of these events, which makes it almost impossible to integrate them in research projects with short funding periods. Further, forest geneticists are confronted with numerous challenges that hamper meaningful association studies. While technical advances provided us with a roadmap on how to move from forest genetics to forest genomics (Neale and Kremer, 2011), next generation phenotyping for trees is still in its infancy. The ideal phenotype should be meaningful in terms of fitness and easy to measure in a large number of trees in natural populations. Here, the analysis of wood cores has a number of clear advantages. First, wood cores can be collected with an acceptable investment of time and money in the field, (a well-trained person might collect up to 30-40 cores per day). All subsequent steps can be conducted in the lab. Second, wood cores provide data over long time series, and thus permit to characterize tree reaction to climate, as well as to focus on particularly extreme events. Thereby, we can exploit the internal archive of trees and study extreme episodes in their past which are not accessible directly using other measures and cannot easily be integrated within short-term research grants (typically 3 to 5 year-long). Third, the analysis is by far not limited to the mere measurement of tree-ring width but can be expanded in various directions such as wood anatomical features, for example cell wall thickness or lumen area, which are considered a proxy for physiological adaptations to external factors (Carrer et al., 2010; Ziaco et al., 2016), as well as isotope measures, which, among other things, provide evidence for the water use efficiency of a tree (Seibt et al., 2008). Microdensitometry can supplement ring width data with information of wood density and thereby provide a more complete picture of growth, for example during extreme events (e.g. Martinez-Meier et al., 2008). All these measures can be conducted for time series of several years, or with a focus on particular years of interest.

A major challenge for wood core analysis is that the individual growth reactions integrate over numerous signals such as age, microclimatic conditions, competition and health, in addition to the reaction to climatic conditions. Thus, a characterization of micro-environmental conditions is essential which remains notoriously difficult to determine at a scale which is relevant for individual

trees in natural populations. This is also highly relevant for the study of adaptational processes, which might occur at smaller spatial scales than previously considered (Scotti et al., 2015). A few studies already provided evidence that local adaptation can occur at very small spatial scales, contradicting the previous notion that this is impeded by gene flow in trees (e.g. Budde et al., 2014; Eckert et al., 2015). Thus, with the goal of understanding the underlying mechanisms of adaptational processes in forest trees, we should not solely rely on the major advances made in sequencing technologies, but also broaden our focus to promising phenotyping approaches such as provided by dendrochronology and dendroecology in combination with a more detailed view of environmental conditions.

Contributions to Manuscript or Data

Design BF, BZ, GV, KH, LO, SL, UB

Data collection CB, HJ, KH, LO, MW, SVL

Data analysis AP, DB, KH, LO

Data interpretation AP, DB, KH, LO

Writing the manuscript DB, KH, LO

Acknowledgements

We thank the Bavarian Forest National Park for supporting the field work. KH was funded by the ERAnet BiodivERsA project 'TipTree' (ANR-12-EBID-0003 granted to BZ, LO, BF, GV, SL) funded by the German Federal Ministry of Education and Research (Grant 01LC1202A to). MW and SL were supported by ERASMUS scholarships during their stay at WSL Birmensdorf. Thanks to Lena Hellmann, Diego Galván Candela and Ricardo Ochoa Pereira for their help during field work, and to Anne Verstege for her support with tree-ring measurements.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res* **25**(17), 3389–3402.

Babushkina, E. A., Vaganov, E. A., Grachev, A. M., Oreshkova, N. V., Belokopytova, L. V., Kostyakova, T. V. and Krutovsky, K. V. (2016), 'The effect of individual genetic heterozygosity on general homeostasis, heterosis and resilience in Siberian larch (Larix sibirica Ledeb.) using dendrochronology and microsatellite loci genotyping', *Dendrochronologia* **38**, 26–37.

Behringer, D., Zimmermann, H., Ziegenhagen, B. and Liepelt, S. (2015), 'Differential Gene Expression Reveals Candidate Genes for Drought Stress Response in Abies alba (Pinaceae)', *PLoS ONE* **10**(4), e0124564.

Bosela, M., Popa, I., Gömöry, D., Longauer, R., Tobin, B., Kyncl, J., Kyncl, T., Nechita, C., Petráš, R., Sidor, C. G., Šebeň, V. and Büntgen, U. (2016), 'Effects of post-glacial phylogeny and genetic diversity on the growth variability and climate sensitivity of European silver fir', *J Ecol* **104**(3), 716–724.

Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012), 'Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics', *WIREs Data Mining Knowl Discov* **2**(6), 493–507.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y. and Buckler, E. S. (2007), 'TASSEL: software for association mapping of complex traits in diverse samples', *Bioinformatics* **23**(19), 2633–2635.

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**(1), 5–32.

Briffa, K. R., van der Schrier, G. and Jones, P. D. (2009), 'Wet and dry summers in Europe since 1750: evidence of increasing drought', *International Journal of Climatology* **29**(13), 1894–1905.

Browning, B. L. and Browning, S. R. (2016), 'Genotype Imputation with Millions of Reference Samples', *The American Journal of Human Genetics* **98**(1), 116–126.

Budde, K. B., Heuertz, M., Hernández-Serrano, A., Pausas, J. G., Vendramin, G. G., Verdú, M. and González-Martínez, S. C. (2014), 'In situ genetic association for serotiny, a fire-related trait, in Mediterranean maritime pine (Pinus pinaster)', *New Phytol* **201**(1), 230–241.

Bunn, A. G. (2008), 'A dendrochronology program library in R (dplR)', *Dendrochronologia* **26**(2), 115–124.

Büntgen, U., Tegel, W., Kaplan, J. O., Schaub, M., Hagedorn, F., Bürgi, M., Brázdil, R., Helle, G., Carrer, M., Heussner, K.-U., Hofmann, J., Kontic, R., Kyncl, T., Kyncl, J., Camarero, J. J., Tinner, W., Esper, J. and Liebhold, A. (2014), 'Placing unprecedented recent fir growth in a European-wide and Holocene-long context', *Frontiers in Ecology and the Environment* **12**(2), 100–106.

Carrer, M. (2011), 'Individualistic and Time-Varying Tree-Ring Growth to Climate Sensitivity', *PLOS ONE* **6**(7), e22813.

Carrer, M., Nola, P., Motta, R. and Urbinati, C. (2010), 'Contrasting tree-ring growth to climate responses of Abies alba toward the southern limit of its distribution area', *Oikos* **119**(9), 1515–1525.

Cook, E. R. (1985), 'A time series analysis approach to tree ring standardization'.

Cook, E. R. and Kairiukstis, L. A. (2013), *Methods of Dendrochronology: Applications in the Environmental Sciences*, Springer Science & Business Media. Google-Books-ID: C7TnCAAAQBAJ.

Cook, E. R. and Peters, K. (1981), 'The smoothing spline: a new approach to standardizing forest interior tree-ring width series for dendroclimatic studies', *Tree-ring bulletin* .

Cropper, J. P. (1979), 'Tree-ring skeleton plotting by computer', *Tree-Ring Bulletin* .

Eckert, A. J., Bower, A. D., Wegrzyn, J. L., Pande, B., Jermstad, K. D., Krutovsky, K. V., St. Clair, J. B. and Neale, D. B. (2009), 'Association Genetics of Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-Hardiness Related Traits', *Genetics* **182**(4), 1289–1302.

Eckert, A. J., Maloney, P. E., Vogler, D. R., Jensen, C. E., Mix, A. D. and Neale, D. B. (2015), 'Local adaptation at fine spatial scales: an example from sugar pine (Pinus lambertiana, Pinaceae)', *Tree Genetics & Genomes* **11**(3), 1–17.

Elling, W., Dittmar, C., Pfaffelmoser, K. and Rötzer, T. (2009), 'Dendroecological assessment of the complex causes of decline and recovery of the growth of silver fir (Abies alba Mill.) in Southern Germany', *Forest Ecology and Management* **257**(4), 1175–1187.

Eveno, E., Collada, C., Guevara, M. A., Léger, V., Soto, A., Díaz, L., Léger, P., González-Martínez, S. C., Cervera, M. T., Plomion, C. and Garnier-Géré, P. H. (2008), '"Contrasting Patterns of Selection at Pinus pinaster Ait. Drought Stress Candidate Genes as Revealed by Genetic Differentiation Analyses"', *Mol Biol Evol* **25**(2), 417–437.

Fritts, H. C. (2001), *Tree Rings and Climate*, The Blackburn Press, Caldwell, N.J.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C. (2015), 'VSURF: An R Package for Variable Selection Using Random Forests', *The R Journal* **7**(2), 19–33.

González-Martínez, S. C., Ersoz, E., Brown, G. R., Wheeler, N. C. and Neale, D. B. (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), 1915–1926.

González-Martínez, S. C., Huber, D., Ersoz, E., Davis, J. M. and Neale, D. B. (2008), 'Association genetics in Pinus taeda L. II. Carbon isotope discrimination', *Heredity* **101**(1), 19–26.

Grissino-Mayer, H. D. (2001), 'Evaluating Crossdating Accuracy: A Manual and Tutorial for the Computer Program COFECHA', *Tree-Ring Research* .

Groover, A. T. (2005), 'What genes make a tree a tree?', *Trends in Plant Science* **10**(5), 210–214.

Hamanishi, E. T. and Campbell, M. M. (2011), 'Genome-wide responses to drought in forest trees', *Forestry* **84**(3), 273–283.

IPCC (2014), Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp., Technical report.

Jombart, T. and Ahmed, I. (2011), 'adegenet 1.3-1: new tools for the analysis of genome-wide SNP data', *Bioinformatics* **27**(21), 3070–3071.

King, G. M., Gugerli, F., Fonti, P. and Frank, D. C. (2013), 'Tree growth response along an elevational gradient: climate or genetics?', *Oecologia* **173**(4), 1587–1600.

Krause, G. H. M., Arndt, U., Brandt, C. J., Bucher, J., Kenk, G. and Matzner, E. (1986), 'Forest decline in Europe; Development and possible causes', *Water, Air, and Soil Pollution* **31**(3-4), 647–668.

Kursa, M. B. and Rudnicki, W. R. (2010), 'Feature selection with the Boruta package', *Journal of Statistical Software* **36**(11).

Lander, E. S. and Schork, N. J. (1994), 'Genetic dissection of complex traits', *Science* **265**(5181), 2037–2048.

Latijnhouwers, M., Xu, X.-M. and Møller, S. G. (2010), 'Arabidopsis stromal 70-kDa heat shock proteins are essential for chloroplast development', *Planta* **232**(3), 567–578.

Liaw, A. and Wiener, M. (2002), 'Classification and regression by randomForest', *R news* **2**(3), 18–22.

Lloret, F., Keeling, E. G. and Sala, A. (2011), 'Components of tree resilience: effects of successive low-growth episodes in old ponderosa pine forests', *Oikos* **120**(12), 1909–1920.

Martinez-Meier, A., Sanchez, L., Pastorino, M., Gallo, L. and Rozenberg, P. (2008), 'What is hot in tree rings? The wood density of surviving Douglas-firs to the 2003 drought and heat wave', *Forest Ecology and Management* **256**(4), 837–843.

Maurel, C., Verdoucq, L., Luu, D.-T. and Santoni, V. (2008), 'Plant Aquaporins: Membrane Channels with Multiple Integrated Functions', *Annu. Rev. Plant Biol.* **59**(1), 595–624.

McLaughlin, S. B. (1985), 'A Critical Review', *Journal of the Air Pollution Control Association* **35**(5), 512–534.

Neale, D. B. and Kremer, A. (2011), 'Forest tree genomics: growing resources and applications', *Nat Rev Genet* **12**(2), 111–122.

Pinto, P. E., Gégout, J.-C., Hervé, J.-C. and Dhôte, J.-F. (2007), 'Changes in environmental controls on the growth of Abies alba Mill. in the Vosges Mountains, north-eastern France, during the 20th century', *Global Ecology and Biogeography* **16**(4), 472–484.

Pluess, A. R. and Weber, P. (2012), 'Drought-Adaptation Potential in Fagus sylvatica : Linking Moisture Availability with Genetic Diversity and Dendrochronology', *PLOS ONE* **7**(3), e33636.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000), 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics* **155**(2), 945–959.

R Core Team (2016), 'R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna Austria. URL https://www.R-project.org/'.

Roschanski, A. M., Csilléry, K., Liepelt, S., Oddou-Muratorio, S., Ziegenhagen, B., Huard, F., Ullrich, K. K., Postolache, D., Vendramin, G. G. and Fady, B. (2015), 'Evidence of divergent selection for drought and cold tolerance at landscape and local scales in Abies alba Mill. in the French Mediterranean Alps', *Mol Ecol* pp. n/a–n/a.

Roschanski, A. M., Fady, B., Ziegenhagen, B. and Liepelt, S. (2013), 'Annotation and Re-Sequencing of Genes from De Novo Transcriptome Assembly of Abies alba (Pinaceae)', *Applications in Plant Sciences* **1**(1), 1200179.

Schmieden, U. and Wild, A. (1995), 'The contribution of ozone to forest decline', *Physiologia Plantarum* **94**(2), 371–378.

Scotti, I., González-Martínez, S. C., Budde, K. B. and Lalagüe, H. (2015), 'Fifty years of genetic studies: what to make of the large amounts of variation found within populations?', *Annals of Forest Science* pp. 1–7.

Seibt, U., Rajabi, A., Griffiths, H. and Berry, J. A. (2008), 'Carbon isotopes and water use efficiency: sense and sensitivity', *Oecologia* **155**(3), 441.

Silvius, J. E., Ingle, M. and Baer, C. H. (1975), 'Sulfur Dioxide Inhibition of Photosynthesis in Isolated Spinach Chloroplasts', *Plant Physiol.* **56**(3), 434–437.

Sjögren, L. L. E., Stanne, T. M., Zheng, B., Sutinen, S. and Clarke, A. K. (2006), 'Structural and Functional Insights into the Chloroplast ATP-Dependent Clp Protease in Arabidopsis', *Plant Cell* **18**(10), 2635–2649.

The UniProt Consortium (2015), 'UniProt: a hub for protein information', *Nucleic Acids Res* **43**(D1), D204–D212.

Torre, D. L., R, A., Lin, Y.-C., Van de Peer, Y. and Ingvarsson, P. K. (2015), 'Genome-Wide Analysis Reveals Diverged Patterns of Codon Bias, Gene Expression, and Rates of Sequence Evolution in Picea Gene Families', *Genome Biol Evol* **7**(4), 1002–1015.

van der Maaten-Theunissen, M., van der Maaten, E. and Bouriaud, O. (2015), 'pointRes: An R package to analyze pointer years and components of resilience', *Dendrochronologia* **35**, 34–38.

Wigley, T. M. L. and Atkinson, T. C. (1977), 'Dry years in south-east England since 1698', *Nature* **265**(5593), 431–434.

Wilson, R. and Elling, W. (2004), 'Temporal instability in tree-growth/climate response in the Lower Bavarian Forest region: implications for dendroclimatic reconstruction', *Trees* **18**(1), 19–28.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M. and Buckler, E. S. (2010), 'Mixed linear model approach adapted for genome-wide association studies', *Nat Genet* **42**(4), 355–360.

Zhao, Y., Chen, F., Zhai, R., Lin, X., Wang, Z., Su, L. and Christiani, D. C. (2012), 'Correction for population stratification in random forest analysis', *Int. J. Epidemiol.* **41**(6), 1798–1806.

Ziaco, E., Biondi, F. and Heinrich, I. (2016), 'Wood Cellular Dendroclimatology: Testing New Proxies in Great Basin Bristlecone Pine', *Front Plant Sci* **7**.

Supplementary Material

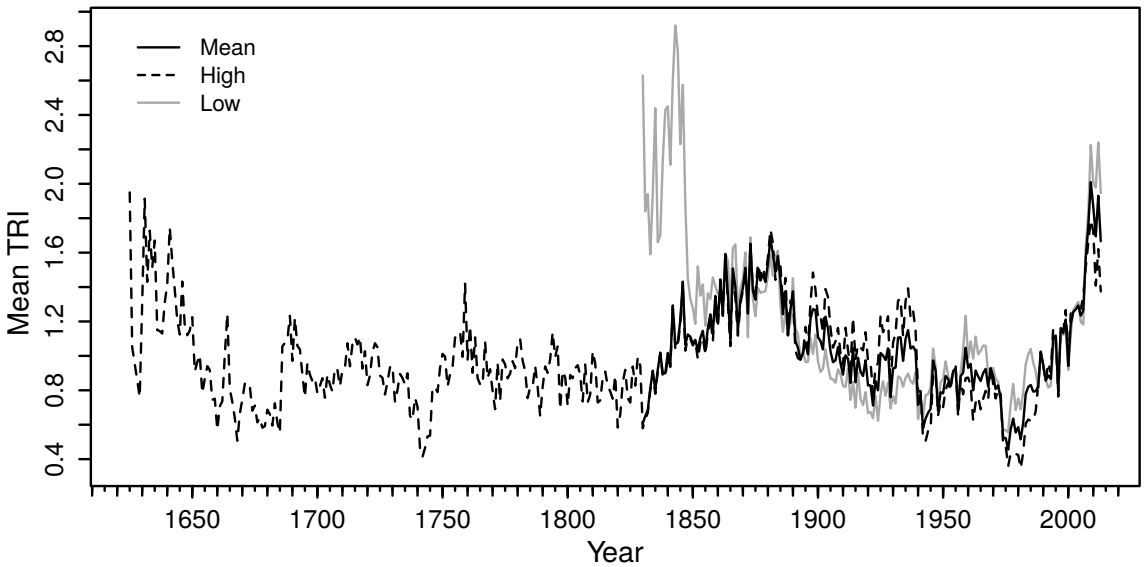

Figure S1. All available mean tree-ring index (TRI) data (detrended tree-ring width data) for the silver fir trees from both sites (High and Low). Also shown is the weighted mean between both sites (Mean).

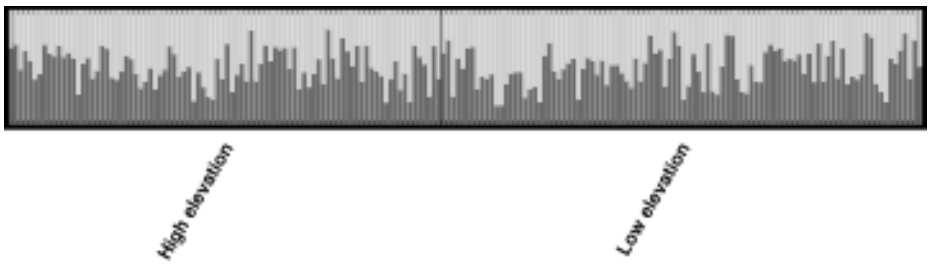

Figure S2. Result from the STRUCTURE analysis of the SNP data for the two sampling sites (High and Low elevation) with K = 2 clusters. Each bar represents one individual tree and the different colors correspond to the membership coefficients for each cluster.

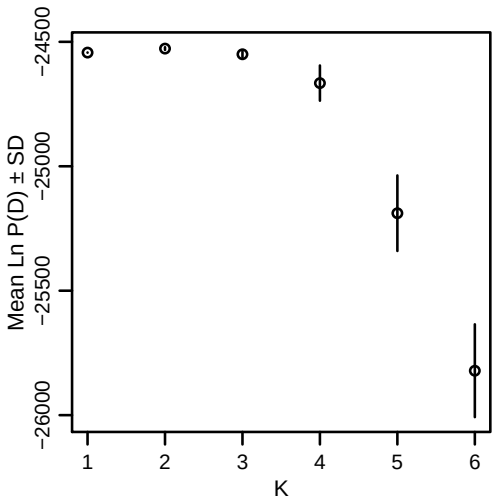

Figure S3. Mean Ln P(D) with corresponding standard deviation (SD) from the Structure analysis of the SNP data between the two sampling sites for each K from 1 to 6.

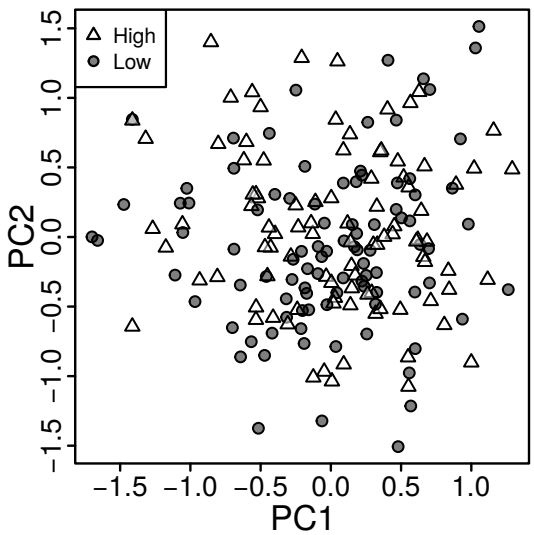

Figure S4. First two axes of the principal component analysis of the SNP data from silver fir trees on the two sampled sites (High and Low). There is no apparent genetic difference visible between the sites.

Table S1. Summary statistics for the temperatures in 2014 and 2015 from dataloggers on the two sites (Low and High) and results of Welch's *t*-tests between the two sites. The mean temperatures (Mean Temp) are not significantly different between the sites, both for 2014 and 2015. The maximum temperatures (Max Temp), on the other hand, differ significantly between sites in both years, with higher temperatures at the lower site.

	2014		2015	
	Mean Temp	Max Temp	Mean Temp	Max Temp
Low (mean ± SD)	8.35 ± 5.78	13.33 ± 6.95	7.85 ± 7.28	12.78 ± 8.74
High (mean ± SD)	7.74 ± 5.50	10.67 ± 6.11	7.09 ± 7.40	10.01 ± 8.26
t	1.32	5.01	1.28	4.03
df	604	596	606	604
p-value	0.1873	<0.001	0.1999	<0.001

SD: Standard deviation, *t*: *t*-statistic, df: degrees of freedom

Table S2. SNPs associated with scaled dendrophenotypes using three different association methods (TASSEL GLM, Boruta=B, VSURF=V). Resistance=Rt, Recovery=Rc, Resilience=Rs, AP=air pollution, EndDep=End of depression period.

SNP	Rt to AP	EndDep	Rc 1976	Rc from AP	Rs 1976	Rt 1976	Slope 1973-74
241.16							
251.177				V			
509.192		V					
628.351	BV			V			
716.144	BV			V			
945.39							V
1113.563							
1181.592							V
1455.289							V
1468.274					V		
1468.511							
1505.255							V
2088.183					V		V
2161.199							
2161.64							
2190.239							
2190.265							BV
2315.209	V						
2387.318							
2495.349							
2831.313							
2937.405						V	
2937.651							
2986.147							
2986.327							
3196.56		V					
3765.215							
3803.208	V						
3908.414							
3942.73							
3942.88							
4336.1071							
4538.344							
4538.47					V	BV	
4911.323							
4921.779							
5004.249							V
5004.671							
5181.245							
5306.661	V						
5502.128							
5502.167							V
5502.725							
5566.13							
5566.56							
5945.417				V			
6119.125							
6331.174							
6641.5							
6757.46							
6968.51		V		V	V		
7067.53							
7082.263							
7082.314	V						
7150.658							
7377.148						V	V
7377.313							
8092.366				BV			
8206.176							
8206.228							
8213.487							
8213.555							
8248.183							
8255.1089							
8649.617						V	
8759.481							

SNP	Rt to AP	EndDep	Rc 1976	Rc from AP	Rs 1976	Rt 1976	Slope 1973-74
8815.372						V	
8855.137		V		BV			
8855.98					V		
9065.486							
9065.715							V
9197.63						BV	V
9263.403						V	V
9373.1559							V
10568.484		V		BV			
11291.4439							V
12178.301					BV	V	V
12182.435							
12900.345							
14101.217		V					
14311.903					V		
14455.953							
14514.32							
14580.627				BV	V		
14623.303							
14915.331							
15135.469		V				V	
15256.604	BV	V			BV	BV	
15256.66		V		V			
15452.813					V		V
15484.667							
15484.725							
15596.971					V		
15944.276							
15944.288							
16332.419					BV		
16356.332	V						
16411.197					V	BV	
16430.504				BV		BV	
16727.97							
16756.89				B			
16774.46							
16822.456		V					
17540.195							V
18067.318		B		V			
18465.35							V
18599.28				V	V		
18599.772							V
18676.245				V			
18676.31							
18676.73							
18762.199							
19054.28		V		V	V		
19173.238							
20694.312							
20694.92					V		
21141.311							
23660.1194	B			V			
23982.493							
24318.117	V			V	V	BV	
24902.128							
24902.464							V
24902.466							V
26572.464							
26602.585							V
26764.484							
27000.1448	V						
30715.525							
30715.72				V			
31121.47			V				

CHAPTER 6

A critical note on random forest based feature selection and interaction analyses in genetic association studies

David Behringer

Manuscript (2017)

A critical note on random forest based feature selection and interaction analyses in genetic association studies

David Behringer

Faculty of Biology, Conservation Biology, Philipps-University Marburg, D–35043 Marburg, Germany

Abstract

Random forest analyses in combination with feature selection techniques are increasingly used to reduce high data dimensionality in genetic association studies. Often the goal is to find a small set of variables that explains the largest portion of the variation in a given phenotype. Such approaches are generally successful in selecting a highly predictive set of variables (e.g. single nucleotide polymorphisms (SNPs)) but suffer from the fact that multiple, equally predictive sets are available. Holliday et al. (2012) developed such an approach called 'backward purging'. Using the fact that random forests capture both marginal and interaction effects of all predictors, they also presented an approach to test for the interaction within a given set of SNPs. This study will show that the backward purging approach produces non-unique results and that the interaction between predictors is highly dependent upon the set of predictors under investigation. In combination, both methods can produce statistically significant but diametrically different results based solely on random sampling. As such, biological interpretation of the SNPs that are identified by backward purging is heavily impeded and in most cases the interaction between predictors cannot be sensibly evaluated. To obtain some useful information from the backward purging procedure an extension of the method, called 'repeated backward purging', is proposed.

Keywords: Machine learning, single nucleotide polymorphism, dendrophenotype, *Abies alba*, *Picea sitchensis*, epistasis

---

Introduction

Machine learning procedures are gaining traction in genetic association studies. This is largely due to the high dimensionality of the data (often coined 'curse of dimensionality' or 'large p, small n problem') (Bellman, 1957) and the resulting need for sufficient sample sizes (Bureau et al., 2005; Cordell, 2009). Taking into account a large number of predictors and their interaction, puts constraints on the applicability of univariate methods. Machine learning algorithms offer an alternative approach to association analysis using datasets with a relatively low number of samples (e.g. individuals) compared to the amount of predictors (e.g. single nucleotide polymorphisms (SNPs)). Among the different machine learning algorithms, random forest procedures (Breiman, 2001) are heavily used in the life sciences and a large body of literature exists on the topic (e.g.

Bureau et al., 2005; Goldstein et al., 2010; Chen and Ishwaran, 2012). Random forests consist of a number of decision trees with the goal to find predictors that explain most of the variation in a given response variable (e.g. some phenotype). Predictors are then ranked based on their importance in explaining variation averaged over all trees. Usually, the next goal is to identify a small set of predictors that explains a relatively large amount of variation. This feature selection involves iteratively removing the least important predictor(s) and re-calculating the importance of the remaining predictors (Díaz-Uriarte and Andrés, 2006). Variations of this approach are already implemented in some R packages, for example **varSelRF** (Díaz-Uriarte and Andrés, 2006) and **AUCRF** (Calle et al., 2011) but were intended for the use in microarray experiments and thus can only handle classification-based problems, i.e. the response variable has to be factorial. Many environmental and/or phenotypic variables are, however, numeric and necessitate a regression-based approach.

Holliday et al. (2012) developed a feature selection procedure, called 'backward purging', that can handle numeric response variables. The approach was developed to associate SNPs with phenological data in Sitka spruce (*Picea sitchensis*) and has been implemented in a number of recent studies (Brieuc et al., 2015; Hornoy et al., 2015; Hess et al., 2016). The backward purging procedure is a powerful tool but has a major caveat: each backward purging run usually results in a slightly different set of predictors with comparable explanatory power. This is a common problem of feature selection techniques aiming at identifying a set of predictors that yield the minimal prediction error and considerably impedes the biological interpretation of the results. To ameliorate this effect I propose an extension of the technique I call 'repeated backward purging', to filter for predictors that are part of most or even all sets. Using the fact that random forests account for the interaction of the predictors, as well as for their marginal effects, Holliday et al. (2012) further tested for possible epistatis (interaction) among the most important SNPs. As will be shown in this paper, this approach is seriously biased by the predictors under investigation and thus, in part, a function of the random selection of a set of the most important predictors. Consequently the biological interpretation of the interaction results can be very problematic.

Materials and Methods

To demonstrate the possible problems with the backward purging procedure I used data from Holliday et al. (2012) that was given as supplementary material. Since Holliday et al. (2012) do not state how they processed their data and only supply genotypes for 251 SNPs (instead of the 339 SNPs they analyzed in their study), a real replication is not possible. Since random forest cannot handle missing data and in order to retain as much SNPs as possible, I removed all individuals with missing values which resulted in 251 SNPs in 109 individuals from three different populations. When substructure exists in the data, random forest can identify SNPs as significant based on their allele frequencies between populations, even when there is no association with the phenotype (Zhao et al., 2012). Thus, I only used the adjusted phenotypes for timing of budset and cold hardiness. Additionally, I used data from Chapter 5 in which random forest procedures were implemented to identify links between SNPs in candidate genes and certain dendrophenotypes associated with air pollution in silver fir (*Abies alba*) in the Bavarian Forest National Park. The dataset consists of 130 SNPs and seven dendrophenotypes for 193 individuals located at two

sites at elevations around 770 m a.s.l. and 1120 m a.s.l, respectively. No population structure is present and the dendrophenotypes were centered and scaled among the sites. For comparison with the Sitka spruce data I only analyzed two dendrophenotypes, the recovery from air pollution and the resistance in the drought year 1976. However, this selection is arbitrary since this study only attempts to highlight certain pitfalls in the backward purging approach and interaction analysis.

Random forest

Random forest is a machine-learning algorithm that grows multiple decision trees, which are closely related to identification keys used to determine the species of a given organism. A forest, then, consists of a collection of decision trees. The random part is introduced at two different levels: bagging and random feature selection (Breiman, 2001). Bagging stands for Bootstrap aggregating, meaning every tree that is grown in the forest gets a different bootstrapped sample of the data. This new dataset has the same dimension as the original set and usually consists of around 63% unique data and 37% duplicates. Random feature selection introduces randomness exclusively on the level of the predictor (bagging applies to both samples and predictors), in order to prevent the repeated selection of overly dominant predictors. For each tree a random subset of predictors is chosen and evaluated which predictor splits the data best, i.e. has the clearest association with the response variable. This process is repeated until the terminal nodes (leaves) contain only samples from the same variable level (classification problems) or the entire dataset is covered (regression problems). The results of all trees are then averaged over the entire forest.

Since every tree is grown based on roughly two-third of the data, the remaining one-third, that were not bagged, hence 'out-of-bag' (OOB) samples, can be used to estimate the accuracy of the prediction. While growing the forest, at each bootstrap iteration the bagged samples are used as training data to predict the OOB samples for the current tree (Liaw and Wiener, 2002). The resulting OOB error (prediction accuracy for regression models) is then repeatedly calculated while each predictor is randomized separately. For each predictor the differences between the 'real' and the randomized OOB error are then averaged over all trees and normalized by the standard deviation. This gives a measure of variable importance (VI) for each predictor in the set.

All random forest calculations presented in this paper were conducted in R (R Core Team, 2016) with the **randomForest** package (Liaw and Wiener, 2002). Commented R-scripts and custom R-functions for all used procedures as well as all datasets can be found on the data CD submitted with this thesis.

Backward purging

Generally, the goal in genetic association studies is to find all predictors (e.g. SNPs) that are significantly associated with a given phenotype (e.g. a disease state or a growth parameter). In terms of machine-learning algorithms, the goal is to select all relevant features. This is generally very hard to achieve (Nilsson et al., 2007). Finding a small set of predictors that explains most of the variation in a given response variable, on the other hand, is easier to manage. Unfortunately, these minimal-optimal approaches tend to result in non-unique sets of predictors with equally good predictive power.

The procedure in Holliday et al. (2012) is such a minimal-optimal feature selection approach and is conducted as follows:

- Step 1: Select a number of SNPs based on previous analyses (I followed the procedure in Holliday et al. (2012) and used the most important SNPs from a full random forest model).
- Step 2: Run three random forest procedures on the data and record the VI for each SNP, as well as the average variance explained over the three models.
- Step 3: Calculate the mean VI for each SNP over the three random forest models and remove the SNP with the lowest mean VI.
- Step 4: Rerun from Step 2 with the new data until there are only two SNPs left.
- Step 5: Pick the set of SNPs with the highest average variance explained.

While the predictive power of this set is maximized, different runs can result in somewhat different sets of SNPs, each with comparable predictive power. To discern between predictors that would be identified in almost each backward purging run and those that were only included based on the random bootstrap samples, multiple runs have to be conducted.

To demonstrate this, I ran four backward purging procedures with the Sitka spruce data from Holliday et al. (2012), as well as with the silver fir data from Chapter 5. The same top 20 SNPs, based on the VI from one full random forest procedure for each variable, were used in the backward purging runs. The number of trees to grow was set to 1500 in all cases.

Repeated backward purging

By conducting only one backward purging run, it is impossible to discern if a predictor would be part of (almost) any run or is highly unique and thus relatively uninformative. In order to determine this relationship I repeated the backward purging procedure multiple times and recorded the occurrence of all predictors in the respective runs. Thus, it was possible to select only those predictors that were present in the majority of runs.

To evaluate the effectiveness of this approach I compared the results from the repeated backward purging procedure with the results from all-relevant feature selection techniques based on random forest, implemented in the R packages **Boruta** (Kursa and Rudnicki, 2010) and **VSURF** (Genuer et al., 2015). VSURF differentiates between variables identified for interpretation and prediction. The latter are always a subset of the former and since the SNPs commonly identified by repeated backward purging are not intended for prediction I used the larger interpretation set for comparison. For the dendrophenotypes of the silver fir data I used the results in Chapter 5. For the phenology data in Sitka spruce I ran Boruta with 2000 trees in each forest and VSURF with the default settings, following the procedure in Chapter 5. SNPs were only considered as identified when they were selected by both Boruta and VSURF since any real study will usually employ more than one method and emphasize on SNPs that were detected based on more than one approach.

Since Holliday et al. (2012) did not provide their entire dataset for replication and the annotation for many genes is missing (including annotated genes in which they discovered SNPs in their study, e.g. xth1), a comparison between the results was not conducted.

Interaction

To test for interaction within the set of most important SNPs, Holliday et al. (2012) iteratively removed one SNP from the set and calculated the VI of the other SNPs and repeated this procedure for the entire set. The exact steps of the procedure are conducted as follows:

- Step 1. Calculate the real VI for each SNP using a random forest model.
- Step 2. Repeat Step 1 multiple times (at least five times).
- Step 3. Remove one SNP and calculate the VI of all other SNPs and repeat this step for all SNPs.
- Step 4. Repeat Step 3 multiple times (at least five times).
- Step 5. Perform Welch's unequal variances *t*-test for each pair of real VI and changed VI upon removal of another SNP (Holliday et al. (2012) used Student's *t*-test but without checking for homogeneity of variance Welch's *t*-test is preferable).
- Step 6. Extract the mean changes in VI with the corresponding 95% confidence intervals and *p*-values from the tests.
- Step 7. To control for the inflating alpha error due to multiple testing, adjust the *p*-values by controlling the false discovery rate as described in Benjamini & Hochberg (1995).

I tested the interaction of the most important SNPs for the Sitka spruce and the silver fir data using the results from the previous backward purging analyses with 20 repetitions and 2000 trees grown in each forest.

Results

Backward purging

For both the Sitka spruce and the silver fir data, the four repeated backward purging procedures resulted in some SNPs that were part of every identified set, while some SNPs were only part of a subset of runs (Fig. 1). The different sets for the budset phenotype (Fig. 1 A) contained seven to 11 SNPs with an average explained variance ranging from 32.2% to 35.4%. For cold hardiness (Fig. 1 B), the sets consisted of nine to 14 SNPs with 40.5% to 42.9% average explained variance. The silver fir data showed a similar fluctuation in identified SNPs over all sets, consisting of 10 to 14 SNPs for the recovery from air pollution (Fig. 1 C) and 10 to 15 SNPs for the resistance in 1976 (Fig. 1 D). Accordingly, the average explained variance ranged from 17.9% to 20.2% and 15.2% to 19.9%, respectively. While the majority of identified SNPs in a single run was part of every backward purging result for the recovery from air pollution, this was not the case for the resistance in 1976, budset timing or cold hardiness.

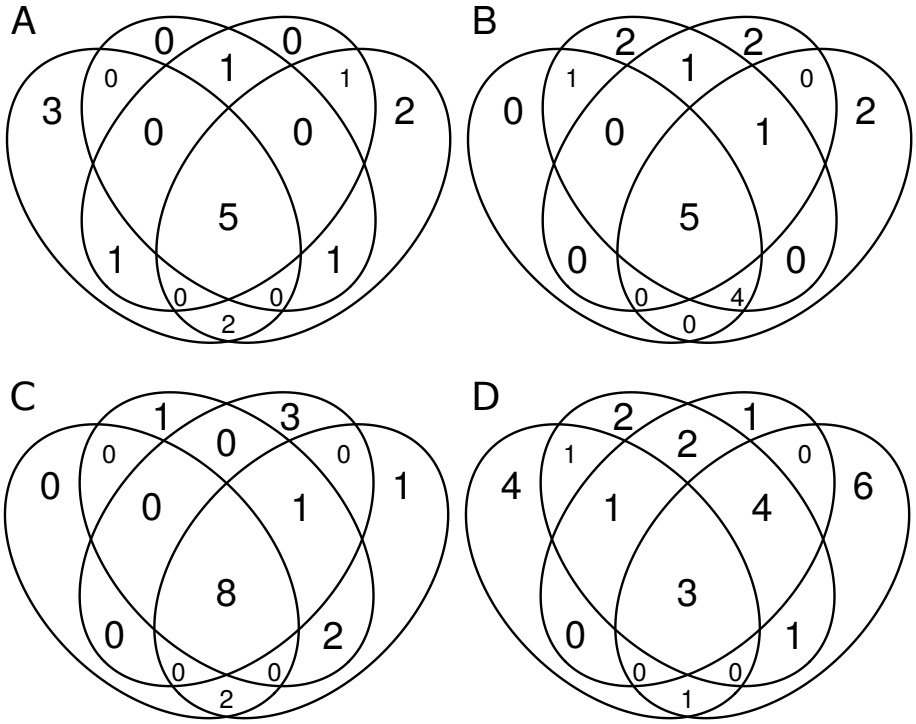

Figure 1. Venn diagrams for the results of four different backward purging runs with SNP data and the phenotypes A) budset and B) cold hardiness from Holliday et al. (2012) and C) the recovery from air pollution and D) the resistance in 1976 from Chapter 5. Each run identifies a different set of SNPs as most predictive with unique SNPs for each run. Some SNPs are commonly identified in all runs and could be selected by the repeated backward purging procedure.

Repeated backward purging

The three different feature selection methods showed a relatively large overlap for budset timing in the Sitka spruce data (Table 1). Repeated backward purging identified four SNPs that were also selected by Boruta and VSURF but failed to find two other SNPs. However, repeated backward purging had one SNP in common with VSURF that Boruta did not select and found another SNP that was not identified by the other methods. For cold hardiness the overlap was less distinct. Only two SNPs were commonly identified by all methods while repeated backward purging did not identify three SNPs that were selected by Boruta and VSURF. Also, repeated backward purging and VSURF commonly selected two SNPs that were not identified by Boruta.

The SNPs that were commonly identified by all backward purging runs in the dendronphenotypes of the silver fir data included some, but not all, SNPs that were identified in Chapter 5 (Table 2). Using Boruta and VSURF, Chapter 5 identified five SNPs as important for the recovery from air pollution and six for the resistance in 1976. Repeated backward purging found all SNPs for the recovery from air pollution and three of six for the resistance in 1976. Thus, repeated backward purging lacked to identify three SNPs that were found with other methods for the resistance in 1976 and none for the recovery from air pollution but found three SNPs that were not identified by other methods for the recovery from air pollution.

Table 1. SNPs identified by Boruta, VSURF and repeated backward purging (repBackPurge) for the two phenotypes in Sitka spruce from Holliday et al. (2012). Four SNPs are commonly identified by all methods for budset timing and two SNPs for cold hardiness.

SNP	Budset timing	Cold hardiness
258_207_S	-	Boruta, VSURF
162_350_S	Boruta, VSURF, repBackPurge	Boruta, VSURF, repBackPurge
19_567_S	Boruta, VSURF, repBackPurge	-
237_231_S	VSURF, repBackPurge	-
169_375_NS	-	VSURF, repBackPurge
114_144_S	Boruta, VSURF, repBackPurge	-
162_39_S	repBackPurge	-
232_195_S	-	Boruta, VSURF
62_359_NS	-	VSURF, repBackPurge
257_105_S	Boruta, VSURF	-
266_573_S	Boruta, VSURF, repBackPurge	Boruta, VSURF
162_289_S	Boruta, VSURF	Boruta, VSURF, repBackPurge

Table 2. SNPs identified by Boruta, VSURF and repeated backward purging (repBackPurge) for the two dendrophenotypes in silver fir from Chapter 5. Five SNPs are commonly identified by all methods for the recovery from air pollution and three SNPs for the resistance in 1976.

SNP	Recovery from air pollution	Resistance in 1976
04538.470	-	Boruta, VSURF
08092.366	Boruta, VSURF, repBackPurge	-
08855.137	Boruta, VSURF, repBackPurge	-
09197.63	-	Boruta, VSURF
14580.627	Boruta, VSURF, repBackPurge	-
15256.604	-	Boruta, VSURF, repBackPurge
15256.660	repBackPurge	-
16411.197	-	Boruta, VSURF
16756.89	repBackPurge	-
24318.117	-	Boruta, VSURF, repBackPurge
05945.417	repBackPurge	-
10568.484	Boruta, VSURF, repBackPurge	-
16430.504	Boruta, VSURF, repBackPurge	Boruta, VSURF, repBackPurge

Interaction

Repeated tests for interaction within one SNP set yielded relatively stable results. This can be seen by the 95% confidence intervals for the mean of 20 repetitions (Fig. 2 and 3). However, the interaction between different predictors was clearly influenced by the selection of SNPs under investigation, both for the Sitka spruce data (Fig. 2) and the silver fir data (Fig. 3). Some pairwise SNP combinations showed positive interaction (the mean VI of one SNP was reduced upon the removal of the other SNP) in one set and negative interaction (the mean VI of one SNP was increased upon the removal of the other SNP) in another set. In many cases the interaction was statistically significant at the $\alpha = 0.05$ level. For example SNP 62_359_NS upon removal of SNP 169_375_NS for cold hardiness in the Sitka spruce data (Fig. 2 C). For both SNPs the 95% confidence intervals

did not overlap with zero and even after adjusting for multiple testing the effect was still significant (adjusted p-values $<< 0.001$). The same inconsistency could be observed in the silver fir data, for example for the resistance in 1976 (Fig. 3 C). Removing SNP 15256.604 resulted in an increase in importance of SNP 24318.117 in one run and a decrease in another but equivalent run. Again, the mean shift in VI was statistically significant within each run (adjusted p-values $<< 0.001$). Consequently two different studies, both conducting one backward purging run with identical data, would have drawn a statistically significant yet diametrically different conclusion.

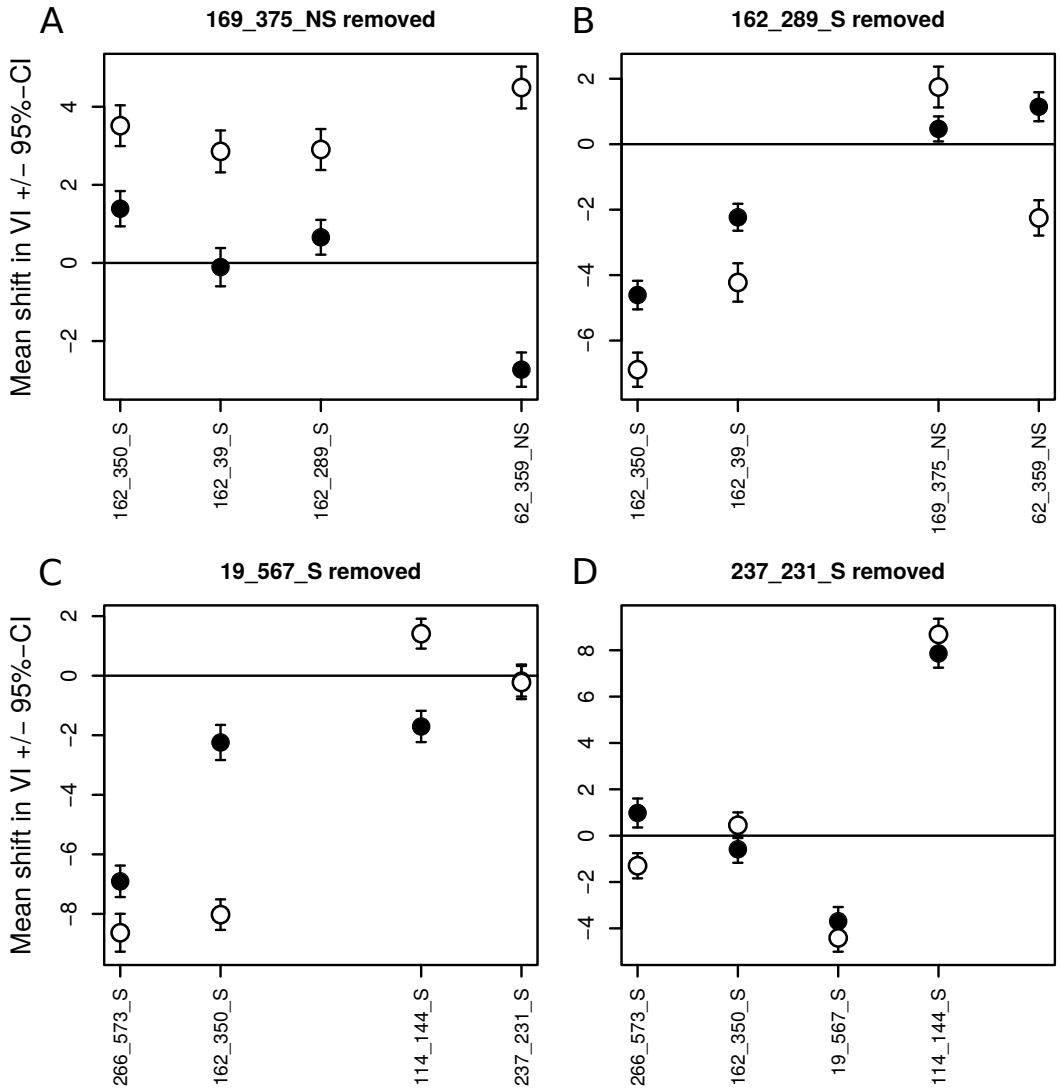

Figure 2. Changes in mean variable importance (VI) and corresponding 95% confidence intervals (CI) for all SNPs that are part of the selection in all previously conducted backward purging runs for (A and B) cold hardiness and (C and D) budset timing when (A) SNP 169_375_NS and (B) SNP 162_289_S is removed and (C) SNP 19_567_S and (D) SNP 237_231_S is removed. Depending on the partly arbitrary selection in each run, some SNPs show positive interaction in one run (black) and negative interaction in another, equivalent run (white) and vice verca.

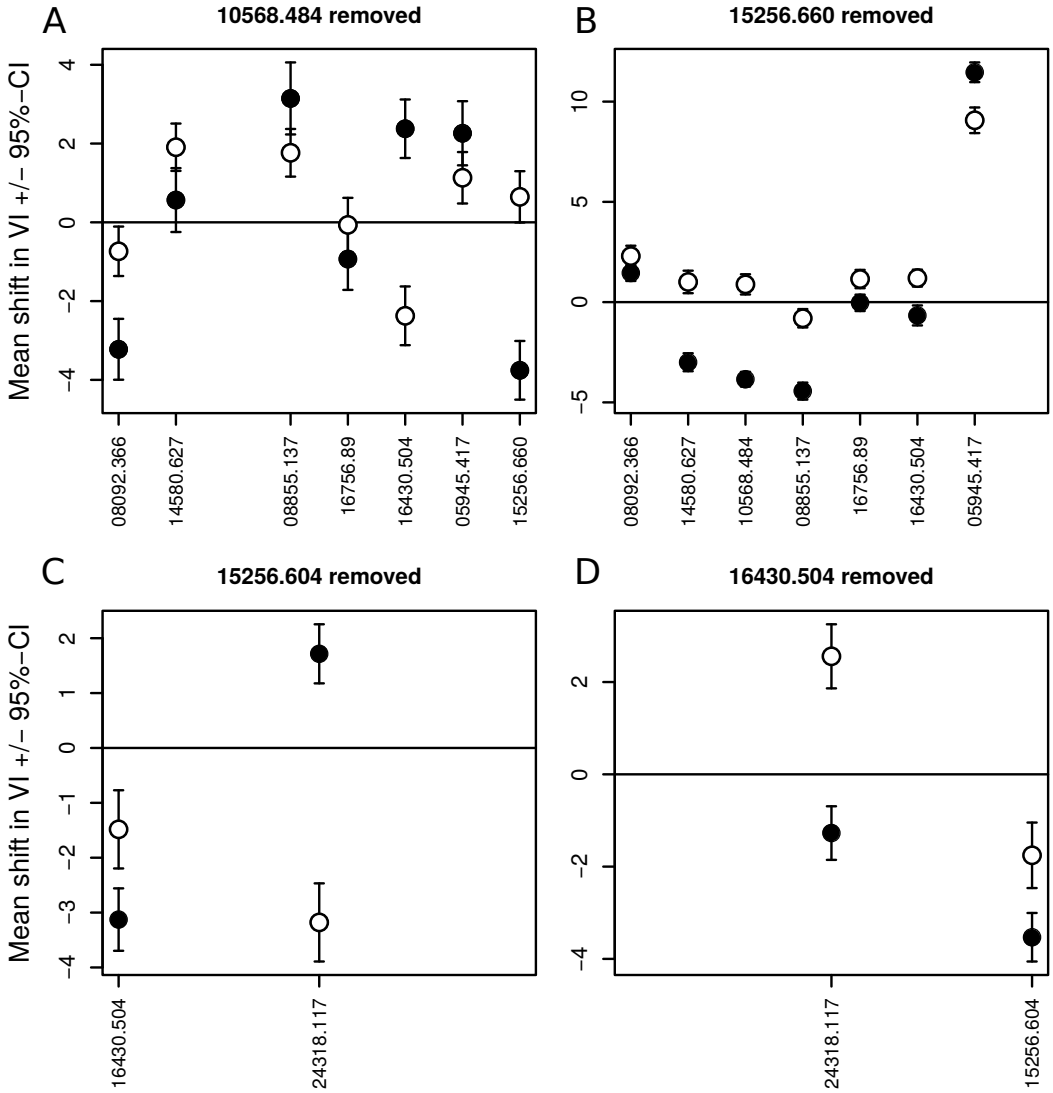

Figure 3. Changes in mean variable importance (VI) and corresponding 95% confidence intervals (CI) for all SNPs that are part of the selection in all previously conducted backward purging runs for (A and B) the recovery from air pollution and (C and D) the resistance in 1976 when (A) SNP 10568.484 and (B) SNP 15256.660 is removed and (C) SNP 15256.604 and (D) SNP 16430.504 is removed. Depending on the partly arbitrary selection in each run, some SNPs show different interactions depending on the run (black and white).

Discussion

Backward purging is a powerful tool and certainly manages to identify a small set of predictors that explains a large portion of the variation in any variable. However, it can also tempt the investigator to falsely interpret the results. With its sole purpose being the maximization of prediction accuracy, the backward purging procedure as described in Holliday et al. (2012) should not be nonchalantly used for interpretation. With only one run performed, a single set of SNPs cannot be labeled as important regarding a certain phenotype. Because if this set is deemed important, why not the set from the second run, or the third? All backward purging can do, is approximate the potential number of predictors that can explain most of the variation in a given variable.

As the analyses presented in this paper show, many runs identify the same SNPs as most important. Such SNPs should be more trustworthy, since their VI seems independent of bootstrap

sampling. Thus, it stands to reason that among most backward purging sets are some truly important predictors, mixed in with a number of weakly interacting predictors whose selection is due to random sampling. Repeating the backward purging procedure multiple times and selecting only commonly identified predictors can help to identify SNPs for interpretation. However, the repeated backward purging procedure gives mixed results when compared with other all-relevant feature selection techniques, ranging from a 50/50 to a complete overlap. Given that repeated backward purging is a reduction of the minimal-optimal backward purging approach, a lack in overlap with all-relevant feature selection techniques is to be expected. To stress this point, repeated backward purging should not be able to identify all relevant features and rather tend to identify predictors with a large marginal effect.

In any case, four repetitions are certainly not enough to reliably discern between SNPs, although they give a small level of confidence, especially when filtering for SNPs that are part of every run. The number of repetitions necessary is hard to assess and seems to be heavily reliant upon the dataset under investigation. While the Sitka spruce and silver fir data will surely show further variation beyond the four replications that were presented in this study, other datasets might not deviate substantially from the findings in a few repetitions. The 'Ozone' data (Lichman, 2013), for example, is rather low-dimensional (one dependent variable with 203 entries and 12 predictors) and running the backward purging procedure 100 times gives the same eight predictors in 92% of the runs (data not shown). The remaining 8% of the runs result in six and once in seven predictors that are all part of the other runs.

It should be noted that commonly identified SNPs have, as a set, lower predictive accuracy than the previously identified backward purging sets. This can be directly derived from the backward purging procedure. Since all commonly identified SNPs are part of each backward purging run they represent a reduction of the set with the largest variance explained. Reducing this set will inevitably lead to lower prediction accuracy (Fig. S1).

As a minimal-optimal feature selection technique, backward purging is a sound approach for identifying sets of SNPs for the prediction of a given phenotype. The lack of uniqueness, however, casts doubt on the applicability of the results since multiple, equally good sets are available. A possible solution could be the approximation of the most predictive set by repeating the backward purging procedure multiple times and filtering for the set with the largest variance explained. The major drawback here is the dimensionality of the data. A large number of runs with a high dimensional dataset would take a very long time to process and the difference in variance explained might be so minuscule between sets that this approach seems unrewarding.

Even given the apparent drawbacks, especially the lack of biological interpretability, backward purging still produces relatively straightforward results. The interaction analysis in Holliday et al. (2012), on the other hand, is more problematic. Holliday et al. (2012) argued that interaction can be detected by the change in VI of all SNPs upon removing one SNP at a time. As was shown in this study, by removing one SNP and re-calculating the VI of all other SNPs, the interaction between the removed SNP and all other members of the SNP set can indeed be quantified relatively reliably. This interaction, however, is highly dependent upon the SNPs under investigation. Applied to a system with a fixed set of predictors, the interaction analysis can be used to quantify the interaction of all these predictors. Given a system with a variable number of predictors that can

change in composition, on the other hand, will inevitably lead to different conclusions. Genomic datasets generally consist of a large number of predictors that usually only represent a fraction of all existing predictors. Any study designed to associate SNPs with phenotypes will consequently be undersampled. Further, association studies try to reduce dimensionality in the data to select the predictors that are most strongly associated with a phenotype. Such a selection will very often produce variable results since different methods are employed and no single approach is trustworthy on its own. Backward purging is such an approach and the instability of the results was presented. However, other methods fare equally and will always show variation in the results. This fact alone makes the interaction analysis, at best, questionable in the context of association studies. Combined with the backward purging approach, the results are all but guaranteed to be, in part, a product of random sampling.

In conclusion, backward purging should be used with caution, especially regarding the interpretation of the results. Repeated backward purging might be a rough fix to extract some biological information but should always be combined with other methods. Interaction between predictors should only be studied in systems where the selection of predictors will not lead to variable outcomes, as is the case for most genetic association studies.

Acknowledgments

I would like to thank Katrin Heer, Lars Opgenoorth, Sascha Liepelt and Birgit Ziegenhagen for all the fruitful discussions.

References

Bellman, R. (1957), *Dynamic Programming*, Princeton University Press.

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289–300.

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**(1), 5–32.

Brieuc, M. S. O., Ono, K., Drinan, D. P. and Naish, K. A. (2015), 'Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (Oncorhynchus tshawytscha)', *Mol Ecol* **24**(11), 2729–2746.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. and Van Eerdewegh, P. (2005), 'Identifying SNPs predictive of phenotype using random forests', *Genet. Epidemiol.* **28**(2), 171–182.

Calle, M. L., Urrea, V., Boulesteix, A.-L. and Malats, N. (2011), 'AUC-RF: A New Strategy for Genomic Profiling with Random Forest', *Human Heredity* **72**(2), 121–132.

Chen, X. and Ishwaran, H. (2012), 'Random forests for genomic data analysis', *Genomics* **99**(6), 323–329.

Cordell, H. J. (2009), 'Detecting gene–gene interactions that underlie human diseases', *Nat Rev Genet* **10**(6), 392–404.

Díaz-Uriarte, R. and Andrés, S. A. d. (2006), 'Gene selection and classification of microarray data using random forest', *BMC Bioinformatics* **7**(1), 3.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C. (2015), 'VSURF: An R Package for Variable Selection Using Random Forests', *The R Journal* **7**(2), 19–33.

Goldstein, B. A., Hubbard, A. E., Cutler, A. and Barcellos, L. F. (2010), 'An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings', *BMC Genetics* **11**, 49.

Hess, J. E., Zendt, J. S., Matala, A. R. and Narum, S. R. (2016), 'Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing', *Proc. R. Soc. B* **283**(1830), 20153064.

Holliday, J. A., Wang, T. and Aitken, S. (2012), 'Predicting Adaptive Phenotypes From Multilocus Genotypes in Sitka Spruce (Picea sitchensis) Using Random Forest', *G3* **2**(9), 1085–1093.

Hornoy, B., Pavy, N., Gérardi, S., Beaulieu, J. and Bousquet, J. (2015), 'Genetic Adaptation to Climate in White Spruce Involves Small to Moderate Allele Frequency Shifts in Functionally Diverse Genes', *Genome Biol Evol* **7**(12), 3269–3285.

Kursa, M. B. and Rudnicki, W. R. (2010), 'Feature selection with the Boruta package', *Journal of Statistical Software* **36**(11).

Liaw, A. and Wiener, M. (2002), 'Classification and regression by randomForest', *R news* **2**(3), 18–22.

Lichman, M. (2013), 'UCI Machine Learning Repository'.

Nilsson, R., Peña, J. M., Björkegren, J. and Tegnér, J. (2007), 'Consistent feature selection for pattern recognition in polynomial time', *Journal of Machine Learning Research* **8**(Mar), 589–612.

R Core Team (2016), 'R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna Austria. URL https://www.R-project.org/'.

Zhao, Y., Chen, F., Zhai, R., Lin, X., Wang, Z., Su, L. and Christiani, D. C. (2012), 'Correction for population stratification in random forest analysis', *Int. J. Epidemiol.* **41**(6), 1798–1806.

Supplementary Material

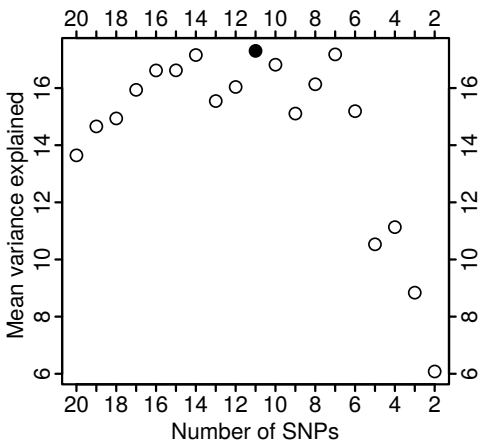

Figure S1. Mean variance explained for each set of SNPs in the course of a backward purging run for the resistance in 1976 in the silver fir data from Chapter 5. This plot can be produced by setting `Plot=T` in the backward purging function (R-script on the data CD).

CHAPTER 7

Synthesis

A short recapitulation

In this thesis two approaches to candidate gene selection were presented. First, an experimental approach attempted to set up a highly controlled environment (Chapter 2) to capture the gene regulatory response of silver fir seedlings to drought stress (Chapter 4). As such, gene expression represents the phenotypic reaction of the plants towards water shortage. The association between the differentially expressed genes and the water status is straightforward, with the assumption of a cause-effect relationship between drought stress and consequent regulation of the necessary genes to cope with the situation. Apart from the obvious caveat of missing biological replication, the experimental setup is a viable approach towards functional candidate gene selection. Since no pre-selection of genes is necessary and the entire transcriptome is sampled, novel insights into previously unknown genes are possible. An obvious example are the putative uncharacterized proteins (Chapter 4) that are differentially expressed as a reaction to drought stress in silver fir seedlings but are not yet described in the literature. This opens up the possibility to find genes in non-model organisms that might not be accessible in model organisms. For example, orphan genes that are unique to particular groups (Tautz and Domazet-Lošo, 2011) could be identified in such an approach. This is especially interesting since many stress-inducible genes are shared between model organisms, such as rice and Arabidopsis (Rabbani et al., 2003).

Highly controlled environments are very complicated to set up and maintain. Confounding factors have to be controlled for while ensuring the necessary conditions to reduce any stress for the subject that is not the goal of the study. Depending on the organism this can be virtually impossible to achieve. Forest trees are such an example where experimental setups will mostly be unfit for adult individuals. Seedlings on the other hand can easily be managed in a laboratory and thus used in experimental studies. As the terahertz monitoring shows, this allows for a very tight control over the plant water status and consequent phenotyping (Chapter 2). As such, the experimental approach necessitates a rather complicated and resource-intensive setup prior to data collection. Any influence of other factors on the measured signal, for example changing chlorophyll content (Chapter 3), has to be accounted for. The collection of data itself is then relatively straightforward and the data are 'easy' to analyze (Chapter 4).

Observational studies on the other hand cannot highlight cause-effect relationships but are necessary to gain a much more 'realistic' picture. Since experimental studies are highly controlled, they necessarily do not represent the actual situation of living organisms and especially adults. There, positive and negative interactions both with other organisms such as mycorrhiza or bacteria as well as habitat specific abiotic factors, shape the genomic reaction of every individual. As such, the setup and sampling for an observational study are relatively straightforward, especially studying large, sessile organisms such as forest trees. Due to the myriad of confounding factors, however, any conclusion must be interpreted with caution. Further, since conifer genomes are very large and reference sequences are usually unavailable, financial and organizational restrains warrant a pre-selection of genes under investigation. This inevitably small subset of the much larger genome, in combination with an unknown number of unaccounted, confounding factors, shifts the burden of the observational approach towards data analysis. As could be shown, more computationally intensive approaches, such as random forest, can be more powerful and fitting than classical statistical calculations like simple linear models (Chapter 5). However, these approaches introduce

new problems and pitfalls, mainly increased difficulty in biological interpretation of the results (Chapter 6).

Limitations and outlook

Functional candidate genes

Terahertz spectroscopy is a highly useful technology to monitor water status, especially in coniferous species where optical assessment is very hard. The condition of this approach to be conducted in a laboratory environment, however, limits the generalization of the results. Since only seedlings could be monitored, the functional candidate genes identified in Chapter 4 do not necessarily have to be involved in the drought stress response of adult individuals. The case can be made for previously annotated genes that have been shown to be involved in drought stress response in model organisms but less so for previously unknown genes. This, however, does not imply that the approach is not very promising. Especially silver fir faces many threats particularly during earlier life stages, such as browsing by deer (Wolf, 2003), and consequently the seedling stage is a focus of interest.

A further limitation regarding the generalization of the functional candidate genes concerns the source material. Since the transcriptome analysis was only conducted on needles, the differentially expressed genes are candidates for drought stress response within this part of the seedlings. Microarray studies on Arabidopsis suggest that stress responsive gene regulation differs between roots and leaves (Kreps et al., 2002) and as such, further research into the drought stress response within below-ground material in silver fir seedlings is clearly necessary. If this can be properly conducted using the terahertz approach remains to be seen. An easy first solution would be to repeat the experiment as described in Chapter 2 and harvest the seedlings' roots for transcriptome analysis. However, a more sophisticated approach would include measuring the water content in the roots directly but this might be technically challenging.

The identified functional candidate genes correspond to a specific phase of the drought stress response. Terahertz monitoring allowed to ensure that all seedlings within one group had the same amount of water loss and thus approximately the same level of stress. Given that stress response has different phases, however, it is possible, and even likely, that each phase leads to just slightly or even major differences in the molecular response. Again, in Arabidopsis, stress has been shown to first induce a general response and upon duration of the stress a shift towards a more specific response (Kreps et al., 2002). This suggests the possibility that the candidate genes identified in Chapter 4 are either general stress response genes, specific drought stress response genes or a mixture of both, depending on the phase of stress response of the seedlings. Next steps should thus include the resolution of gene regulation in time. This would include monitoring the water status and harvesting needles at different, but comparable, stress levels. Since the loss of needles will inevitably induce further stress for the plant, this approach should be conducted on older individuals that still comfortably fit inside a laboratory but have enough needles in order to limit the amount of additional stress. Given that in the course of this thesis functional drought stress candidate genes have been identified already, the time-resolving approach could be limited to studying genes that were pre-selected based on the results of Chapter 4. This would have the

benefit of using cheaper and less extensive methods such as microarrays or reverse-transcription quantitative real-time PCR.

While the identification of functional drought stress responsive candidate genes is a legitimate scientific pursuit, some general criticism is warranted. This concerns the applicability of the results regarding actual populations. Gradually analyzing the genetic architecture of stress response in forest trees should be focused on understanding the role of genes for functional and regulatory proteins with the aim of predicting the fate of populations in the face of rapidly changing environments. This includes the interaction of multiple stresses and the influence of associated organisms, such as pathogens or mycorrhiza. Terahertz monitoring could be implemented in this context in different ways. For example, the influence of interacting stressors such as drought and sulfur dioxide (SO_2) on the gene regulatory response would necessitate setting up a closed-chamber system to fumigate seedlings with specific concentrations of SO_2 gas, while simultaneously applying specific levels of drought stress. An even more complex design could incorporate the presence and absence of different mycorrhizal species, competing plants and pathogens, as well as different soil nutrient concentrations, light availability and temperature. Lastly, seedlings from different provenances, potentially adapted to different climates, could be incorporated. This is an important aspect since the identified candidate genes in Chapter 4 could be the result of local adaptation to the specific conditions of the seed material.

Gradually, a habitation model could be constructed for silver fir seedlings, aiding in the understanding of the influence and interaction of different factors on gene expression. This could also add to the growing knowledge of the effect of genotypic interaction (G x G) between different organisms in plant-soil feedback systems (Whitham et al., 2006; Van Nuland et al., 2016). Since an experimental setup allows to control environmental factors and terahertz monitoring can ensure comparable water content, the effect of plant genotypes on the expression of genes in the soil community (the 'extended phenotype' of the plant genes; Dawkins, 1982), and vice versa, can be studied in isolation. Using different provenances might also shed light on locally adapted genotypes, shaping the extended phenotype (Gugerli et al., 2013).

However, this approach has clear limitations. Terahertz monitoring relies on the fact that the plant material does not grow significantly during the measuring period. While this makes silver fir better suited for this technique than other fast-growing plants, it limits the observable phenotype. In Chapter 4 the only phenotypic variation that was measured was the gene expression, i.e. the amount of transcripts. Yet, regarding the fate of individual trees in populations under changing climatic conditions, gene expression might not be a very useful phenotype. Information on the functional association of a gene to a specific stress alone does not directly provide information about the biological relevance of this gene. Further information is necessary, such as the impact of this gene or its variation (e.g. SNPs within the gene) on a tree's performance, e.g. growth, pathogen resistance or seed size. Association with these performance traits must be conducted in a different type of experiment.

Polymorphic candidate genes for trait variation

Associating variation in candidate genes, in the form of SNPs, with trait variation, offers the ability to identify genes whose variants might influence the phenotype of its carrier. As such, the genes

identified in the approach in Chapter 5 of this thesis show variation that is associated with potentially adaptively relevant traits. The dendrophenotypes are derivatives of tree-ring width and a measure of stress coping capability. However, this approach has a number of drawbacks. As mentioned earlier, no direct cause-effect relationships can be established as a consequence of the observational nature of the study. The identified SNPs could in fact be the cause for the variation in the dendrophenotypes but they could also be markers for this variation because they are physically close to the true causal genetic variant. There could also be confounding factors whose influence could result in phenotypic variation. This introduces the problem that variation in a dendrophenotype is not caused by a SNP but by an environmental factor dis-proportionally influencing individuals with a specific SNP genotype.

Apart from this general restriction, the study design in Chapter 5 lacks specificity regarding the exact nature of the stress leading to the growth depression in the individual trees. The dendrophenotypes are measures for individual coping efficiency but it is not entirely clear what stressors caused the depression period. Elling et al. (2009) distinctly identified SO_2 pollution as the main driver of the depression in silver fir and this fits the data in Chapter 5 rather well. Additionally, an interaction with other stressors, mainly drought, are likely but since there is not sufficient local data on SO_2 concentration and water availability, the causal factors remain speculative. Interestingly, most of the genes with an associated trait variation in Chapter 5 were related to photosynthesis and only few to drought. This might indicate that SO_2 pollution was the major driver of the depression period and that drought played a minor role. It would be interesting to compare these results with those from terahertz monitoring-differential expression experiments that incorporate drought stress and SO_2 fumigation. This could highlight if the identified SNPs lie within genes that play a functional role in the combination of drought and SO_2 stress response.

In any case, a low turnout in significant associations was to be expected, given the low preselected, number of SNPs analyzed. Compared to the massive genome, this represents a tiny fraction of possible options and thus a low probability of having selected the 'right' genes. Further, for SNPs causing small effect sizes in phenotypes, a large number of sampled individuals is necessary to detect significant associations. Based on association studies in other conifer species such as *Pinus taeda* (González-Martínez et al., 2006, 2008) and *Pseudotsuga menziesii* var. *menziesii* (Eckert et al., 2009), a single SNP explains less than 5% of the variation in a trait and this suggests that most SNPs have polygenic effects. Consequently, a larger number of individuals would have likely resulted in the significant identification of more, small-effect associations.

Also, the pre-selection of candidate genes should be partially based on results from experimental studies, such as the terahertz-monitoring approach in Chapter 2 and 4. Using primarily functional candidate genes will increase the likelihood of finding significant associations. Non-functional preselection and low-dimensional datasets are often the reason for the inability to reproduce the results of association studies with candidate genes, which has led to severe criticism of this approach (Tabor et al., 2002).

Another major problem that association studies have to deal with is the reliance on the proper statistical methods. Simple approaches, such as analysis of variance (ANOVA), are often not applicable because the conditions of the test, namely normal distribution of the trait values within each genotype, as well as equal variance (Balding, 2006*a*), are not met. Choosing alternative tests can

be problematic, however, since there is little consensus among scientists and there is a plethora of methods to choose from, ranging from frequentist (Balding, 2006*b*) over bayesian (Stephens and Balding, 2009) to machine learning (Libbrecht and Noble, 2015) approaches. For genetic association studies, machine learning procedures are gaining traction and are usually based on the random forest algorithm. A major motivation for applying this approach is the increasing use of SNPs in association studies and the desire to capture both their marginal, as well as their interaction effect (Bureau et al., 2005). As could be shown in Chapter 6, however, some of these techniques can lead to results that are not unique and thus might not have any biological meaning. This makes the interpretation of the results very problematic and highlights the fact, that most random forest procedures are intended for building predictive models. If any consistent prediction will ever be possible based on such results remains to be seen. Stochasticity is ever present in natural tree populations and can lead to discrepancies between model predictions and real outcomes (Aitken et al., 2008). Hence, even if a model might have a relatively good fit and would predict a good performance of a population under increasing temperatures, the warmer climate might also lead to an increase in some herbivore who decimates the tree population.

Aside from the restrictions imposed on this type of study, dendrophenotypes are a promising measure for genetic association. Wood cores provide relatively easily accessible data ranging over long time periods, which is particularly useful for long-lived, sessile organisms, such as trees. Further, dendrophenotypes, as derived measures for performance in response to extreme environmental stress, should arguably be of adaptational consequence. This in turn makes associated genes valuable for further studies regarding selection processes in natural populations. It remains questionable, however, how variable dendrophenotypes are within an individual. Further research must focus on the ratio of intra- and inter-individual variation to gauge the resolution and accuracy of this novel phenotypic measure. Especially regarding association studies, precision in measuring the phenotype is very important (Neale and Savolainen, 2004).

Concluding remarks

Identifying candidate genes for stress response in silver fir is a heavy task. Many of the most promising methodologies are not directly applicable and there is increasing need to develop techniques that allow to unravel the complex genetic architecture of this important forest tree. The work comprised in this thesis is aimed at providing a framework for the identification and analysis of candidate genes for stress response in silver fir. It consists of bringing silver fir into the laboratory for controlled experiments to identify functional candidate genes using a novel terahertz monitoring setup, as well as novel dendro-phenotypic measures for genetic association of potentially adaptive traits within natural populations (Fig. 1).

Both approaches can and should be combined, for example by searching for variation within functional candidate genes and using them in association studies. This will greatly be aided by more and better reference sequences which are currently rare for silver fir (Roschanski et al., 2013). Another option would be the comparison of results, which in the work for this thesis resulted in one drought responsive gene being identified by both approaches (glucan-endo-1,3-beta-glucosidase, Chapter 4 and 5), which adds to the credibility of the association.

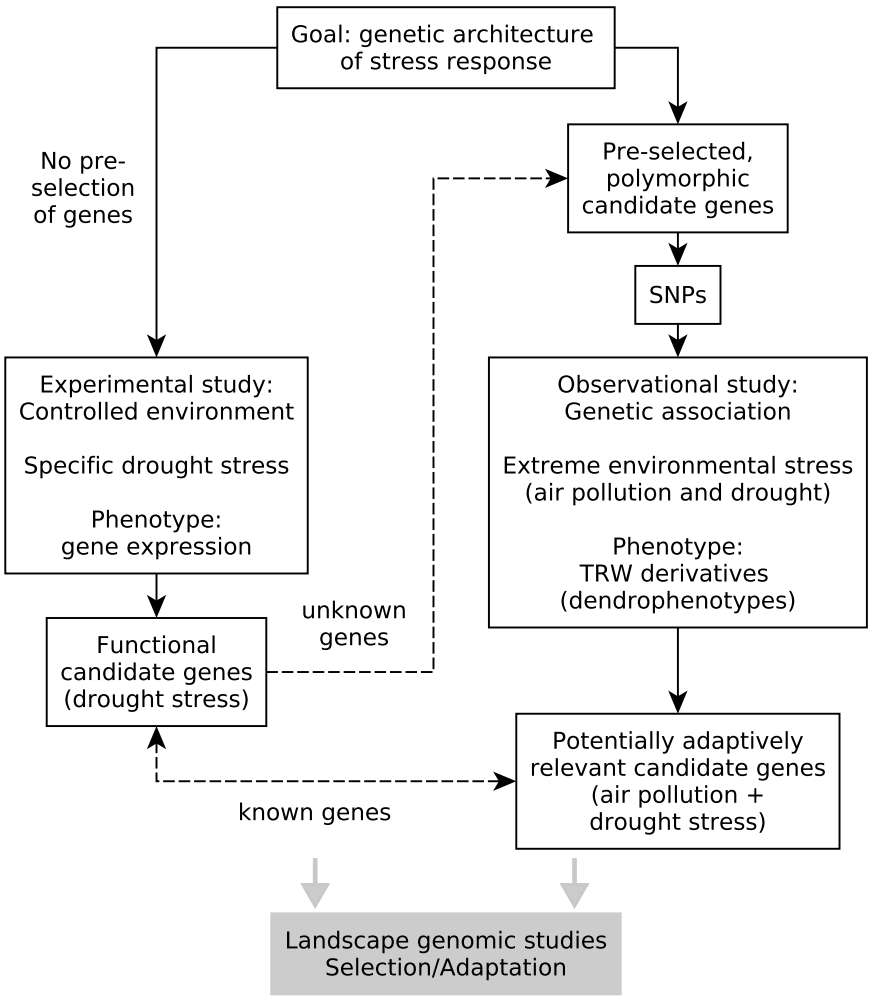

Figure 1. A framework for the selection and analysis of candidate genes for stress response in silver fir. Functional candidate genes for drought stress response can be identified by differential gene expression in an experimental setup using terahertz monitoring. This yields previously unknown genes that should be part of the pre-selection of polymorphic candidate genes for which variations (SNPs) can be identified. These SNPs can then be associated with dendrophenotypes, based on tree-ring width (TRW) data, in an observational study to identify potentially adaptively relevant candidate genes for extreme environmental stress response (e.g. air pollution and drought). Known genes, identified in both approaches, can be compared and used in further studies, aimed at identifying adaptively relevant genes under selection (e.g. landscape genomic studies).

Growing resources in the form of functional and potentially adaptively relevant candidate genes will greatly benefit landscape genomic approaches that incorporate information of phenotype, genotype and local environment across a landscape scale (Sork et al., 2013). While these studies can be of great value regarding forest management and conservation, they rely on knowledge about genomic regions that shape locally adaptive phenotypes. Among other aspects, the lack of knowledge regarding the genetic architecture underlying complex, adaptively relevant traits is one of the main reasons for the low commonality in results of landscape genomic studies on forest trees (Ćalić et al., 2015).

In conclusion, both approaches presented in this thesis are promising tools and should be further integrated in the framework of candidate gene selection and analysis in silver fir and other non-model conifer species, to hopefully contribute to better predict and manage the fate of populations in the face of ever more intensive land-use and rapid climate change.

Bibliography

Adams, H. D., Guardiola-Claramonte, M., Barron-Gafford, G. A., Villegas, J. C., Breshears, D. D., Zou, C. B., Troch, P. A. and Huxman, T. E. (2009), 'Temperature sensitivity of drought-induced tree mortality portends increased regional die-off under global-change-type drought', *PNAS* **106**(17), 7063–7066.

Ahuja, M. R. and Neale, D. B. (2005), 'Evolution of genome size in conifers', *Silvae genetica* **54**(3), 126–137.

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T. and Curtis-McLane, S. (2008), 'Adaptation, migration or extirpation: climate change outcomes for tree populations', *Evolutionary Applications* **1**(1), 95–111.

Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., Kitzberger, T., Rigling, A., Breshears, D. D., Hogg, E. T., Gonzalez, P., Fensham, R., Zhang, Z., Castro, J., Demidova, N., Lim, J.-H., Allard, G., Running, S. W., Semerci, A. and Cobb, N. (2010), 'A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests', *Forest Ecology and Management* **259**(4), 660–684.

Armenise, L., Simeone, M. C., Piredda, R. and Schirone, B. (2012), 'Validation of DNA barcoding as an efficient tool for taxon identification and detection of species diversity in Italian conifers', *Eur J Forest Res* **131**(5), 1337–1353.

Ayres, M. P. and Lombardero, M. J. (2000), 'Assessing the consequences of global change for forest disturbance from herbivores and pathogens', *Science of The Total Environment* **262**(3), 263–286.

Balding, D. J. (2006*a*), 'A tutorial on statistical methods for population association studies', *Nat Rev Genet* **7**(10), 781–791.

Balding, D. J. (2006*b*), 'A tutorial on statistical methods for population association studies', *Nat Rev Genet* **7**(10), 781–791.

Barber, V. A., Juday, G. P. and Finney, B. P. (2000), 'Reduced growth of Alaskan white spruce in the twentieth century from temperature-induced drought stress', *Nature* **405**(6787), 668–673.

Barrett, R. D. and Schluter, D. (2008), 'Adaptation from standing genetic variation', *Trends in Ecology & Evolution* **23**(1), 38–44.

Behringer, D. (2013), Identifying candidate genes for drought stress response in Abies alba seedlings using terahertz time-domain spectroscopy and gene expression profiling, Master thesis, Philipps-Universität Marburg, Marburg.

Bengtsson, J., Nilsson, S. G., Franc, A. and Menozzi, P. (2000), 'Biodiversity, disturbances, ecosystem function and management of European forests', *Forest Ecology and Management* **132**(1), 39–50.

Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., Yuen, M. M. S., Keeling, C. I., Brand, D., Vandervalk, B. P., Kirk, H., Pandoh, P., Moore, R. A., Zhao, Y., Mungall, A. J., Jaquish, B., Yanchuk, A., Ritland, C., Boyle, B., Bousquet, J., Ritland, K., MacKay, J., Bohlmann, J. and Jones, S. J. M. (2013), 'Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data', *Bioinformatics* **29**(12), 1492–1497.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. and Van Eerdewegh, P. (2005), 'Identifying SNPs predictive of phenotype using random forests', *Genet. Epidemiol.* **28**(2), 171–182.

Ćalić, I., Bussotti, F., Martínez-García, P. J. and Neale, D. B. (2015), 'Recent landscape genomics studies in forest trees—what can we believe?', *Tree Genetics & Genomes* **12**(1), 1–7.

Davila Olivas, N. H., Kruijer, W., Gort, G., Wijnen, C. L., van Loon, J. J. A. and Dicke, M. (2017), 'Genome-wide association analysis reveals distinct genetic architectures for single and combined stress responses in Arabidopsis thaliana', *New Phytol* **213**(2), 838–851.

Dawkins, R. (1982), *The Extended Phenotype: The Gene as the Unit of Selection*, Freeman, Oxford.

Doležel, J., Bartoš, J., Voglmayr, H. and Greilhuber, J. (2003), 'Letter to the editor', *Cytometry* **51A**(2), 127–128.

Eckert, A. J., Bower, A. D., Wegrzyn, J. L., Pande, B., Jermstad, K. D., Krutovsky, K. V., Clair, J. B. S. and Neale, D. B. (2009), 'Association Genetics of Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-Hardiness Related Traits', *Genetics* **182**(4), 1289–1302.

Elling, W., Dittmar, C., Pfaffelmoser, K. and Rötzer, T. (2009), 'Dendroecological assessment of the complex causes of decline and recovery of the growth of silver fir (Abies alba Mill.) in Southern Germany', *Forest Ecology and Management* **257**(4), 1175–1187.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowski, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., Prentice, I. C., Ramankutty, N. and Snyder, P. K. (2005), 'Global Consequences of Land Use', *Science* **309**(5734), 570–574.

Frankham, R. and Weber, K. (2000), Nature of quantitative genetic variation, *in* R. Singh and C. Krimbas, eds, 'Evolutionary Genetics: From Molecules to Morphology', Cambridge University Press, Cambridge, UK, pp. 351–368.

González-Martínez, S. C., Ersoz, E., Brown, G. R., Wheeler, N. C. and Neale, D. B. (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), 1915–1926.

González-Martínez, S. C., Huber, D., Ersoz, E., Davis, J. M. and Neale, D. B. (2008), 'Association genetics in Pinus taeda L. II. Carbon isotope discrimination', *Heredity* **101**(1), 19–26.

Gugerli, F., Brandl, R., Castagneyrol, B., Franc, A., Jactel, H., Koelewijn, H.-P., Martin, F., Peter, M., Pritsch, K., Schröder, H., Smulders, M. J. M., Kremer, A., Ziegenhagen, B. and Contributors, E. J. (2013), 'Community genetics in the time of next-generation molecular technologies', *Molecular Ecology* **22**(12), 3198–3207.

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. C. and Shyr, Y. (2013), 'Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data', *PLOS ONE* **8**(8), e71462.

Hannah, L., Carr, J. L. and Lankerani, A. (1995), 'Human disturbance and natural habitat: a biome level analysis of a global data set', *Biodiversity and conservation* **4**(2), 128–155.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O. and Townshend, J. R. G. (2013), 'High-Resolution Global Maps of 21st-Century Forest Cover Change', *Science* **342**(6160), 850–853.

Hobbs, R. J., Arico, S., Aronson, J., Baron, J. S., Bridgewater, P., Cramer, V. A., Epstein, P. R., Ewel, J. J., Klink, C. A., Lugo, A. E., Norton, D., Ojima, D., Richardson, D. M., Sanderson, E. W., Valladares, F., Vilà, M., Zamora, R. and Zobel, M. (2006), 'Novel ecosystems: theoretical and management aspects of the new ecological world order', *Global Ecology and Biogeography* **15**(1), 1–7.

Ingram, J. and Bartels, D. (1996), 'The molecular basis of dehydration tolerance in plants', *Annual review of plant biology* **47**(1), 377–403.

IPCC (2014), Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp., Technical report.

Kahl, G., Molina, C., Rotter, B., Jüngling, R., Frank, A., Krezdorn, N., Hoffmeier, K. and Winter, P. (2012), 'Reduced representation sequencing of plant stress transcriptomes', *J. Plant Biochem. Biotechnol.* **21**(1), 119–127.

Karnosky, D. F., Pregitzer, K. S., Zak, D. R., Kubiske, M. E., Hendrey, G. R., Weinstein, D., Nosal, M. and Percy, K. E. (2005), 'Scaling ozone responses of forest trees to the ecosystem level in a changing climate', *Plant, Cell & Environment* **28**(8), 965–981.

Korte, A. and Farlow, A. (2013), 'The advantages and limitations of trait analysis with GWAS: a review', *Plant Methods* **9**, 29.

Kovats, R., Valentini, R., Bouwer, L., Georgopoulou, E., Jacob, D., Martin, E., Rounsevell, M. and Soussana, J.-F. (2014), Europe. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Barros, V.R., C.B. Field, D.J. Dokken, M.D. Mastrandrea, K.J. Mach, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White (eds.)]., Technical report, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1267-1326.

Krajick, K. (2001), 'Defending Deadwood', *Science* **293**(5535), 1579–1581.

Krause, G. H. M., Arndt, U., Brandt, C. J., Bucher, J., Kenk, G. and Matzner, E. (1986), Forest Decline in Europe: Development and Possible Causes, *in* H. C. Martin, ed., 'Acidic Precipitation', Springer Netherlands, pp. 1701–1722. DOI: 10.1007/978-94-009-3385-9_171.

Kreps, J. A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X. and Harper, J. F. (2002), 'Transcriptome Changes for Arabidopsis in Response to Salt, Osmotic, and Cold Stress', *Plant Physiol.* **130**(4), 2129–2141.

Lander, E. S. and Schork, N. J. (1994), 'Genetic dissection of complex traits', *Science* **265**(5181), 2037–2048.

Levitt, J. (1980), *Responses of Plants to Environmental Stresses, Vol. 1*, Academic Press, New York, NY.

Libbrecht, M. W. and Noble, W. S. (2015), 'Machine learning applications in genetics and genomics', *Nat Rev Genet* **16**(6), 321–332.

Lichtenthaler, H. K. (1998), 'The Stress Concept in Plants: An Introduction', *Annals of the New York Academy of Sciences* **851**(1), 187–198.

Liepelt, S., Cheddadi, R., de Beaulieu, J.-L., Fady, B., Gömöry, D., Hussendörfer, E., Konnert, M., Litt, T., Longauer, R., Terhürne-Berson, R. and Ziegenhagen, B. (2009), 'Postglacial range expansion and its genetic imprints in Abies alba (Mill.) — A synthesis from palaeobotanic and genetic data', *Review of Palaeobotany and Palynology* **153**(1–2), 139–149.

Loarie, S. R., Duffy, P. B., Hamilton, H., Asner, G. P., Field, C. B. and Ackerly, D. D. (2009), 'The velocity of climate change', *Nature* **462**(7276), 1052–1055.

Mauricio, R. (2001), 'Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology', *Nat Rev Genet* **2**(5), 370–381.

Murray, B. G. (1998), 'Nuclear DNA Amounts in Gymnosperms', *Ann Bot* **82**(suppl 1), 3–15.

Neale, D. B. and Kremer, A. (2011), 'Forest tree genomics: growing resources and applications', *Nat Rev Genet* **12**(2), 111–122.

Neale, D. B. and Savolainen, O. (2004), 'Association genetics of complex traits in conifers', *Trends in Plant Science* **9**(7), 325–330.

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., Cardeno, C., Koriabine, M., Holtz-Morris, A. E., Liechty, J. D., Martínez-García, P. J., Vasquez-Gross, H. A., Lin, B. Y., Zieve, J. J., Dougherty, W. M., Fuentes-Soriano, S., Wu, L.-S., Gilbert, D., Marçais, G., Roberts, M., Holt, C., Yandell, M., Davis, J. M., Smith, K. E., Dean, J. F., Lorenz, W. W., Whetten, R. W., Sederoff, R., Wheeler, N., McGuire, P. E., Main, D., Loopstra, C. A., Mockaitis, K., deJong, P. J., Yorke, J. A., Salzberg, S. L. and Langley, C. H. (2014), 'Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies', *Genome Biology* **15**, R59.

Niinemets, Ü. (2010), 'Responses of forest trees to single and multiple environmental stresses from seedlings to mature plants: Past stress history, stress interactions, tolerance and acclimation', *Forest Ecology and Management* **260**(10), 1623–1639.

Nourtier, M., Chanzy, A., Cailleret, M., Yingge, X., Huc, R. and Davi, H. (2012), 'Transpiration of silver Fir (Abies alba mill.) during and after drought in relation to soil properties in a Mediterranean mountain area', *Annals of Forest Science* pp. 1–13.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., 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, Å., Rilakovic, N., Ritland, C., Rosselló, J. A., Sena, J., Svensson, T., Talavera-López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z.-Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T. R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Lee Thompson, S., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P. K., Lundeberg, J. and Jansson, S. (2013), 'The Norway spruce genome sequence and conifer genome evolution', *Nature* **497**(7451), 579–584.

Orr, H. A. (2005), 'The genetic theory of adaptation: a brief history', *Nature Reviews Genetics* **6**(2), 119–127.

Pflieger, S., Lefebvre, V. and Causse, M. (2001), 'The candidate gene approach in plant genetics: a review', *Molecular Breeding* **7**(4), 275–291.

Puizina, J., Sviben, T., Krajačić-Sokol, I., Zoldoš-Pećnik, V., Siljak-Yakovlev, S., Papeš, D. and Besendorfer, V. (2008), 'Cytogenetic and molecular characterization of the Abies alba genome and its relationship with other members of the Pinaceae', *Plant Biology* **10**(2), 256–267.

Rabbani, M. A., Maruyama, K., Abe, H., Khan, M. A., Katsura, K., Ito, Y., Yoshiwara, K., Seki, M., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2003), 'Monitoring Expression Profiles of Rice Genes under Cold, Drought, and High-Salinity Stresses and Abscisic Acid Application Using cDNA Microarray and RNA Gel-Blot Analyses', *Plant Physiol.* **133**(4), 1755–1767.

Rockman, M. V. and Kruglyak, L. (2006), 'Genetics of global gene expression', *Nat Rev Genet* **7**(11), 862–872.

Roschanski, A. M., Fady, B., Ziegenhagen, B. and Liepelt, S. (2013), 'Annotation and Re-Sequencing of Genes from De Novo Transcriptome Assembly of Abies alba (Pinaceae)', *Applications in Plant Sciences* **1**(1), 1200179.

Roth, R., Ebert, I. and Schmidt, J. (1997), 'Trisomy associated with loss of maturation capacity in a long-term embryogenic culture of Abies alba', *Theoretical and applied genetics* **95**(3), 353–358.

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P. and Neale, D. B. (2013), 'Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate', *Tree Genetics & Genomes* **9**(4), 901–911.

Stephens, M. and Balding, D. J. (2009), 'Bayesian statistical methods for genetic association studies', *Nat Rev Genet* **10**(10), 681–690.

Tabor, H. K., Risch, N. J. and Myers, R. M. (2002), 'Candidate-gene approaches for studying complex genetic traits: practical considerations', *Nat Rev Genet* **3**(5), 391–397.

Tautz, D. and Domazet-Lošo, T. (2011), 'The evolutionary origin of orphan genes', *Nat Rev Genet* **12**(10), 692–702.

Tinner, W., Colombaroli, D., Heiri, O., Henne, P. D., Steinacher, M., Untenecker, J., Vescovi, E., Allen, J. R. M., Carraro, G., Conedera, M., Joos, F., Lotter, A. F., Luterbacher, J., Samartin, S. and Valsecchi, V. (2013), 'The past ecology of Abies alba provides new perspectives on future responses of silver fir forests to global warming', *Ecological Monographs* **83**(4), 419–439.

Torre, A. R. D. L., Birol, I., Bousquet, J., Ingvarsson, P. K., Jansson, S., Jones, S. J. M., Keeling, C. I., MacKay, J., Nilsson, O., Ritland, K., Street, N., Yanchuk, A., Zerbe, P. and Bohlmann, J. (2014), 'Insights into Conifer Giga-Genomes', *Plant Physiol.* **166**(4), 1724–1732.

Van Nuland, M. E., Wooliver, R. C., Pfennigwerth, A. A., Read, Q. D., Ware, I. M., Mueller, L., Fordyce, J. A., Schweitzer, J. A. and Bailey, J. K. (2016), 'Plant–soil feedbacks: connecting ecosystem ecology and evolution', *Funct Ecol* **30**(7), 1032–1042.

Wang, Z., Gerstein, M. and Snyder, M. (2009), 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature reviews genetics* **10**(1), 57–63.

Whitham, T. G., Bailey, J. K., Schweitzer, J. A., Shuster, S. M., Bangert, R. K., LeRoy, C. J., Lonsdorf, E. V., Allan, G. J., DiFazio, S. P., Potts, B. M., Fischer, D. G., Gehring, C. A., Lindroth, R. L., Marks, J. C., Hart, S. C., Wimp, G. M. and Wooley, S. C. (2006), 'A framework for community and ecosystem genetics: from genes to ecosystems', *Nat Rev Genet* **7**(7), 510–523.

Wolf, H. (2003), Silver fir - Abies alba, *in* 'EUFORGEN Technical Guidelines for genetic conservation and use for silver fir (Abies alba)', International Plant Genetic Resource Institute, Rome, Italy.

Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marçais, G., Puiu, D., Roberts, M., Wegrzyn, J. L., Jong, P. J. d., Neale, D. B., Salzberg, S. L., Yorke, J. A. and Langley, C. H. (2014), 'Sequencing and Assembly of the 22-Gb Loblolly Pine Genome', *Genetics* **196**(3), 875–890.

Zimmermann, H. (2014), Differential expression of candidate genes for drought stress response in Abies alba Mill. seedlings, Master thesis, Philipps-Universität Marburg, Marburg.

Danksagung – Acknowledgments

Ich danke Norman Born für die sehr gute und erfolgreiche Zusammenarbeit, Ralf Gente für sein offenes Ohr für physikalische Fragen und seine Geduld diese zu beantworten, Heike Zimmermann für ihre eiserne Arbeitsmoral und die unterhaltsame Laborarbeit, Christina Mengel für ihre vielen guten Ratschläge im Labor, Felix Staeps für unzählige clevere und lustige Diskussionen und Franziska Willems für ihr sonniges Gemüt und ihre ansteckende Lebensfreude.

Vielen Dank auch an Nina Farwig für die Zweitkorrektur dieser Dissertation und Michael Bölker und Gerhard Kost für ihre Bereitschaft wertvolle Zeit zu opfern um Mitglieder in der Prüfungskommission zu sein.

Mein besonderer Dank geht an Katrin Heer für ihren analytischen Geist, ihre endlose Geduld und ihre Fähigkeit immer im richtigen Moment weg zu schauen wenn man Studentenfutter klaut, Lars Opgenoorth für seinen ungezügelten Enthusiasmus und seine transzendentalen Tanzkünste, Sascha Liepelt für seine immer offene Tür, seinen Blick fürs Wesentliche und seine etlichen wertvollen Kommentare und Ideen und Birgit Ziegenhagen für die vielen spannenden und anregenden Diskussionen und dafür, dass sie eine Atmosphäre in ihrer Arbeitsgruppe geschaffen hat in der man gerne arbeitet und die jeder, wenn er weg ist, vermisst.

Zuletzt möchte ich meiner Familie danken, meiner Schwester Lisa Kerscher und Stéphane Bölingen und ganz besonders meinen Eltern Jutta und Jürgen Behringer, ohne die ich nicht hätte studieren und ebensowenig promovieren können und ohne deren Unterstützung ich in vielerlei Hinsicht heute nicht hier wäre.

Appendix

Data availability

All data for the published articles and unpublished manuscripts for this thesis can be found in public repositories and on an attached data CD which is structured and labeled according to the outline of the thesis by chapter name as follows:

Chapter 2 PDF document of the published paper, the supplemental data (one movie and one DOC file) and the Master thesis of David Behringer (2013).

Chapter 3 CSV data and R-script to create all figures.

Chapter 4 PDF document of the published paper, the master thesis of Heike Zimmermann (2014). All supporting information can be freely downloaded here:

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124564.

FASTQ files from the two Illumina-based MACE libraries were deposited in the SRA (Short Read Archive, NCBI) with the following accession: PRJNA266095.

Chapter 5 Sub-folder 'SNPsCleaning': CSV files of the raw SNP data (AbiesSNPsRawData.csv), R-script (SNPsCleaning.R) with all necessary custom R-functions to clean the SNPs as described in the manuscript, the results from the linkage disequilibrium analysis (SNPsCleanedLDResults.txt) and the cleaned SNPs in different file formats.

Sub-folder 'SNPsPCA': R-script (SNPsPCA.R) to re-run the analysis with the necessary custom R-function (toLFMM.R) to convert the SNP data into a binary format based on minor allele frequency.

Sub-folder 'TRW': CSV files for the tree-ring width data for both sites including original data for all cores (TRW.High/Low.csv) and detrendend mean data (TRW.mean.growth.High/Low.detrended.csv).

Sub-folder 'Dendro': Dendrophenotypes, unscaled (Dendro.csv) and scaled (Dendro.sc.csv) and the temperature data for both sites in 2014 and 2015 (Tempsums.2014.2015.csv).

Sub-folder 'Tassel': CSV files for the results of the TASSEL GLM analyses with estimates (Tassel.GLM.geno.csv) and statistics (Tassel.GLM.stats.csv).

Chapter 6 CSV files for the Sitka spruce (SNPs and phenotypes: TableS3.csv) and the silver fir (SNPs: SNPsSilverFir.csv, phenotypes: DendroScaled.csv) data and R-scripts (SitkaSpruce.R and SilverFir.R) and necessary custom R-functions (in the subfolder 'RFunctions') to reproduce the analyses.

Erklärung zum Eigenanteil

Die Publikationen aus Chapter 2 und Chapter 4 dieser Dissertation sind aus meiner Masterbarbeit mit dem Titel

„Identifying candidate genes for drought stress response in *Abies alba* seedlings using terahertz time-domain spectroscopy and gene expression profiling“

die ich am 15.07.2013 im Studiengang *Organismische Biologie* an der Philipps-Universität Marburg eingereicht habe, hervorgegangen (die Arbeit findet sich als PDF auf der beigelegten Daten-CD im Ordner 'Chapter 2').

Die Publikation aus Chapter 2 ist die in weiten Teilen überarbeitete und teilweise veränderte Version des erstmals am 17.05.2013 bei dem Journal *Plant Physiology* eingereichten und abgelehnten Manuskripts, das als solches Teil meiner Masterarbeit ist.

Die Publikation aus Chapter 4 basiert teilweise auf den Ergebnissen meiner Masterarbeit und der Masterarbeit von Heike Zimmermann die sie am 14.02.2014 im Studiengang *Organismische Biologie* an der Philipps-Universität Marburg mit dem Titel

„Differential expression of candidate genes for drought stress response in *Abies alba* Mill. seedlings“

eingereicht hat (die Arbeit findet sich als PDF auf der beigelegten Daten-CD im Ordner 'Chapter 4').

Es folgt eine Angabe des Eigenanteils an allen Publikationen und Manuskripten die Teil dieser Dissertation sind:

	Design der Experimente	Durchführung der Experimente	Datenauswertung und Interpretation	Hauptverfasser des Textes *
Chapter 2	**DB**, NB, SL	**DB**, NB	**DB**, NB	**DB**, (NB)
Chapter 3	**DB**	**DB**, MS	**DB**	**DB**
Chapter 4	**DB**, HZ, SL	**DB**, HZ	**DB**, HZ	**DB**, (HZ)
Chapter 5	DB nicht beteiligt	DB nicht beteiligt	AP, **DB**, KH, LO	**DB**, KH, LO
Chapter 6	**DB**	**DB**	**DB**	**DB**

AP Alma Piermattei, BZ Birgit Ziegenhagen, **DB David Behringer**, KH Katrin Heer, HZ Heike Zimmermann, LO Lars Opgenoorth, MS Michael Schwerdtfeger, NB Norman Born, SL Sascha Liepelt

* in Klammern sind Ko-Autoren erwähnt, die größere Textbausteine - meist im Methodenteil - beigesteuert haben, die von DB verändert und in den Text eingepasst wurden

Ort, Datum

Unterschrift (David Behringer)

Ort, Datum

Unterschrift (Birgit Ziegenhagen)

Eidesstattliche Versicherung

Ich versichere hiermit, dass ich die vorliegende Dissertation mit dem Titel

„Candidate genes for stress response in silver fir (*Abies alba* Mill.)“

selbst und ohne fremde Hilfe verfasst habe, nicht andere als die in ihr angegebenen Quellen oder Hilfsmittel benutzt habe und das alle vollständig oder sinngemäß übernommenen Zitate als solche gekennzeichnet sind.
Die Dissertation wurde in der vorliegenden oder einer ähnlichen Form noch bei keiner anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuchs oder zu anderen Prüfungszwecken eingereicht.

Ort, Datum

Unterschrift (David Behringer)