Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

2013•2014
FACULTY OF SCIENCES
Master of Statistics: Bioinformatics

Masterproef

Identification of cold-hardiness related genes in coastal Douglas fir (*Pseudotsuga menziesii*)

Promotor :
Prof. dr. Tomasz BURZYKOWSKI

Promotor :
Prof.dr. FREDERIK HENDRICKX

Carl Vangestel

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics

Maastricht University

2013•2014

FACULTY OF SCIENCES

Master of Statistics: Bioinformatics

Masterproef

Identification of cold-hardiness related genes in coastal Douglas fir (*Pseudotsuga menziesii*)

Promotor :
Prof. dr. Tomasz BURZYKOWSKI

Promotor :
Prof.dr. FREDERIK HENDRICKX

Carl Vangestel

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics

Table of Contents

1. ABSTRACT

Extreme low temperatures can induce substantial stress in trees and may infer increased mortality and yield loss. Populations of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*) show marked differences in resilience to freezing temperatures and such selective pressures determine to a large extent the natural distribution of this species.

We here apply a candidate gene-based approach by screening 228 single-nucleotide polymorphisms (SNPs) located in 117 genes along cold-hardiness related environmental gradients. Genes have been selected *a priori* based on existing databases and relevant metabolic pathways in other species. In total our environmental data encapsulated 28 climate and 3 geographical variables. Climate data appeared to be highly correlated as a single principal component could retain as much as 82% of the total variance. Microsatellite and allozyme data revealed no strong population structure. To explore patterns of adaptive variation along these gradients we used three different approaches: a Bayesian outlier detection method, a multinomial logistic regression and a Bayesian environmental analysis. These methods differ in the extent to which they implement information on genetic population structure and the environment. A total of 33 genotype-environment associations could be detected in 24 distinct genes. While being more conservative, results of the outlier analysis corroborated those of the logistic regression. In contrast, results of the Bayesian environmental analysis were incongruent with those of the regression model and identified almost exclusively a non-overlapping set of putative selective genes. In addition, we were able to highlight three genes that showed consistent relationships between pairwise components of the phenotype-genotype-environment spectrum.

Knowledge on adaptive genetic variation will allow forest managers to have access to improved genetic substantiated reforestation guidelines. This may help mitigate the impact of climate change such as induced shifts in the spatial distribution of coastal Douglas fir and may assist in preserving the evolutionary potential of this economically an ecologically valuable species.

2. INTRODUCTION

Natural selection has been a central theme in evolutionary studies for decades. It is the evolutionary force by which the distribution of trait values within a population can change over generations through differential survival or birth rates. One of the premises underlying this process is the presence of phenotypic variation within a population that relates to fitness i.e. some phenotypes will have lower mortality and/or reproduce more. Additionally, this phenotypic variation needs to be heritable such that beneficial trait characteristics can be passed from parents to offspring. The evolution of trait characteristics that enhance fitness are also called adaptations (Fox and Wolf 2006).

Current predictive models on global climate change have fuelled research in forest genetics as future climatological conditions may deviate strongly from contemporary ones potentially creating local patterns of maladaptations in sessile organisms such as trees. Such environmental shifts will pose serious challenges to forest managers and unveiling adaptive genetic variation in forest trees will be key to maintain productivity and sustainability in the face of these environmental changes (González-Martínez et al. 2006, Neale and Kremer 2011). A detailed knowledge of adaptive genetic variation in natural populations will allow conservation managers to preserve the adaptive genetic diversity and hence evolutionary potential of species and to provide reforestation guidelines for assisted breeding schemes (Neale and Kremer 2011).

The advent of next-generation sequencing (NGS) technology finally provided researchers a tool to address some of the long-standing questions about the molecular basis of local adaptation by allowing them to focus on adaptive rather than neutral genetic variation in natural populations. Temperature and water are among the strongest abiotic selective forces and environmental constraints in trees. Geography creates strong gradients in climatic conditions (extreme temperatures, water availability) and as such can infer clinal variation in adaptive traits and gene frequencies controlling these phenotypes. Although their molecular basis remains yet largely unresolved, forest trees show marked and well supported phenotypic signatures of local adaptation to various environmental factors (Neale and Savolainen 2004). One constraint to performing genome-wide associations studies (GWAS) in trees is their extreme large genome size (up to 160 times larger than the model plant organism *Arabidopsis*) and the small extent of linkage disequilibrium (only several hundreds of base pairs long) such that millions of SNPs would be needed to cover the entire genome. A more feasible alternative is to explore the frequency spectrum of SNPs in targeted genes of interest (Neale and Savolainen 2004). These latter can be identified by screening existing databases such as expressed sequence tags (EST) repositories that highlight genes with known or predicted function (González-Martínez et al. 2006).

Throughout the literature many statistical methods have been described aiming to distinguish selective regions in the genome from neutral ones. Variation in the latter is mainly driven by gene flow (i.e. exchange of genes between populations) and genetic drift (i.e. a random change of allele frequencies across generations due to 'sampling error') and affects all loci simultaneously. In contrast, natural selection acts only upon a single or a few loci, namely those that control the expression of the target phenotype, while the rest of the genome is left unaffected. Positive selection, i.e. a type of selection that increases the frequency of new advantageous mutations, causes elevated levels of genetic differentiation at these loci and as such leaves a distinct and local signature of selection in the genome. Although there are a suit of different statistical approaches to infer selective regions, most of them, in some way, make use of this local deviating pattern of genetic differentiation (Nielsen 2005). Environmental factors often constitute strong selective forces promoting the most adapted phenotype. Therefore researchers nowadays have started to implement their genome-wide scans in a more explicit environmental context by linking the spatial distribution of allele frequencies to environmental gradients, thereby evolving towards a more causative and functional approach of exploring adaptive genetic variation in natural populations (Joost et al. 2007, Schoville et al. 2012).

One of the most ecologically and economically valuable trees in western North America is Douglas fir. Its natural range covers the western part of south Canada to Mexico, from the Pacific coast to the eastern Rocky Mountains. Two varieties are recognized: coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*) which covers the coastal region from California to British Columbia, and the interior Douglas fir (*Pseudotsuga menziesii* var. *glauca*) which can be found more inland in mountainous habitats (St Clair et al. 2005). The range of coastal Douglas fir include a wide range of habitats that differ in both temperature and moisture (St Clair 2006). Especially temperature appears to be one of the strongest abiotic selective forces and environmental constraints in this species as populations along these gradients show marked differences in susceptibility to cold damage (St Clair et al. 2005). Cold-hardiness refers to the ability of a tree to adapt to freezing temperatures and to prevent tissue damage. Two non-mutually exclusive mechanisms to adapt to extreme minimum temperatures are avoidance (prevent freezing) and tolerance (survive freezing). The first comprises for example shifts in life cycli while the latter may refer to the formation of cryoprotective compounds and cold-inducible proteins (Gusta and Wisniewski 2013). Cold-hardiness is assumed to be a quantitative trait which is controlled by many genes each having a small effect rather than by a single, large effect gene.

Quantitative trait values are determined by the interplay of two major components: the environment and genotype. Observed phenotypic variation within a population is the result of environmental and genetic differences between individuals. This is summarized by the well known mathematical equation $P = G+E$, where P represents the phenotypic variance, E the environmental variance and the G the genetic one (Fig. 1). Strictly spoken we can further expand this model by replacing G by its addiditve, dominant and epistatic component and including a G x E interaction (Allendorf et al. 2012).

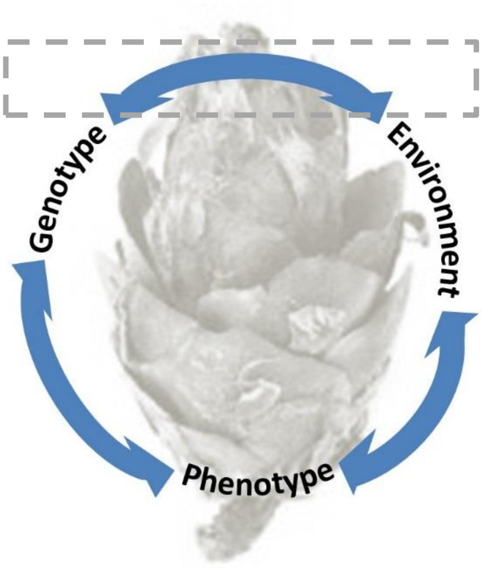

Fig. 1. The observed phenotypes within a population are determined by their genetic constitution, the environment they live in and the interplay between the latter two. This study mainly focusses on the association between genotype and environment (dashed line).

While numerous studies report associations between genotypes and phenotypes, environment and phenotypes or environment and genotypes there's a paucity of studies integrating all these associations simultaneously. Such holistic approach could however provide a powerful methodology to delineate and validate the functional role of genes underlying adaptive traits. The long-term study on coastal Douglas fir provides a unique opportunity to illustrate how combining all components of the phenotype-genotype-environment spectrum may enhance our understanding of genes underlying cold-hardiness associated traits. Genecological studies of Douglas fir using a multi-year common garden approach have addressed the geographical, topographical and climatological patterns in a set of a priori selected cold-hardiness related traits (St Clair et al. 2005, St Clair 2006). The main results indicated winter temperatures and frost dates as key features to local adaptation of Douglas fir throughout its natural range ('phenotype-environment' associations). By complementing earlier common garden experiments on quantitative trait loci (QTL) (Jermstad et al. 2001a, Jermstad et al. 2001b, Jermstad et al. 2003, Wheeler et al. 2005) and using the exact same phenotypic data as the genecological study, Eckert et al. (2009a) were able to highlight a set of 12 genes showing significant associations with 10 cold-hardiness related traits in an association mapping study ('phenotype-genotype' associations). The gene effects were typically small (0.019% - 0.036% variance explained) suggesting cold-hardiness related traits are indeed complex and polygenic.

In this study we aim to expand our current view on the adaptive genetic variation of cold-hardiness related traits in coastal Douglas fir. We apply a candidate-gene based approach by *a priori* selecting SNPs within target genes and associating their allele frequencies with environmental gradients related to cold stress. Three methodologies are contrasted which differ in the extent to which they include genetic background variation and/or environmental data. Finally, we try to integrate these novel results into existing data on respectively genotype-phenotype and environment-phenotype associations, hence providing us a rare opportunity to monitor adaptive variation along the entire genotype-phenotype-environment spectrum of a species.

3. MATERIAL AND METHODS

3.1. Sample collection

A sample of 643 coastal Douglas fir trees in Oregon and Washington was selected from previously published datasets on genecology and association mapping of Douglas fir (St Clair et al. 2005, St Clair 2006, Eckert et al. 2009a, Eckert et al. 2009b). Samples were distributed across the natural range of Douglas fir which covers a suit of environmental gradients. Trees are subdivided in different classes according to the ecoregion they belong to (we refer to St Clair et al. (2006) for a detailed description of the level III ecoregions) and these regions were subsequently further stratified using latitudinal and altitudinal thresholds eventually resulting in 18 distinct regions (Fig. 2).

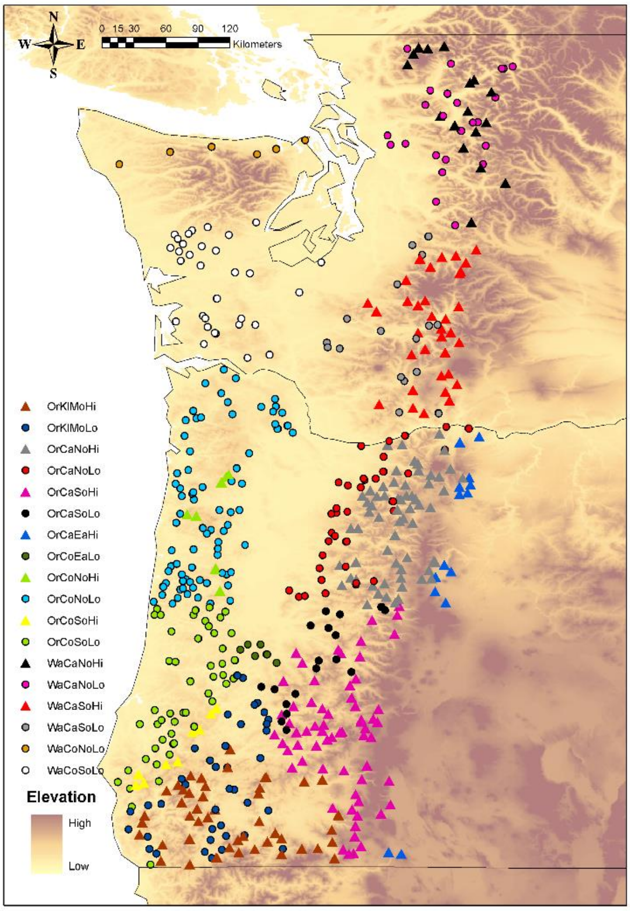

Fig. 2. Source locations of 643 used Douglas fir trees. Samples were assigned to 18 distinct classes based on ecoregion, latitude and elevation (OrKlMo = Klamath Mountains, OrCaNo = Oregon Cascades North, OrCaSo = Oregon Cascades South, OrCaEa = Oregon Cascades East, OrCoNo = Oregon Coast North, OrCoSo = Oregon Coast South, WaCaNo = Washington Cascades North, WaCaSo = Washington Cascades South, WaCoNo = Washington Coast North, WaCoSo = Washington Coast South, OrCoEa = Willamette Valley. 'Hi' or 'Lo' at end of the label refers to respectively the high (above 650 m - triangles) or low (below 650m - circles) stratum). Elevation data was downloaded from the WorldClim database (Hijmans et al. 2005).

3.2. Environmental data

For each sampled tree we collected both ‘geographical’ (latitude, longitude and elevation) and ‘climate’ data. Elevation adjusted climate data was obtained from the gridded 30-second arc ClimateWNA database using the reference period 1961-1990 (Wang et al. 2012). A total of 28 climate variables were assessed. Annual and seasonal climate data encapsulated mean coldest temperature (Tmin), amount of snow (PAS), number of frost-free days (NFFD), degree-days below 0°C (DD0) and degree-days above 5°C (DD5). Annual variables further comprised data on differences between coldest and warmest month (‘continentality’) (TD), begin (bFFP) and end (eFFP) of frost-free period, duration of frost-free period (FFP) and the extreme minimum temperature over 30 years (EMT). Correlations between climate variables were graphically explored using the software package ‘Corrgram’ (Friendly 2002) in the R environment (R Core Team 2013). A subsequent principal component analysis (PCA) was performed to reduce the high dimensionality of the climatic data using the PRINCOMP procedure in SAS 9.3. (SAS Institute 2008, Cary, NC, USA). PCA’s characterized by eigenvalues larger than 1 were retained and correlations between original variables and the 2 most important PCA axes were visualized using biplots. To assess the interaction between geography and climate we calculated Pearson correlation coefficients between elevation, latitude, longitude and all retained PCA axes.

3.3. Candidate gene selection and SNP genotyping

SNP data was obtained from a previous phenotype-genotype association study (Eckert et al. 2009a, Eckert et al. 2009b). In brief, candidate genes putatively involved in cold-tolerance mechanisms were *a priori* selected based on a threefold criterion: i) collocation of genes with QTLs for cold-hardiness (Wheeler et al. 2005), ii) genes involved in physiological metabolic pathways of a cold tolerance response and iii) genes differentially expressed in microarray studies in *A. thaliana* (Lee et al. 2005). Available Douglas fir expressed sequence tag (EST) libraries were screened using a standard BLAST algorithm and a lower limit of 1E-10 was set as a threshold to call putative orthologs. Of the original 384 SNPs that were selected 228 SNPs that met the predefined quality thresholds (0.35 and 0.85 for respectively the indices GenCall$_{50}$ and call rate) were retained. These SNPs were located in 117 distinct genes. Genotyping was carried out on an Illumina GoldenGate genotyping platform at the DNA Technologies Core Facility located at the Genome Center of the University of California, Davis. Gene annotations are described in Eckert et al. (2009a).

3.4. Patterns in neutral genetic variation

The contemporary spatial distribution of neutral alleles, i.e. alleles not affected by evolutionary processes like natural selection, is mainly determined by the interplay between two opposing demographic processes: gene flow and genetic drift (Allendorf et al. 2012). The former tends to homogenize allele frequencies across demes, while the latter indulges an increase in genetic dissimilarity among demes. As not all members of a contemporary gene pool will produce offspring the allelic constitution of future gene pools will deviate from current ones and this effect will be much stronger in small populations. A common metric used to quantify the extent of differences in allele frequencies across spatially separated populations is Wright's F_{ST}-statistic or derivatives thereof (Wright 1951, Balloux and Lugon-Moulin 2002). In his original paper Wright (1951) defined F_{ST} as "the correlation between two alleles chosen at random within subpopulations relative to alleles sampled at random from the total population". We opted to use G_{ST} (Nei 1977, Meirmans and Hedrick 2011), the multi-allelic extension of the traditional F_{ST}, to assess the neutral population structure using a set of 25 allozymes and 6 microsatellites (data obtained from Krutovsky et al. (2009)). $G_{ST} = \frac{cH_T - cH_S}{cH_T}$, where cH_T and cH_S are respectively the corrected expected heterozygosity if all populations were pooled and the corrected average within-population expected heterozygosity. Large positive values indicate strong genetic differentiation or population subdivision. In contrast to the aforementioned trait-based selected SNPs, these markers are believed to better represent the neutral population structure (Selkoe and Toonen 2006). Pairwise G_{st} indices were calculated using GenAlEx 6.501 (Peakall and Smouse 2006) and visualized using the R package heatmap.plus.

We further explored whether genetic similarity between individuals decreased with geographical distance ("isolation-by-distance") by measuring the correspondence between two distance matrices. The first matrix comprised pairwise measures of genetic distance between all individuals where genetic distance for a single di-allelic locus was defined as the squared distance d^2 and calculated as $d^2(ii,ii)=0$, $d^2(ij,ij)=0$, $d^2(ii,ij)=1$, $d^2(ij,ik)=1$, $d^2(ij,kl)=2$, $d^2(ii,jk)=3$ or $d^2(ii,jj)=4$ for the i^{th}, j^{th}, k^{th} and l^{th} different alleles. The total pairwise genetic distance between two individuals is obtained by summing over all loci. The second matrix contained all pairwise geographical distances. Correspondence between the two matrices was quantified by means of a Mantel test (Mantel 1967) conducted in GenAlEx 6.501 (Peakall and Smouse 2006) and significance was assessed using 999 permutations.

3.5. Characterization of adaptive genetic variation

Adaptive genetic variation across an environmental gradient can be identified through loci that deviate strongly from patterns of neutral genetic differentiation or through a direct association of allele frequencies and environment. We here apply three analytical methods i) an outlier-detection method that screens for aberrant variation in neutral genetic differentiation but disregards the environmental variation, ii) a multinomial logistic regression that links allele frequencies to environmental variables but neglects expectations of neutral genetic differentiation and iii) a Bayesian linear model that associates allele frequencies to environmental variables, while taking patterns of neutral genetic variation into account.

3.5.1. Bayesian outlier detection analysis

According to the neutral theory of molecular evolution (Kimura 1983) genetic drift and gene flow are expected to affect all loci simultaneously (although mutation rates may differ across genomic regions their effects are often neglegiable compared to demographic processes at an ‘ecological time scale’). In contrast, local adaptation will only strongly increase (positive directional selection) or decrease (balancing selection) levels of differentiation among populations in those loci subjected to natural selection. Traditional outlier detection methods make use of these theoretical properties by evaluating observed F_{ST} values across a null distribution of F_{ST} indices (i.e. in the absence of natural selection). Loci characterized by observed F_{ST} values at the extreme ends of this ‘null envelop’ are hence interpreted as potential targets of natural selection. Many authors however have urged caution as outcomes can be flawed under several demographic scenarios such as the presence of a small number of strongly isolated demes or a history of recent population expansions. Under these conditions increased levels of false positives have been observed (Beaumont 2005, Schoville et al. 2012). We here use a novel Bayesian approach (Foll and Gaggiotti 2008) that incorporates locus- and population-specific F_{ST} effects and appears to be more robust to deviations from underlying demographical assumptions (Foll and Gaggiotti 2008, Pérez-Figueroa et al. 2010, Narum and Hess 2011). The rationale behind this approach (Foll and Gaggiotti 2008) is that all current allele frequencies of *J* populations originate from an ancestral frequency. As selection regimes may differ among loci and effective population sizes, and consequently genetic drift, and migration rates may be very heterogeneous among populations, contemporary allele frequencies are allowed to deviate from ancestral ones in a population- and locus-specific manner. Denote the ancestral allele frequency of allele *k* at locus *i* as p_{ik} (for *K* number of alleles at the *i*th locus), the current allele frequecies at locus *i* and subpopulation *j* as $\widetilde{\boldsymbol{p}_{\imath\jmath}}$ = {$\widetilde{p_{\imath\jmath 1}}$, $\widetilde{p_{\imath\jmath 2}}$, …, $\widetilde{p_{\imath\jmath K\imath}}$}, the number of alleles *k* at locus *i* and population *j* as $\mathbf{a_{ij}}$ = {a_{ij1}, a_{ij2}, …, a_{ijKi}}, and the sample size at locus *i* in population *j* as n_{ij} $= \sum_k a_{ijk}$. Then, the allele frequencies at locus *i* and subpopulation *j* follow a Dirichlet distribution with $\widetilde{\boldsymbol{p}_{\imath\jmath}}$ ~ Dir ($\theta_{ij}p_{i1}$, …, $\theta_{ij}p_{iKi}$) where θ_{ij} = $1/F_{ST}^{ij}$ -1. The observed allele frequencies are considered to be samples from the true allele frequency $\widetilde{\boldsymbol{p}_{\imath\jmath}}$ and therefore outcomes of a multinomial distribution, $\mathbf{a_{ij}}$ ~ Multinomial (n_{ij}; $\widetilde{p_{\imath\jmath 1}}$, $\widetilde{p_{\imath\jmath 2}}$, …, $\widetilde{p_{\imath\jmath K\imath}}$). As ancestral allele frequencies are unknown the authors chose a noninformative Dirichlet prior to obtain their estimates, $\mathbf{p_i}$ ~Dir (1, …, 1). Locus- and population-specific effects θ_{ij} are modelled using a logistic regression where the logits of genetic differentiation between the ancestral gene pool and population *j* at locus *i*, $F_{ST}^{(ij)}$, is decomposed into a population-specific component β_j (shared by all loci) and a locus-specific component α_i (shared by all populations): $\log\left(\frac{F_{ST}^{ij}}{1-F_{ST}^{ij}}\right) = \alpha_i + \beta_j$. The values F_{ST}^{ij} have a similar interpretation as those outlined by Wright (1951). We make use of the fact that the Dirichlet distribution is a conjugate prior of the multinomial distribution to eliminate the (nuisance) parameters $\widetilde{\boldsymbol{p}_{\imath\jmath}}$ and to obtain a multinomial-Dirichlet distribution:

P($\mathbf{a_{ij}}$ | $\mathbf{p_i}$, α_i, β_j) = $\frac{n_{ij}!\,\Gamma(\theta_{ij})}{\Gamma(n_{ij}+\theta_{ij})}\prod_{k=1}^{K_i}\frac{\Gamma(a_{ijk}+\theta_{ij}p_{ik})}{a_{ijk}!\,\Gamma(\theta_{ij}p_{ik})}$. By multiplying across all loci and populations we obtain the likelihood L($\mathbf{p}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) = $\prod_{i=1}^{I}\prod_{j=1}^{J}P(\boldsymbol{a_{ij}}|\,\boldsymbol{p_i},\alpha_i,\beta_j)$. To identify loci under selection two alternative models were defined, one that includes α_i (with selection) and one that excludes it (no selection). Bayes' factors (BF), i.e. the ratio of the likelihoods of respectively a model with and without selection (Lesaffre and Lawson 2012), were interpreted according to Jeffreys' scale (Table 1) (Jeffreys 1961).

Table 1. Jeffreys' scale of evidence for the Bayes factors.

BF	$\log_{10}$ BF	Interpretation
1 < BF < 3	0 < $\log_{10}$ BF < 0.48	Barely worth mentioning
3 < BF < 10	0.48 < $\log_{10}$ BF < 1	Substantial
10 < BF < 32	1 < $\log_{10}$ BF< 1.5	Strong
32 < BF < 100	1.5 < $\log_{10}$ BF < 2	Very strong
100 < BF < ∞	2 < $\log_{10}$ BF < ∞	Decisive

Analyses were conducted using the software package BayeScan (Foll and Gaggiotti 2008). Sample size was set to 5000 and the thinning interval to 10 with a burn-in of 50000 iterations.

3.5.2. Multinomial logistic regression

Although numerous studies have demonstrated the value of outlier detection methods when attempting to discover key genomic regions involved in the adaptation process, they remain ignorant about the specific (environmental) selection pressures. Joost et al. (2007) directly associated allelic frequencies with environmental variation by fitting binomial logistic regressions for each locus-environmental variable pair. We here apply a slightly modified method by modelling genotypic rather than allelic frequencies, hence replacing the binomial logistic regression by a multinomial one (when more than 2 genotypic classes are available within the sampled region), namely a generalized logits model. Multinomial logistic regression models were fitted using the LOGISTIC procedure in SAS 9.3. (SAS Institute 2008, Cary, NC, USA). Environmental variables included all three geographical variables and the most important principal components (i.e. eigenvalues > 1). To account for the multitude of tests being conducted we specified the false discovery rate at 0.01 using the QVALUE package in R (Storey 2002, Storey and Tibshirani 2003).

3.5.3. Bayesian environmental analysis

One caveat to environmental association analyses such as those outlined above is their incapacity to exclude potential confounding effects of geography and demography. Environmental gradients often co-occur with geographical ones (i.e. latitude, longitude, elevation, etc.). In addition, species characterized by an 'isolation-by-distance' dispersal mode, i.e. the rate of gene flow is a decreasing function of distance, tend to synchronize allele frequencies among neighboring populations resulting in a gradual change of neutral allele frequencies across the landscape. As a consequence, spatial patterns of allele frequencies along an environmental cline at a neutral locus may emerge (Fig. 3) and may hence falsely be assigned as an adaptive genomic region (Novembre and Di Rienzo 2009).

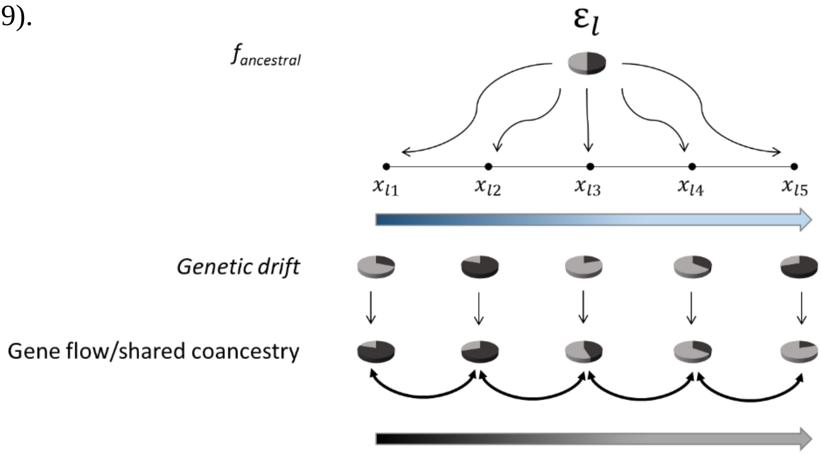

Fig. 3. Correlation between allele frequencies and environmental variables at neutral loci. Consider a single di-allelic locus *l* with ancestral frequencies respectively ε_l and 1- ε_l. Assume, for simplicity, 5 populations have been sampled along a linear environmental gradient (blue arrow). Allele frequencies in each population, x_{li} (i=1,…5), will randomly drift away from the ancestral frequency due to finite sample sizes. High gene flow between neighbouring populations (curved arrows) however will counter drift effects and tend to homogenize allele frequencies in adjacent populations. As a result allele frequencies follow a linear gradient in correspondence to the environmental one (grey arrow).

We here apply a Bayesian linear model that links allele frequencies to environmental variation while correcting for the neutral background pattern of allele frequencies (Coop et al. 2010). First, we estimate a null model in which allele frequencies may spatially covary but are not yet associated with environmental variation. Denote *L* independent loci in *K* populations, n_{kl} and m_{kl} the observed counts of respectively the first and second allele (only di-allelic SNPs are considered) at locus *l* in population *k* and x_{kl} the unknown population frequency of the first allele (arbitrarily chosen). We can then assume that the observed counts of each allele are drawn from a binomial distribution with parameter x_{kl}. Following the same philosophy as the approach outlined in section 3.5.1., these unknown population frequencies originate from an ancestral or global frequency, ε_l, from which they deviate through genetic drift while different population-specific deviations may covary through gene flow or shared coancestry. By circumventing the constraint that these frequencies are between 0 and 1 we assume that x_{kl} is normally distributed around an ancestral allele frequency ε_l ('truncated Gaussian distribution') (Nicholson et al. 2002) and that densities of x_{kl} outside the [0,1] interval are replaced by point masses 0 or 1. We further assume that the variance of this normal distribution equals the product of a population-specific constant, C_k, and a locus-specific term $\varepsilon_l(1-\varepsilon_l)$. This constant C_k can be viewed as a population-specific estimate of F_{ST}. Hence the unknown population frequency x_{kl} is linked by a transform g() to a surrogate population allele frequency θ_{kl} which is not constrained to the interval [0,1]:

$$x_{kl} = g(\theta_{kl}) = \begin{cases} 0 & \text{if } \theta_{kl} < 0 \text{ (probability that allele has been lost)} \\ \theta_{kl} & \text{if } 0 \leq \theta_{kl} \leq 1 \\ 1 & \text{if } \theta_{kl} > 1 \text{ (probability that allele has been fixed)} \end{cases}$$

Hence, for each locus *l* θ_{kl} has a marginal distribution ~ $N(\varepsilon_l, \varepsilon_l(1\text{-}\varepsilon_l)C_k)$. To model the joint distribution of allelic frequencies across all populations we assume the θ_l= $\theta_{1l,\ldots,}\theta_{Kl}$ has a multivariate normal distribution $P(\theta_l \mid \Omega, \varepsilon_l) \sim N(\varepsilon_l, \varepsilon_l(1\text{-}\varepsilon_l)\ \Omega)$, with variance-covariance matrix Ω. The joint posterior of θ_l, Ω and ε_l at locus *l* is then given by $P(\theta_l, \Omega, \varepsilon_l \mid \mathbf{n_l}, \mathbf{m_l}) \propto P(\mathbf{n_l}, \mathbf{m_l} \mid x_l = g(\theta_l))\ P(\theta_l \mid \Omega, \varepsilon_l)\ P(\Omega)\ P(\varepsilon_l)$. As the variance-covariance matrix is not locus-specific the joint posterior for all loci is stated as $P(\theta_l, \ldots, \theta_L, \Omega, \varepsilon_l, \ldots, \varepsilon_L \mid \mathbf{n_1}, \mathbf{m_1}, \ldots, \mathbf{n_L}, \mathbf{m_L}) \propto \{\prod_{l=1}^{l=L} P(\boldsymbol{n_l}, \boldsymbol{m_l} \mid \boldsymbol{x_l} = g(\theta_l))\ P(\theta_l \mid \Omega, \varepsilon_l)\ P(\varepsilon_l)\ \}\ P(\Omega)$, where P(Ω) and $P(\varepsilon_l)$ are respectively the priors for the variance-covariance matrix and ancestral frequency. Next we fit a model in which we allow allele frequencies to covary with an environmental variable *Y*, in other words *Y* constitutes a fixed effect on the mean of the multivariate normal distribution of the transformed population frequencies, hence $P(\theta_l \mid \Omega, \varepsilon_l, \beta) \sim N(\varepsilon_l + \beta Y, \varepsilon_l(1\text{-}\varepsilon_l)\ \Omega)$. The posterior is then estimated as $P(\theta_l, \Omega, \varepsilon_l, \beta \mid \mathbf{n_l}, \mathbf{m_l}) \propto P(\mathbf{n_l}, \mathbf{m_l} \mid \boldsymbol{x_l} = g(\theta_l))\ P(\theta_l \mid \Omega, \varepsilon_l, \beta)\ P(\Omega)\ P(\varepsilon_l)\ P(\beta)$, where P(β) is the prior of the coefficient β and for each locus the posterior of variance-covariance matrix of the null model is used as a prior of the covariance matrix in the alternative model. Tentative support for the alternative ('environmental') model over the null model is evaluated by means of Bayes factors.

Analyses were performed using the software package BayEnv (Coop et al. 2010) on the STEVIN Supercomputer Infrastructure at Ghent University. To estimate the variance-covariance matrix Ω three different chains (different seeds) were run each for 10 million iterations while discarding the first 50000 as a burn-in and thinning was set at 5000 (default). Convergence was checked by exploring trace plots for each element of Ω and estimating Geweke's Z- statistics between the first 25% and last 35% of the Markov chain using the CODA package in R (Plummer et al. 2006). While Coop et al. (2010) suggested to use a single draw of Ω when testing genotype-environment associations, we opted to be conservative and used a custom made python script to obtain the average of all draws of Ω. Although no control SNP dataset was at our disposal and hence adaptive SNPs were included when estimating the null model, it is assumed that selective SNPs constitute only a minor fraction of the total number of SNPs and therefore exert a minimal impact on Ω estimates (Coop et al. 2010). To test whether Ω represented the neutral population structure a mantel test was performed between the correlation matrix calculated from Ω and the pairwise G_{ST} matrix using the software package GenAlEx 6.501 (Peakall and Smouse 2006). Bayes factors, $\frac{P(M_1|n_l,m_l)}{P(M_0|n_l,m_l)}$ with M_1 and M_0 respectively the 'environmental' and 'null' model, were again interpreted according to Jeffreys' scale (Table 1) to differentiate between neutral and putative selective loci (Jeffreys 1961). The same environmental variables as those used in the multinomial regression were included in the Bayesian environmental analysis.

4. RESULTS

4.1. Environmental analysis

Climate variables showed strong correlations among all pairwise comparisons resulting in two clusters of climate variables where within-cluster comparisons were positive and between-cluster ones negative. The first cluster comprised Tmin, DD5, MCMT, EMT, NFFD and eFFP, while the other cluster encapsulated PAS, DD0, TD and bFFP (Fig. 4).

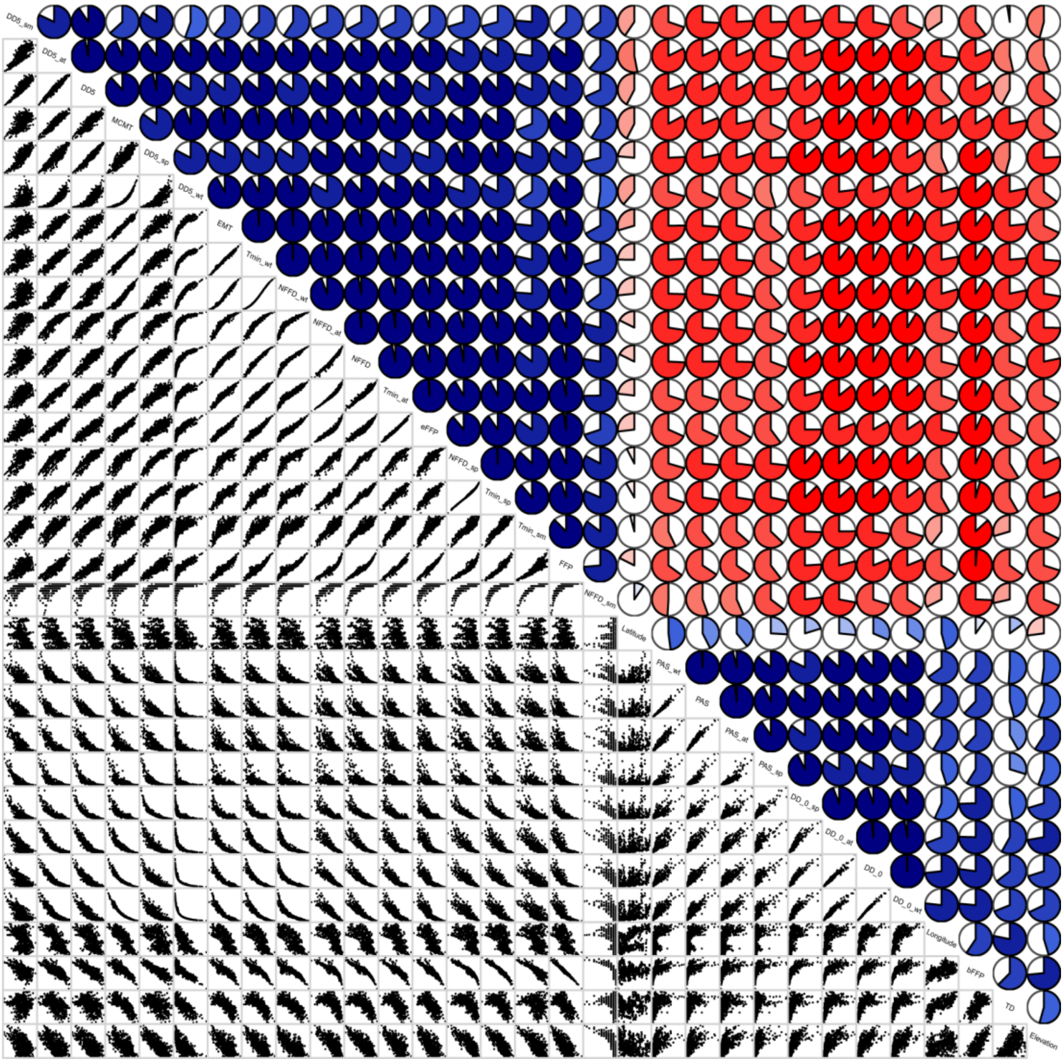

Fig. 4. A sorted correlation matrix of geographical and climate variables. Cells below the principal diagonal depict a bivariate scatterplot. Blue and red pie charts refer to respectively a positive and negative correlation. Magnitude of the correlation is represented by the size and darkness of the filled portion of the pie. Seasonality is indicated by respectively ‘at’ (autumn), ‘wt’ (winter), ‘sp’ (spring) and ‘sm’ (summer). Climate variables are abbreviated as follows: DD5= degree-days above 5°C, DD0= degree-days below 0°C, TD= continentality (°C), bFFP= begin frost-free period (Julian date), eFFP= end frost-free period (Julian date), FFP= frost-free period, NFFP= number of frost-free days, PAS= precipitation as snow (mm), EMT= extreme minimum temperature over 30 years (°C), Tmin= mean minimum temperature (°C), MCMT= mean coldest month temperature (°C).

In concordance with the strong correlations among climate variables the first principal component (PC) explained as much as 81.57% of the total variance (eigenvalue λ_1=22.84). We further retained the second and third PC as a composite environmental variable which additionally captured respectively 7.23% (λ_2=2.02) and 5.15% (λ_3=1.44) of the overall variance. All seasonal variables showed strong positive correlations with PCA1 with the exception of DD0 and PAS which were negatively correlated (0.76< |r| <0.97 with median value of 0.93). Similar trends could be observed for the annual variables with positive correlations for MCMT, DD5, NFFD, eFFP, FFP and EMT and negative ones for TD, DD0, bFFP and PAS (0.64< |r| <0.98 with median value of 0.94). DD5_sm and PAS (annual and seasonal) were positively correlated with PCA2 and negatively with TD (0.42< |r| <0.53). PC3 represented mainly the climate at summer as it showed strong positive associations with NFFD_sm, TD and Tmin_sm (0.64< |r| <0.98) (Fig.5, Fig. 6). A summary of the interpretation of each PC is given in Table 2.

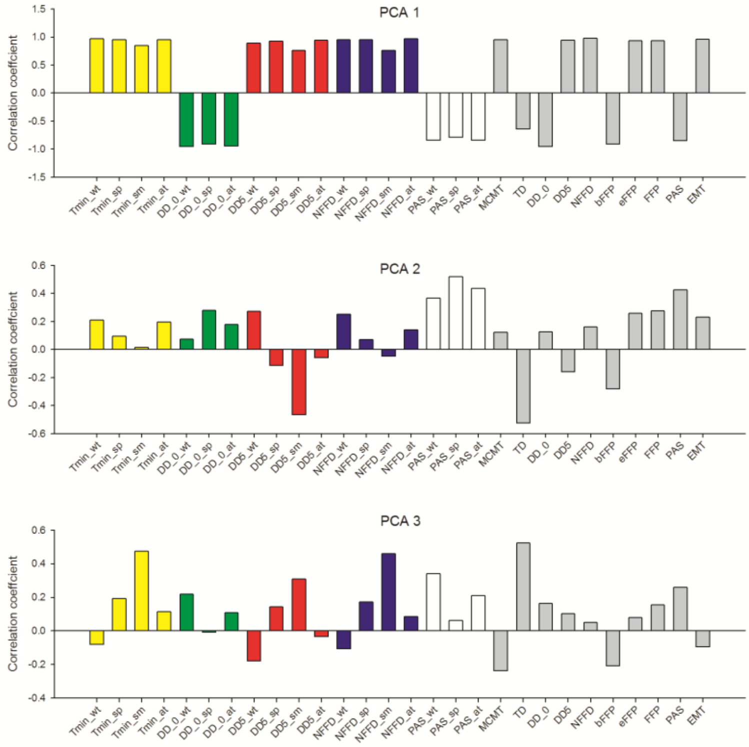

Fig. 5. Correlations between original environmental variables and three retained principal components. Variables are grouped by color (yellow: seasonal minimum temperature, green: seasonal degree-days below 0°C, red: seasonal degree-days above 5°C, blue: seasonal number of frost-free days, white: seasonal amount of snow, grey: annual climate variables).

PCA scores of individual trees belonging to the same population clustered together along the first PC axis, although some overlap between adjacent populations remained (Fig. 6). A t-test indicated that in 77% of all pairwise population comparisons a significant difference between mean PCA1 values could be detected.

Table 2. Direction and summary characteristics of the most important environmental variables associated with each principal component

PCA axis	Direction	Environmental variable	Characteristics of high positive PCA values
1	-	Degree-days below 0 °C	Low amount of snow fall and frost, high minimum
		Begin frost free period	temperatures, small differences between minimum
		Amount of snow fall	and maximum temperature and many degree-days
		Difference between coldest and warmest month	above 5°C and few below 0°C
	+	Degree-days above 5°C	
		Number of frost free days	
		Minimum temperature	
		End frost free period	
		Mean coldest month temperature	
		Extreme minimum temperature	
2	-	Difference between coldest and warmest month	High amount of snow fall, few degree-days above
		Degree-days above 5°C in summer	5°C in summer and small differences between
	+	Amount of snow fall	minimum and maximum temperature
3	+	Number of frost free days in summer	Many warm, frost free days in summer and large
		Minimum temperature in summer	differences between minimum and maximum
		Difference between coldest and warmest month	temperature

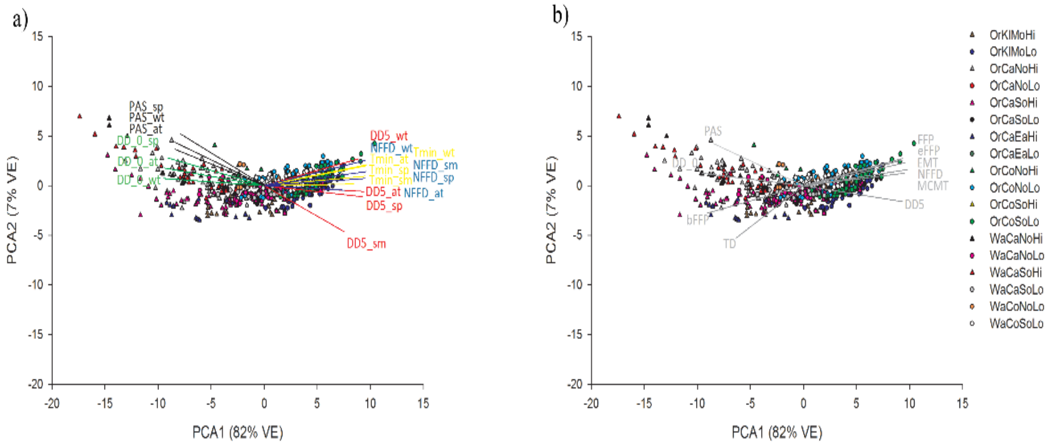

Fig. 6. Biplot showing the PCA scores of all 643 sampled trees along the 2 most important PCA axes (VE= variance explained). Orthogonal projection of each line onto the x- and y-axis represents the correlation of the focal variable with respectively the first and second PCA axis (for illustrative purposes these projections have been scaled by a factor x10) - (a) seasonal climate variables; b) annual climate variables).

As expected, climate and geography were strongly correlated. The first PC showed a strong negative association with longitude (Pearson r= -0.72, p<.0001) and elevation (Pearson r= -0.74, p<.0001). Longitude and elevation on their turn were positively correlated (Pearson r=0.45, p<.0001) illustrating the mountainous region at the east side of the Douglas fir distribution (Table 3, Fig. 7a). Populations at these locations faced harsh winter conditions (low winter temperatures, high amount of snow fall and extensive frost periods). The second PC seemed less affected by geography (Pearson |r|≤0.23) (Table 3, Fig. 7b) while the third PC denoted a moderate correlation with longitude (Pearson r=0.43, p<.0001) and latitude (Pearson r=0.38, p<.0001) indicating the warmer summers at the northeast (Table 3, Fig. 7c).

Table 3. Correlation between geographical and climate variation. Pearson correlation coefficients are given below diagonal, P-values above diagonal. Light-grey area refers to correlation coefficients within a geographical set of variables, dark-grey to the within-climate set and white to geography-climate associations.

	Latitude	Longitude	Elevation	PCA1	PCA2	PCA3
Latitude	1	<.0001	<.0001	<.0001	<.0001	<.0001
Longitude	0.458	1	<.0001	<.0001	<.0001	<.0001
Elevation	-0.271	0.454	1	<.0001	0.012	0.006
PCA1	-0.289	-0.717	-0.740	1	1	1
PCA2	0.227	-0.230	-0.099	0	1	1
PCA3	0.379	0.434	-0.108	0	0	1

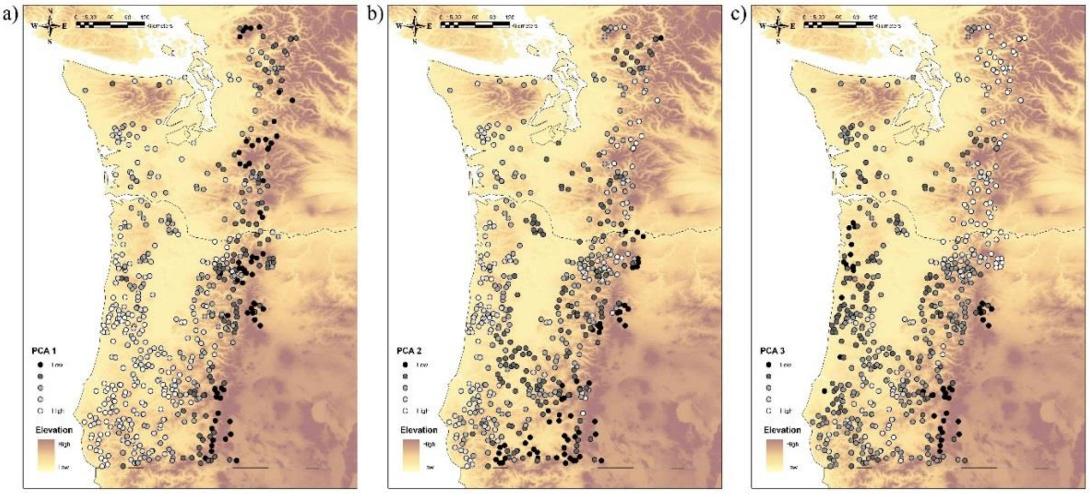

Fig. 7. Association between geography (latitude, longitude and elevation) and climate (a) first principal component, b) second principal component, c) third principal component).

4.2. Neutral population structure

We discarded 9 specimens that had microsatellite date but no allozyme scores (1 of OrCoNoLo, 1 of WaCaNoLo and 7 of WaCaSoLo). Pairwise measures of genetic differentiation indicated little population substructure among the 18 populations as G_{ST}-indices ranged from -0.004 to 0.025. Among all populations OrCoSoHi appeared the most genetic distinct population, followed by WaCoSoLo. However, overall genetic differentiation remained low resembling rather panmictic populations with substantial amount of gene flow between populations (Fig. 8). The 'isolation-by-distance' analysis further indicated that gene flow did not predominantly take place between adjacent populations as no relationship between genetic and geographical distance could be observed (r = 0.018, p=0.16).

	OrKlMoHi	OrKlMoLo	OrCaNoHi	OrCaNoLo	OrCaSoLo	OrCaEaHi	OrCaSoHi	OrCoNoHi	OrCoNoLo	OrCoSoHi	OrCoSoLo	WaCaNoHi	WaCaNoLo	WaCaSoHi	WaCaSoLo	WaCoNoLo	WaCoSoLo	OrCoEaLo
OrKlMoHi	0.000	0.000	0.001	0.002	0.001	0.003	0.001	-0.003	0.000	0.013	0.000	0.004	0.004	0.002	0.003	0.009	0.010	0.001
OrKlMoLo	0.000	0.000	0.000	0.002	0.000	0.003	0.001	-0.001	0.003	0.018	0.000	0.006	0.002	0.002	0.003	0.007	0.011	0.002
OrCaNoHi	0.001	0.000	0.000	0.001	0.000	0.000	0.001	-0.002	0.001	0.018	0.001	0.006	0.002	0.001	0.001	0.006	0.011	0.001
OrCaNoLo	0.002	0.002	0.001	0.000	0.002	0.003	0.000	0.003	0.002	0.021	0.001	0.005	0.003	0.002	0.002	0.006	0.013	0.001
OrCaSoLo	0.001	0.000	0.000	0.002	0.000	0.000	0.001	0.001	0.001	0.012	-0.001	0.005	0.002	0.001	0.001	0.007	0.011	-0.002
OrCaEaHi	0.003	0.003	0.000	0.003	0.000	0.000	0.001	0.001	0.003	0.018	0.004	0.009	0.004	0.000	0.003	0.006	0.012	0.005
OrCaSoHi	0.001	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.002	0.018	0.001	0.004	0.004	0.001	0.002	0.006	0.013	0.002
OrCoNoHi	-0.003	-0.001	-0.002	0.003	0.001	0.001	0.001	0.000	-0.004	0.006	-0.002	0.001	0.000	-0.001	-0.003	0.005	0.004	-0.002
OrCoNoLo	0.000	0.003	0.001	0.002	0.001	0.003	0.002	-0.004	0.000	0.015	0.001	0.005	0.004	0.002	0.003	0.008	0.009	-0.001
OrCoSoHi	0.013	0.018	0.018	0.021	0.012	0.018	0.018	0.006	0.015	0.000	0.015	0.018	0.022	0.018	0.015	0.025	0.025	0.007
OrCoSoLo	0.000	0.000	0.001	0.001	-0.001	0.004	0.001	-0.002	0.001	0.015	0.000	0.004	0.002	0.002	0.002	0.007	0.009	-0.002
WaCaNoHi	0.004	0.006	0.006	0.005	0.005	0.009	0.004	0.001	0.005	0.018	0.004	0.000	0.003	0.006	0.003	0.005	0.013	0.005
WaCaNoLo	0.004	0.002	0.002	0.003	0.002	0.004	0.004	0.000	0.004	0.022	0.002	0.003	0.000	0.001	0.000	0.003	0.010	0.003
WaCaSoHi	0.002	0.002	0.001	0.002	0.001	0.000	0.001	-0.001	0.002	0.018	0.002	0.006	0.001	0.000	0.000	0.007	0.011	0.001
WaCaSoLo	0.003	0.003	0.001	0.002	0.001	0.003	0.002	-0.003	0.003	0.015	0.002	0.003	0.000	0.000	0.000	0.003	0.010	0.008
WaCoNoLo	0.009	0.007	0.006	0.006	0.007	0.006	0.006	0.005	0.008	0.025	0.007	0.005	0.003	0.007	0.003	0.000	0.016	0.011
WaCoSoLo	0.010	0.011	0.011	0.013	0.011	0.012	0.013	0.004	0.009	0.025	0.009	0.013	0.010	0.011	0.010	0.016	0.000	0.014
OrCoEaLo	0.001	0.002	0.001	0.001	-0.002	0.005	0.002	-0.002	-0.001	0.007	-0.002	0.005	0.003	0.001	0.008	0.011	0.014	0.000

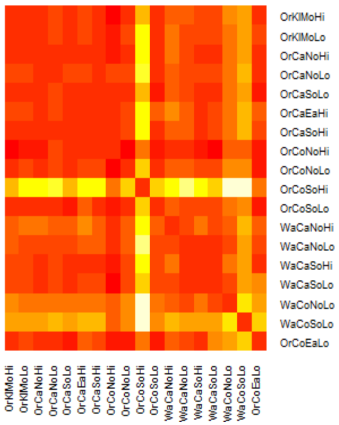

Fig. 8. Matrix of pairwise indices of differentiation (G_{ST}) of 18 populations (left panel). Visualization of the pairwise matrix where dark colors indicate no differentiation and light colors stronger genetic differentiation (right panel).

4.3. Adaptive genetic variation

The Bayesian outlier detection method classified 5 SNPs located in 3 distinct genes as putative selective. Three SNPs of the ribosomal protein 60s RPL31a showed a pattern consistent with that of a locus under selection. According to Jeffreys' thresholds (Table 1) there was decisive evidence for directional selection for SNPs at base pair locations 55 and 418 while this pattern was strongly supported at base pair location 295. Very strong evidence was found for directional selection in erd15, an early response to dehydration protein. BayeScan further identified Pm_CL1994Contig1, a caffeate O-methyltransferase, as a potential locus under stabilizing selection as it showed an unusual low level of genetic differentiation (Fig. 9).

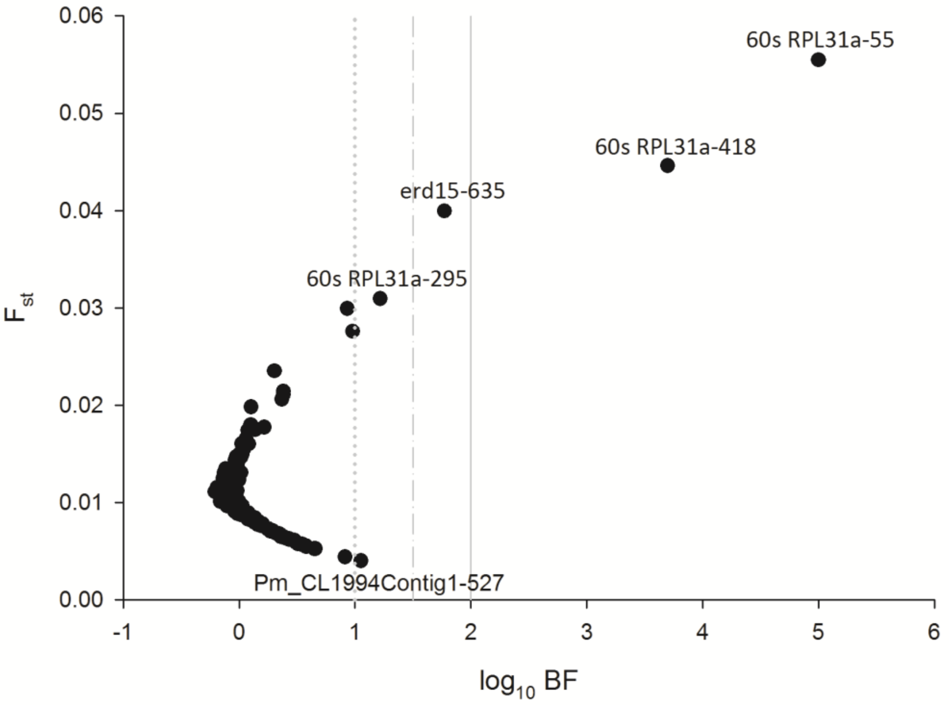

Fig. 9. A scatterplot of the logarithm of the Bayes factors ('selective' versus 'neutral' model) and mean F_{ST} values across all populations for each SNP. Vertical lines at $\log_{10}$BF equal to 1, 1.5 and 2 indicate respectively strong, very strong and decisive evidence for a selective locus, High F_{ST} values represent directional selection while small values denote balancing selection.

Regressing genotypes over environmental variables revealed 18 putative selective SNPs distributed over 14 genes. In congruence with the Bayesian outlier detection method both the genes 60s RPL31a (all three SNPs) and erd15 showed significant associations with the environment. Genotype frequencies at ribosomal protein 60s RPL31a were associated with longitude, elevation as well as PCA1 which is, given the strong climate-geography correlation (see Table 3), not unexpected. In addition, genotype frequencies at location 55 and 295 covaried with PCA3. There was an absence of environmental associations with genotypes of Pm_CL1994Contig1, a gene characterized by less genetic differentiation among populations than expected under a neutral scenario. Of all significant genotype-environment associations 50% (9/18) included geographical variables, 11% (2/18) climate ones and 39% (7/18) both geographical and climate variables (Table 4). Test statistics for the main environmental effects and parameter estimates of the multinomial logistic regression for each SNP are given respectively in Appendix 1 and Appendix 2.

Table 4. Overview of significant genotype-environmental associations identified by a multinomial logistic regressions at FDR=0.01.

SNP	Environmental variables
4CL1-520	Latitude
60s RPL31a-295	Elevation, longitude, PCA1, PCA3
60s RPL31a-418	Elevation, longitude, PCA1
60s RPL31a-55	Elevation, longitude, PCA1, PCA3
aba-609	Latitude
apx-288	Longitude
CN639480.1-430	Latitude
erd15-327	Latitude
erd15-635	Latitude, longitude, PCA3
f3h2-54	Longitude, PCA1
LEA-EMB11-227	PCA1
LEA-EMB11-263	Latitude
ES420560.1-203	Latitude
Pm_CL234Contig1-156	Longitude, PCA1
Pm_CL783Contig1-212	PCA1
sSPcDFD040B03103-274	Latitude, PCA3
sSPcDFE049E11411-220	Longitude
tbe-1259	Longitude

Typically for SNPs where all three genotype classes were present frequencies of one homozygote increased with higher environmental values, while the frequencies of the other homozygote showed an opposite trend, i.e. decreasing frequencies with increasing environmental values. Heterozygote frequencies were maximal at intermediate environmental values and lowest at the extremes of the environmental range. However, as our samples did not include the outer edges of the natural distribution of Douglas fir these extreme values were not always covered (Fig. 10).

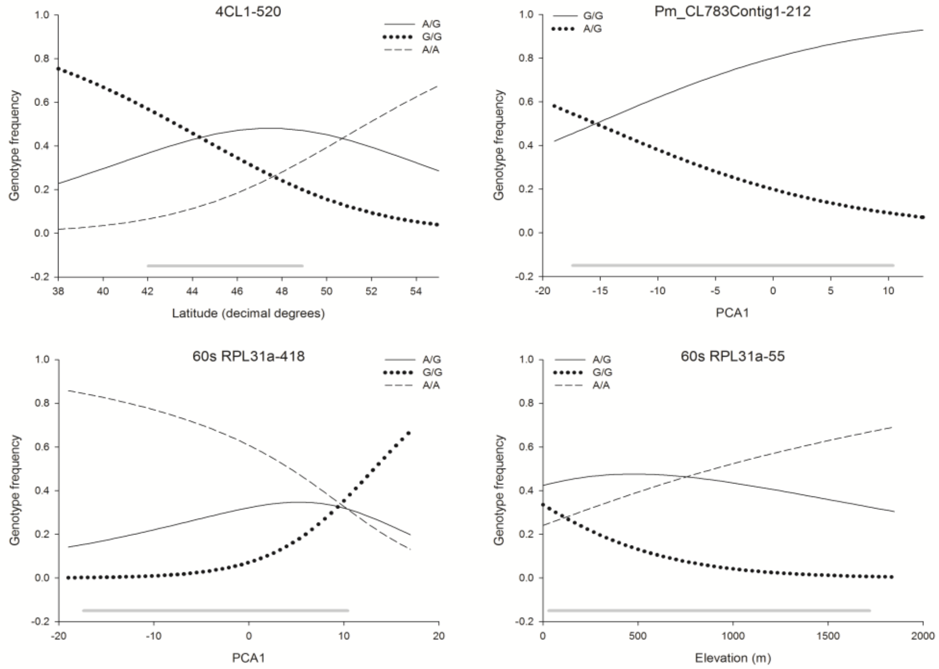

Fig. 10. Estimated genotype frequencies along environmental gradients for four SNPs (4CL1-520, Pm_CL783Contig1-212, 60s-RPL 31a-418 and 60s-RPL 31a-55) from multinomial logistic regressions. The grey horizontal bar above the x-axis indicates the observed range of environmental values within the dataset. Values outside this range are extrapolations and hence not supported by data.

In the Bayesian environmental analysis we did not encounter any convergence issues when estimating the variance-covariance matrix. Trace plots for all three chains showed no abrupt shifts in posterior parameter estimates, values of all chains overlapped extensively and Gewekes' Z-statistics showed no systematic patterns of deviating values. In total, 5% (8/171), 4% (7/171) and 2% (4/171) of the test statistics of respectively the first, second and third chain appeared above/below the absolute value of 2. The variability of allele frequencies within populations (diagonal elements of the matrix) were larger than covariances between population allele frequencies indicating little genetic population structure within our study area. This could however not be confirmed by a Mantel test as there was a lack of correspondence between the average posterior variance-covariance matrix and the matrix of pairwise measures of genetic differentiation (G_{ST}) (r = -0.46, p=0.40) (Fig. 11). This environmental analysis highlighted 16 putative selective SNPs distributed over 13 distinct genes. The majority of selective SNPs were associated with PCA1, PCA3, latitude or longitude (84%). Interestingly, the BayEnv software package identified almost exclusively new target genes under selection (12 out 13 genes) (Fig. 12, Fig. 13). Consistent with the previous two analytical procedures the Bayesian environmental model identified the early response to dehydration protein, erd15, as a potential target of selection. There was again strong/ decisive support for an association between Erd15 and PCA1, latitude and longitude. In contrast to the multinomial regression the SNP at base pair position 327 in this gene was also strongly linked to PCA2.

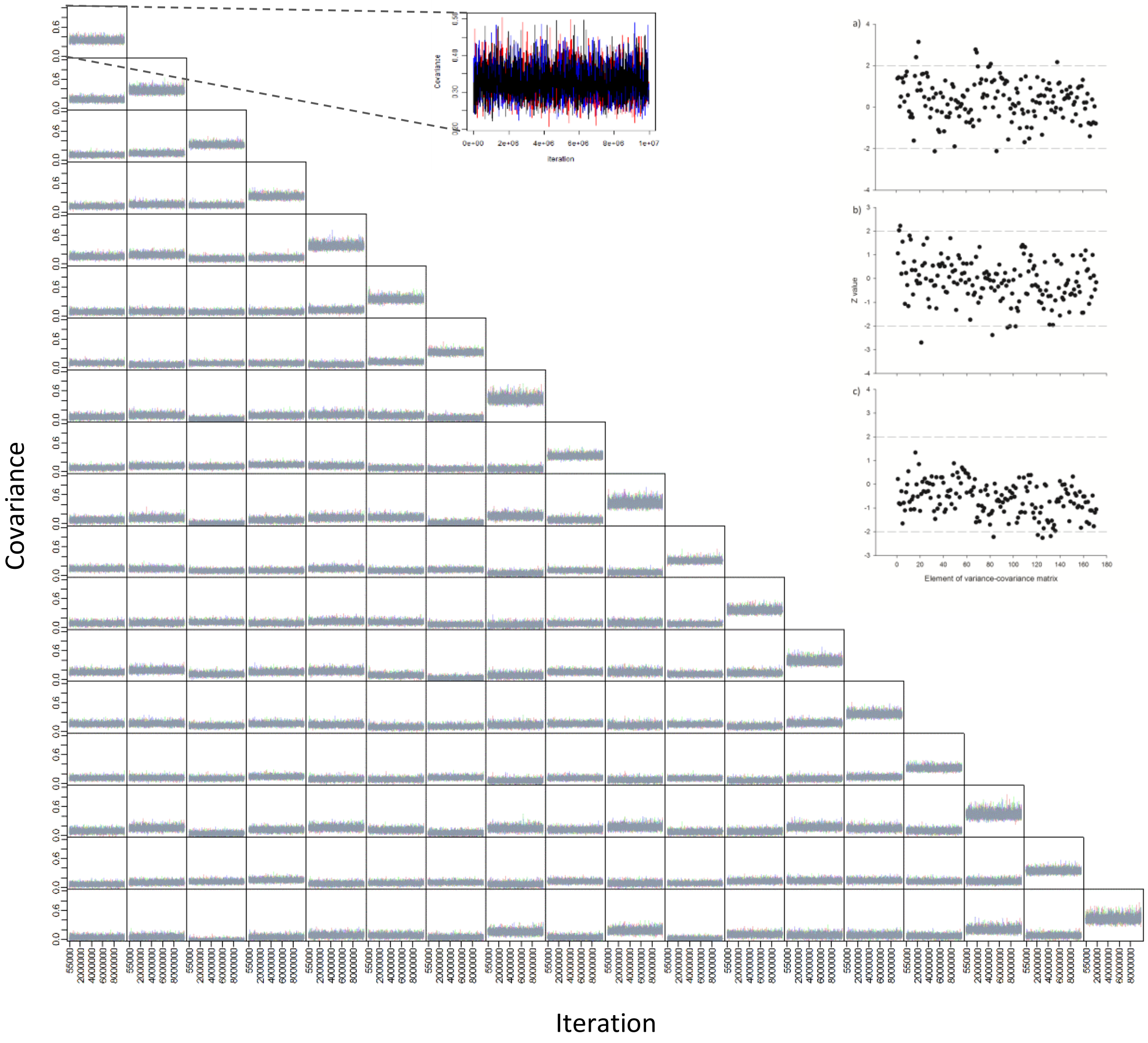

Fig. 11 Trace plots of each element of the variance-covariance matrix for 3 chains (red, black, blue), only the lower portion of the symmetric matrix is shown. Dimensions of matrix equals the number of populations (n=18). Horizontal axis denotes the iteration number, the vertical axis the covariance of allele frequencies between populations. In the middle upper part of the graph the first cell of matrix is enlarged. Right panel shows the distribution of Gewekes' Z-statistic of each cell of the matrix for the first (a), second (b) and third (c) chain.

In total the combined results of our three methods resulted in the identification of 33 SNPs in 24 distinct genes. The outlier detection method identified the fewest SNPs while the multinomial logistic regression and the Bayesian environmental analysis discovered approximately an equal number of SNPs (respectively 18 and 16). The correlative nature among environmental variables was reflected by the results of both the multinomial logistic regression and the Bayesian environmental analyses as respectively in 39% and 44% of the cases a single SNP was associated with more than one climate and/or geographical variable. Genotype-environment associations were dominated by geographical variables in the logistic regression (65% of all significant associations), while they were more balanced in the Bayesian analysis (52% all SNP associations with a BF>1 comprised a climate variable) (Fig. 13).

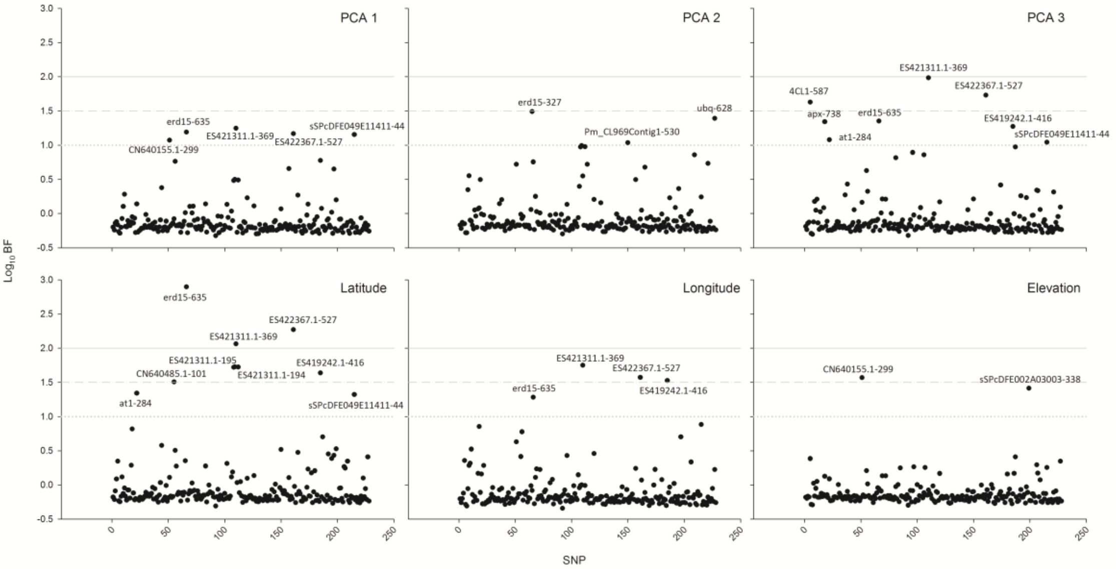

Fig. 12. Logarithms of the Bayes factors for each SNP indicating the evidence in favor of a 'selective' model over a 'neutral' one. Upper panel highlights the distribution of Bayes factors for environmental models including climate variables (PCA1, PCA2 and PCA3), lower panel those that include geographical variables (latitude, longitude and elevation).

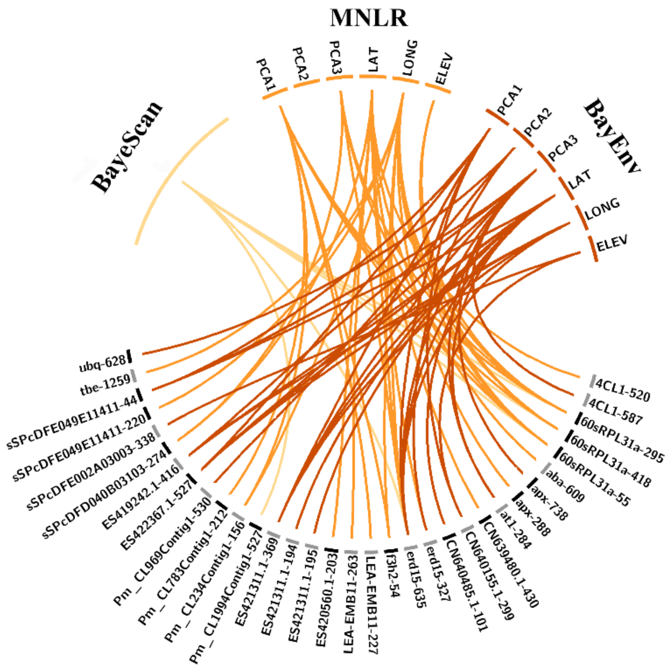

Fig. 13. Comparison among three analytical procedures to delineate putative selective genes: a Bayesian outlier detection method (BayeScan), a multinomial logistic regression (MNLR) and a Bayesian environmental analysis (BayEnv). Links represent significant associations (FDR=0.01) or Bayes factors larger than 1

4.4. The phenotype-genotype-environment spectrum

Finally, we attempted to integrate our results ('genotype-environment') into those studies previously conducted on Douglas fir on 'genotype-phenotype' (Eckert et al. 2009a) and 'environment-phenotype' associations (St Clair et al. 2005) (see Fig. 1). To make our results more comparable we focused on the same environmental variables as those used in St Clair et al. (2005), namely latitude, elevation, frost related variables, minimum temperature in winter and the difference between the coldest and warmest month.

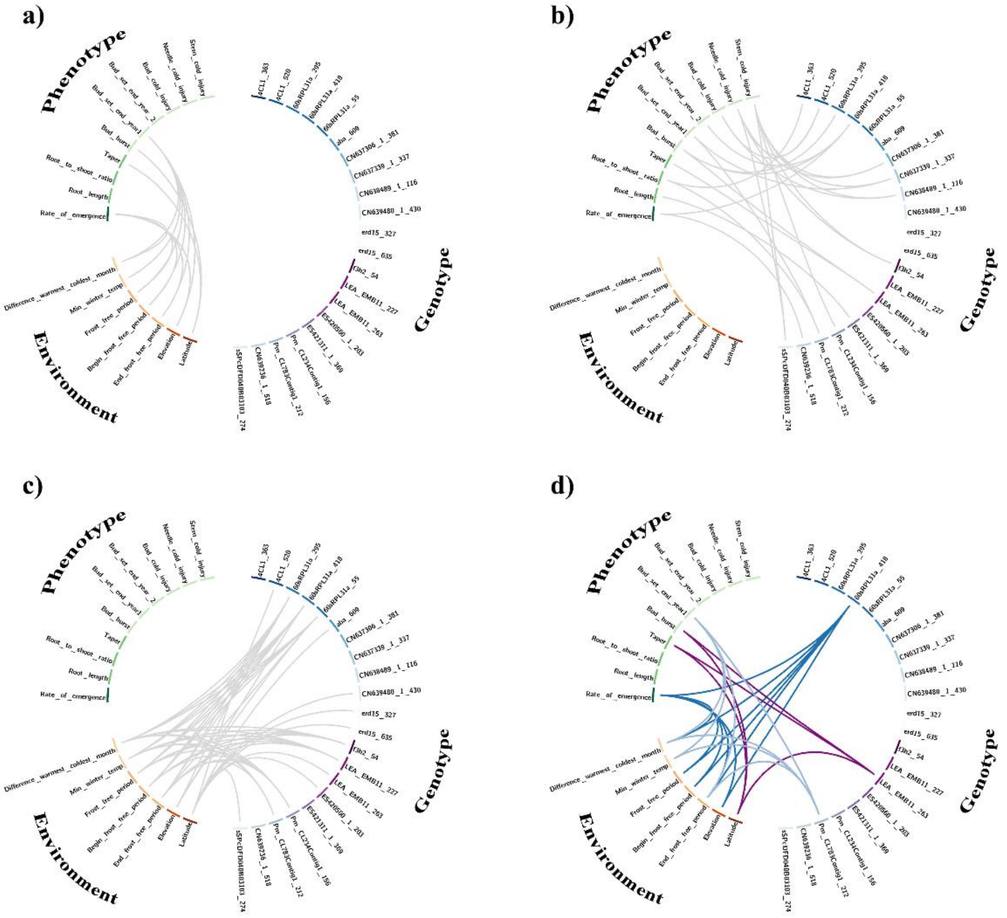

Fig. 14. Summary of the main findings among cold-hardiness related traits, genes and environment. Significant associations are depicted as a connection between corresponding phenotype, genotype or environmental variables. Similar colors below the label represent respectively traits belonging to the same phenotypic class (emergence, growth and resource partitioning or phenology and cold tolerance), SNPs within the same gene and variables part of a similar environmental class (latitude, elevation, frost date and winter temperature). While the original papers refer to multivariate composite phenotypes as well, for reasons of graphical clarity we restrict ourselves to the most important univariate variables - a) phenotype-environmental associations (only correlations of |r|>0.4 for the most important variables are shown) (St Clair et al. 2005, 2006); b) phenotype-genotype associations (Eckert et al. 2009); c) genotype-environmental associations (this study) and d) integrating phenotype-genotype-environmental associations for three selected genes (60sRPL31a, LEA_EMB11 and Pm_CL783Contig1).

We excluded environmental composite measures as these were lacking in the study of St. Clair et al. (2005, 2006) . As both previous studies applied general(ized) linear models we opted to fit a multinomial logistic regression to explore the association between georeferenced genotypes and these univariate environmental variables. Corrections were again made for multiple testing (FDR=0.01). From the original 117 screened genes we were able to identify 11 genes related to 7 environmental variables while of those genes no less than 64% were associated to cold-related traits (Eckert et al. 2009a) (Fig. 14). Assembling the main findings of all three studies resulted in a comprehensive and highly interlinked 'pheno-geno-environmental' network. Such a network highlights closed circuits where both components of the phenotype-genotype association are linked to the same environmental variable. Three distinct genes are an integral part of such circuits (60s RPL31a (a 60s ribosomal proteinL31a), LEA-EMB11 (a late embryogenesis abundant protein (LEA)) and Pm_CL783Contig1 (a SOUL heme-binding family protein)) (Fig. 14). Genotype frequencies of LEA-EMB11 for example were related to latitude, taper (ratio of diameter to height, see Eckert et al. (2009) for details) and timing of the first green needles from the terminal bud. The latter two traits were on their turn strongly related to latitude thereby closing the pheno-geno-environmental circuit.

5. DISCUSSION

Multinomial logistic regression models revealed 18 genes that are potentially a target of natural selection. An outlier detection method confirmed similar patterns for two genes while the Bayesian environmental analysis could support an adaptive signature for one of those two genes, the early-response dehydration protein. In general, results from the multinomial logistic regression and the Bayesian environmental analysis did not corroborate. As much as 15% of the candidate SNPs were identified as putative selective in at least one of the applied methods, which could be expected based on the fact that SNPs were not randomly chosen but carefully screened in advance (Prunier et al. 2012). Combining data on cold-hardiness related phenotypes, genotypes and environment also identified three genes that may code for important cold-hardiness related traits.

While traditional outlier detection methods are particularly prone to falsely identifying adaptive genes, the BayeScan algorithm is assumed to be rather conservative (Narum and Hess 2011). Our results provided further support for this as only three genes were selected as potential targets. The outlier detection method did not take advantage of the environmental gradient and such agnostic approaches have been shown to have reduced statistical power when screening putative adaptive loci (Schoville et al. 2012). As such, in these type of analyses some important genotype-environment associations may have remained cryptic. The multinomial logistic regression on its turn did take into account environmental data but can be seriously flawed by confounding demographical processes which may mimic the pattern induced by natural selection. In addition, from a statistical point of view populations may not be treated as independent units as gene flow between adjacent populations may cause allele frequencies to covary at a local scale resulting in inflated degrees of freedom and concomitant increased rates of type I errors when no remedial actions are taken (Coop et al. 2010). We consider this to be unlikely in our study as neutral population structure across this species' range was low to nonexistent. Lack of genetic structure is not unusual in conifers as populations often tend be large, thereby minimizing the effects of genetic drift, and exhibit substantial levels of gene flow as no barriers are apparent (Neale and Savolainen 2004, Buschiazzo et al. 2012). However we still need to urge some caution when interpreting the results as pairwise measures of genetic differentiation are summary statistics across all loci while the extent to which ancestral history is reflected in each genomic region may vary substantially. One way to circumvent these criticism is to account for the demographic history by including scores obtained from a principal coordinate analysis on a genetic distance matrix (or any method that quantifies the extent of neutral population structure) as covariates in the logistic model. While neutral population structure may infer false patterns of selection, accounting for these demographic events may cause substantial loss of statistical power in some cases. When neutral population structure covaries strongly with environmental gradients, 'neutral' (i.e. null model) and 'selective' model will be equally likely and hence results in low Bayes factors. The extent to which this occurs will most likely depend on the strength of the correlation between neutral allele frequencies and environment. Notably the lack of genetic structure in this study may have resulted in a non-significant Mantel test between the genetic distance matrices of respectively the pairwise measures of neutral populations structure and the Bayesian variance-covariance matrix estimated in the Bayesian environmental analysis.

In line with our expectations outlier detection methods were not able to identify many putative selective SNPs. These detection methods gain most power when adaptive allele frequency distributions are characterized by local ‘bell’ shaped curves at specific distinct environmental values. Such a pattern gives rise to high levels of overall genetic differentiation in contrast to situations characterized by a smooth shift from one allelic variant to the other along an environmental gradient (Narum and Hess 2011, Prunier et al. 2012). As samples in our study were intentionally collected along environmental gradients adaptive SNPs most likely will be characterized by a clinal pattern rather than a ‘bell’ shaped one and hence may miss the adequate statistical power.

This study highlighted a small number of genes that showed consistent patterns either across the genotype-phenotype-environment spectrum or across all methodologies applied in this study. Frequencies of SNPs at the late embryogenesis abundant (LEA) protein were highly dependent on latitude and linked to both timing of needle appearance (bud burst) and the ratio of diameter to height of 2nd year seedlings (taper). Assigning a direct causative relationship between allelic variation at the LEA protein and both traits remains however elusive, even when the extent of linkage disequilibrium in natural conifer populations are known to be extremely small (Neale and Savolainen 2004) and target SNPs could therefore very likely be situated within a gene. Conifers could adapt to extreme freezing conditions by both avoiding freezing and reducing its sensitivity to frost. At higher latitude needles at the terminal bud appear to emerge later and taper is reduced. These traits refer to a trade off between postponing the start of the growing season and reducing the probability of extensive frost damage during early spring. However, exposure to low temperatures often induces biochemical changes such as an up-regulation of LEA proteins in a variety of plants (Hannah et al. 2005, Lorenz et al. 2006). These hydrophilic proteins purportedly mitigate the disruptive effects on lipid bilayers during frost through membrane stabilization and prevention of cellular crystallization (Garay-Arroyo et al. 2000). As all trait responses may be interrelated it remains problematic to assign the true phenotypic effect of allelic variation observed at a locus in such observational studies. However, these explorative genome scans can give rise to well defined and testable hypotheses. Using only a reduced set of SNPs and a small set of samples raised in common gardens in climate chambers, clear-cut experiments at reduced financial costs become feasible. Another interesting gene was the early responsive to dehydration (Erd) 15 gene. All three methodologies identified this gene as putative selective and allele/genotype frequencies were associated with latitude, longitude, PCA1 and PCA3. In Arabidopsis Erd15 indirectly controls freezing tolerance by modulating the sensitivity of the plant to the phytohormone abscisic acid (ABA). It is well known that ABA plays a vital role in controlling responses to various abiotic stressors such as severe drought and frost (Kariola et al. 2006).

Complex traits such as cold-hardiness appear to be polygenic where the trait is controlled by many genes with each locus contributing only small effects to the total phenotype (Eckert et al. 2009a, Neale and Kremer 2011). Due to this small effect size and the multitude of tests being conducted, these association studies inevitably suffer from a certain number of false negative and false positive associations. As such, unequivocally assigning a specific gene as a true important determinant of an adaptive trait remains cumbersome. However, results from multinomial logistic regressions (for example) provide us an opportunity to further investigate, with a minimal extra effort, the generality of previously identified environment-genotype associations. Collecting a few extra samples outside the current sampled environmental range allows us to assess whether allele frequencies of truly independent data sets coincide with extrapolated frequencies based on the logistic model. In addition, consistent and recurrent appearance of a gene throughout a combined genotype-phenotype-environment analysis may help validate its true function and limit the number of genes that may warrant further investigation.

In conclusion, using three different analytical procedures we identified 33 SNPs located in 24 distinct genes. There was however low congruence between the two analytical approaches that included environmental gradients. One exception is the early responsive to dehydration protein that showed a remarkable consistent pattern of directional selection across all three methods. Combining results of the multinomial logistic regression with preexisting data on cold-hardiness related phenotypes revealed three promising genes: a late embryogenesis abundant protein, a SOUL heme-binding family protein and a 60s ribosomal proteinL31a.

6. REFERENCES

Allendorf, F. W., S. N. Aitken, and G. Luikart. 2012. *Conservation and the Genetics of Populations*. Wiley-Blackwell Publishing.

Balloux, F. and N. Lugon-Moulin. 2002. The estimation of population differentiation with microsatellite markers. Molecular Ecology **11**:155-165.

Beaumont, M. A. 2005. Adaptation and speciation: what can F-St tell us? Trends in Ecology & Evolution **20**:435-440.

Buschiazzo, E., C. Ritland, J. Bohlmann, and K. Ritland. 2012. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. BMC Evolutionary Biology **12**:8.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard. 2010. Using environmental correlations to identify loci underlying local adaptation. Genetics **185**:1411-1423.

Eckert, A. J., A. D. Bower, J. L. Wegrzyn, B. Pande, K. D. Jermstad, K. V. Krutovsky, J. B. S. Clair, and D. B. Neale. 2009a. Asssociation genetics of coastal Douglas fir (*Pseudotsuga menziesu* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. Genetics **182**:1289-1302.

Eckert, A. J., J. L. Wegrzyn, B. Pande, K. D. Jermstad, J. M. Lee, J. D. Liechty, B. R. Tearse, K. V. Krutovsky, and D. B. Neale. 2009b. Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*). Genetics **183**:289-298.

Foll, M. and O. Gaggiotti. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics **180**:977-993.

Fox, C. W. and J. B. Wolf. 2006. *Evolutionary genetics: concepts and case studies*. Oxford University Press.

Friendly, M. 2002. Corrgrams: exploratory displays for correlation matrices. The American Statistician **56**:316-324.

González-Martínez, S. C., K. V. Krutovsky, and D. B. Neale. 2006. Forest-tree population genomics and adaptive evolution. New Phytologist **170**:227-238.

Gusta, L. V. and M. Wisniewski. 2013. Understanding plant cold hardiness: an opinion. Physiologia Plantarum **147**:4-14.

Hannah, M. A., A. G. Heyer, and D. K. Hincha. 2005. A global survey of gene regulation during cold acclimation in *Arabidopsis thaliana*. PLoS Genet **1**:e26.

Hijmans, R. J., J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology **25**:1965-1978.

Jeffreys, H. 1961. *Theory of probability*. Oxford Univ. Press.

Jermstad, K. D., D. L. Bassoni, K. S. Jech, G. A. Ritchie, N. C. Wheeler, and D. B. Neale. 2003. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas fir. III. Quantitative trait loci-by-environment interactions. Genetics **165**:1489-1506.

Jermstad, K. D., D. L. Bassoni, K. S. Jech, N. C. Wheeler, and D. B. Neale. 2001a. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. I. Timing of vegetative bud flush. Theoretical and Applied Genetics **102**:1142-1151.

Jermstad, K. D., D. L. Bassoni, N. C. Wheeler, T. S. Anekonda, S. N. Aitken, W. T. Adams, and D. B. Neale. 2001b. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. II. Spring and fall cold-hardiness. Theoretical and Applied Genetics **102**:1152-1158.

Joost, S., A. Bonin, M. W. Bruford, L. Despres, C. Conord, G. Erhardt, and P. Taberlet. 2007. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Molecular Ecology **16**:3955-3969.

Kariola, T., G. Brader, E. Helenius, J. Li, P. Heino, and E. T. Palva. 2006. Early responsive to dehydratation 15, a negative regulator of abscisic acid responses in Arabidopsis. Plant Physiology **142**:1559-1573.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge Univ. Press.

Krutovsky, K. V., J. B. S. Clair, R. Saich, V. D. Hipkins, and D. B. Neale. 2009. Estimation of population structure in coastal Douglas-fir *Pseudotsuga menziesii* (Mirb.) *Franco var. menziesii* using allozyme and microsatellite markers. Tree Genetics & Genomes **5**:641-658.

Lee, B. H., D. A. Henderson, and J.-K. Zhu. 2005. The Arabidopsis cold-responsive transcriptome and its regulation by ICE1. Plant Cell **17**:3155-3175.

Lesaffre, E. and A. B. Lawson. 2012. *Bayesian biostatistics*. Wiley and Sons.

Lorenz, W. W., F. Sun, C. Liang, D. Kolychev, H. Wang, X. Zhao, M. M. Cordonnier-Pratt, L. H. Pratt, and J. F. Dean. 2006. Water stress-responsive genes in loblolly pine (Pinus taeda) roots identified by analyses of expressed sequence tag libraries. Tree Physiology **26**:1-16.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research **27**:209-220.

Meirmans, P. G. and P. W. Hedrick. 2011. Assessing population structure: FST and related measures. Molecular Ecology Resources **11**:5-18.

Narum, S. R. and J. E. Hess. 2011. Comparison of F(ST) outlier tests for SNP loci under selection. Molecular Ecology Resources **11**:184-194.

Neale, D. B. and A. Kremer. 2011. Forest tree genomics: growing resources and applications. Nature Reviews Genetics **12**:111-122.

Neale, D. B. and O. Savolainen. 2004. Association genetics of complex traits in conifers. Trends in Plant Science **9**:325-330.

Nei, M. 1977. F-statistics and analysis of gene diversity in subdivided populations. Annals of Human Genetics **41**:225-233.

Nicholson, G., A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson, and P. Donnelly. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society Series B-Statistical Methodology **64**:695-715.

Nielsen, R. 2005. Molecular signatures of natural selection. Annual Review of Genetics **39**:197-218.

Novembre, J. and A. Di Rienzo. 2009. Spatial patterns of variation due to natural selection in humans. Nature Reviews Genetics **10**:745-755.

Peakall, R. and P. E. Smouse. 2006. Genalex 6: genetic analysis in excel. Population genetic software for teaching and research. Molecular Ecology Notes **6**:288-295.

Pérez-Figueroa, A., M. J. GarcÍA-Pereira, M. Saura, E. RolÁN-Alvarez, and A. Caballero. 2010. Comparing three different methods to detect selective loci using dominant markers. Journal of Evolutionary Biology **23**:2267-2276.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News **6**:7-11.

Prunier, J., S. Gerardi, J. Laroche, J. Beaulieu, and J. Bousquet. 2012. Parallel and lineage-specific molecular adaptation to climate in boreal black spruce. Molecular Ecology **21**:4270-4286.

Schoville, S. D., A. Bonin, O. Francois, S. Lobreaux, C. Melodelima, and S. Manel. 2012. Adaptive genetic variation on the landscape: methods and cases. Pages 23-43 *in* D. J. Futuyma, editor. Annual Review of Ecology, Evolution, and Systematics, Vol 43.

Selkoe, K. A. and R. J. Toonen. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecology Letters **9**:615-629.

St Clair, J. B. 2006. Genetic variation in fall cold hardiness in coastal Douglas-fir in western Oregon and Washington. Canadian Journal of Botany-Revue Canadienne De Botanique **84**:1110-1121.

St Clair, J. B., N. L. Mandel, and K. W. Vance-Boland. 2005. Genecology of Douglas fir in western Oregon and Washington. Annals of Botany **96**:1199-1214.

Storey, J. D. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B-Statistical Methodology **64**:479-498.

Storey, J. D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America **100**:9440-9445.

Wang, T., A. Hamann, D. Spittlehouse, and T. N. Murdock. 2012. ClimateWNA – High-resolution spatial climate data for western North America. . Journal of Applied Meteorology and Climatology **61**:16-29.

Wheeler, N. C., K. D. Jermstad, K. Krutovsky, S. N. Aitken, G. T. Howe, J. Krakowski, and D. B. Neale. 2005. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. IV. Cold-hardiness QTL verification and candidate gene mapping. Molecular Breeding **15**:145-156.

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics **15**:323-354.

7. APPENDIX

Appendix 1. Test statistics, p- and q- values of all significant genotype-environment associations per SNP at a FDR=0.01.

SNP	Variable	df	Wald Chi-Square	p-value	q-value
4CL1-520	Latitude	2	31.3269	1.58E-07	1.10E-05
60s RPL31a-295	Elevation	1	15.8678	6.79E-05	3.58E-03
	Longitude	1	27.1074	1.93E-07	8.20E-06
	PCA1	2	29.7973	4.72E-04	8.57E-03
	PCA3	1	14.8141	1.19E-04	7.38E-03
60s RPL31a-418	Elevation	2	20.7653	3.10E-05	2.45E-03
	Longitude	2	31.4371	1.49E-07	8.20E-06
	PCA1	2	20.3247	3.86E-05	1.63E-03
60s RPL31a-55	Elevation	2	29.7973	3.39E-07	5.35E-05
	Longitude	2	48.517	1.00E-10	1.28E-08
	PCA1	2	33.1797	6.24E-08	7.93E-06
	PCA3	2	19.1067	7.10E-05	7.38E-03
aba-609	Latitude	2	16.109	3.18E-04	6.36E-03
apx-288	Longitude	2	15.0721	5.33E-04	7.57E-03
CN639480.1-430	Latitude	2	21.7926	1.85E-05	5.19E-04
erd15-327	Latitude	2	21.9673	1.70E-05	5.19E-04
erd15-635	Latitude	2	34.9738	2.54E-08	3.56E-06
	Longitude	2	18.4401	9.90E-05	1.81E-03
	PCA3	2	39.8386	2.20E-09	4.62E-07
f3h2-54	Longitude	2	23.803	6.78E-06	2.16E-04
	PCA1	2	4.6756	3.06E-05	1.63E-03
LEA-EMB11-227	PCA1	2	7.5613	1.20E-04	3.82E-03
LEA-EMB11-263	Latitude	1	19.0563	1.27E-05	5.19E-04
ES420560.1-203	Latitude	2	15.7131	3.87E-04	6.78E-03
Pm_CL234Contig1-156	Longitude	2	20.8955	2.90E-05	7.41E-04
	PCA1	2	10.2414	1.72E-04	4.38E-03
Pm_CL783Contig1-212	PCA1	1	6.1289	3.97E-04	8.41E-03
sSPcDFD040B03103-274	Latitude	1	13.4166	2.49E-04	5.82E-03
	PCA3	1	14.497	1.40E-04	7.38E-03
sSPcDFE049E11411-220	Longitude	2	17.4483	1.63E-04	2.60E-03
tbe-1259	Longitude	1	11.3229	7.66E-04	9.78E-03

Appendix 2. Parameter estimates of the multinomial logistic regression for each SNP where the main effect of the environmental variable is significant at FDR=0.01.

SNP	Genotype	Variable	DF	Estimate	SE	Wald	Pr >	SNP	Variable	Genotype	DF	Estimate	SE	Wald	Pr >
						Chi-Square	Chi-Square							Chi-Square	Chi-Square
4CL1-520	A/G	Intercept	1	10.129	3.053	11.010	<.001	4CL1-520	Latitude	A/G	1	-0.200	0.067	8.809	0.003
4CL1-520	G/G	Intercept	1	18.475	3.214	33.038	<.001	4CL1-520	Latitude	G/G	1	-0.388	0.071	29.709	<.001
60s RPL31a-295	G/G	Intercept	1	0.704	0.214	10.806	0.001	60s RPL31a-295	Elevation	G/G	1	0.001	0.000	15.868	<.001
60s RPL31a-295	G/G	Intercept	1	96.383	18.244	27.910	<.001	60s RPL31a-295	Longitude	G/G	1	0.771	0.148	27.107	<.001
60s RPL31a-295	G/G	Intercept	1	1.560	0.107	211.303	<.001	60s RPL31a-295	PCA1	G/G	1	-0.105	0.030	12.223	<.001
60s RPL31a-295	G/G	Intercept	1	1.577	0.109	208.972	<.001	60s RPL31a-295	PCA3	G/G	1	0.606	0.157	14.814	<.001
60s RPL31a-418	A/G	Intercept	1	-0.313	0.203	2.373	0.123	60s RPL31a-418	Elevation	A/G	1	-0.001	0.000	3.166	0.075
60s RPL31a-418	G/G	Intercept	1	-0.430	0.326	1.743	0.187	60s RPL31a-418	Elevation	G/G	1	-0.003	0.001	19.551	<.001
60s RPL31a-418	A/G	Intercept	1	-39.408	14.397	7.493	0.006	60s RPL31a-418	Longitude	A/G	1	-0.316	0.117	7.256	0.007
60s RPL31a-418	G/G	Intercept	1	-207.200	39.296	27.793	<.001	60s RPL31a-418	Longitude	G/G	1	-1.667	0.318	27.396	<.001
60s RPL31a-418	A/G	Intercept	1	-0.635	0.093	46.982	<.001	60s RPL31a-418	PCA1	A/G	1	0.062	0.025	6.059	0.014
60s RPL31a-418	G/G	Intercept	1	-2.143	0.186	132.879	<.001	60s RPL31a-418	PCA1	G/G	1	0.222	0.054	17.008	<.001
60s RPL31a-55	A/G	Intercept	1	0.566	0.204	7.700	0.006	60s RPL31a-55	Elevation	A/G	1	-0.001	0.000	7.084	0.008
60s RPL31a-55	G/G	Intercept	1	0.331	0.294	1.270	0.26	60s RPL31a-55	Elevation	G/G	1	-0.003	0.001	28.246	<.001
60s RPL31a-55	A/G	Intercept	1	-60.716	14.187	18.317	<.001	60s RPL31a-55	Longitude	A/G	1	-0.495	0.116	18.364	<.001
60s RPL31a-55	G/G	Intercept	1	-198.700	30.911	41.303	<.001	60s RPL31a-55	Longitude	G/G	1	-1.604	0.251	40.950	<.001
60s RPL31a-55	A/G	Intercept	1	0.101	0.091	1.244	0.265	60s RPL31a-55	PCA1	A/G	1	0.094	0.024	14.988	<.001
60s RPL31a-55	G/G	Intercept	1	-1.410	0.160	77.270	<.001	60s RPL31a-55	PCA1	G/G	1	0.239	0.046	26.665	<.001
60s RPL31a-55	A/G	Intercept	1	0.086	0.090	0.927	0.336	60s RPL31a-55	PCA3	A/G	1	-0.201	0.117	2.954	0.086
60s RPL31a-55	G/G	Intercept	1	-1.410	0.157	81.038	<.001	60s RPL31a-55	PCA3	G/G	1	-0.983	0.227	18.794	<.001
aba-609	A/G	Intercept	1	9.720	3.083	9.938	0.002	aba-609	Latitude	A/G	1	-0.193	0.069	7.898	0.005
aba-609	G/G	Intercept	1	0.744	3.065	0.059	0.808	aba-609	Latitude	G/G	1	0.004	0.068	0.003	0.959
apx-288	A/C	Intercept	1	-96.119	26.435	13.221	<.001	apx-288	Longitude	A/C	1	-0.762	0.215	12.605	<.001
apx-288	C/C	Intercept	1	-136.700	80.190	2.908	0.088	apx-288	Longitude	C/C	1	-1.075	0.650	2.734	0.098
CN639480.1-430	A/G	Intercept	1	9.295	2.361	15.498	<.001	CN639480.1-430	Latitude	A/G	1	-0.214	0.053	16.372	<.001
CN639480.1-430	G/G	Intercept	1	-5.296	3.256	2.646	0.104	CN639480.1-430	Latitude	G/G	1	0.085	0.072	1.406	0.236
erd15-327	A/G	Intercept	1	26.615	12.892	4.262	0.039	erd15-327	Latitude	A/G	1	-0.531	0.274	3.757	0.053
erd15-327	G/G	Intercept	1	42.693	12.494	11.676	<.001	erd15-327	Latitude	G/G	1	-0.834	0.265	9.888	0.002
erd15-635	A/C	Intercept	1	-17.711	21.726	0.665	0.415	erd15-635	Latitude	A/C	1	0.427	0.464	0.845	0.358
erd15-635	C/C	Intercept	1	35.369	20.033	3.117	0.077	erd15-635	Latitude	C/C	1	-0.649	0.429	2.287	0.13
erd15-635	A/C	Intercept	1	-159.900	272.700	0.344	0.558	erd15-635	Longitude	A/C	1	-1.332	2.239	0.354	0.552
erd15-635	C/C	Intercept	1	-354.600	269.000	1.737	0.188	erd15-635	Longitude	C/C	1	-2.950	2.209	1.782	0.182
erd15-635	A/C	Intercept	1	2.175	1.140	3.642	0.056	erd15-635	PCA3	A/C	1	0.202	0.831	0.059	0.808
erd15-635	C/C	Intercept	1	6.389	1.081	34.943	<.001	erd15-635	PCA3	C/C	1	-1.436	0.797	3.242	0.072
f3h2-54	A/C	Intercept	1	-61.038	12.952	22.209	<.001	f3h2-54	Longitude	A/C	1	-0.494	0.105	21.959	<.001
f3h2-54	C/C	Intercept	1	27.748	39.389	0.496	0.481	f3h2-54	Longitude	C/C	1	0.252	0.321	0.614	0.433
f3h2-54	A/C	Intercept	1	-0.376	0.083	20.555	<.001	f3h2-54	PCA1	A/C	1	0.103	0.023	20.737	<.001
f3h2-54	C/C	Intercept	1	-3.127	0.256	149.535	<.001	f3h2-54	PCA1	C/C	1	0.053	0.069	0.593	0.441
LEA-EMB11-227	A/G	Intercept	1	-0.513	0.084	37.495	<.001	LEA-EMB11-227	PCA1	A/G	1	0.098	0.023	18.052	<.001
LEA-EMB11-227	G/G	Intercept	1	-4.864	0.580	70.413	<.001	LEA-EMB11-227	PCA1	G/G	1	0.030	0.154	0.039	0.844
LEA-EMB11-263	C/C	Intercept	1	15.074	3.045	24.503	<.001	LEA-EMB11-263	Latitude	C/C	1	-0.293	0.067	19.056	<.001
ES420560.1-203	A/G	Intercept	1	-14.917	20.676	0.521	0.471	ES420560.1-203	Latitude	A/G	1	0.415	0.467	0.789	0.374
ES420560.1-203	G/G	Intercept	1	-0.874	20.500	0.002	0.966	ES420560.1-203	Latitude	G/G	1	0.146	0.463	0.100	0.752
Pm_CL234Contig1-156	A/T	Intercept	1	-64.378	14.016	21.098	<.001	Pm_CL234Contig1-156	Longitude	A/T	1	-0.520	0.114	20.780	<.001
Pm_CL234Contig1-156	T/T	Intercept	1	-15.841	25.910	0.374	0.541	Pm_CL234Contig1-156	Longitude	T/T	1	-0.111	0.211	0.278	0.598
Pm_CL234Contig1-156	A/T	Intercept	1	-0.512	0.089	32.965	<.001	Pm_CL234Contig1-156	PCA1	A/T	1	0.100	0.024	17.049	<.001
Pm_CL234Contig1-156	T/T	Intercept	1	-2.179	0.171	162.860	<.001	Pm_CL234Contig1-156	PCA1	T/T	1	0.010	0.044	0.053	0.818
Pm_CL783Contig1-212	G/G	Intercept	1	1.392	0.101	191.024	<.001	Pm_CL783Contig1-212	PCA1	G/G	1	0.090	0.026	12.547	<.001
sSPcDFD040B03103-274	G/G	Intercept	1	23.674	5.638	17.633	<.001	sSPcDFD040B03103-274	Latitude	G/G	1	-0.448	0.122	13.417	<.001
sSPcDFD040B03103-274	G/G	Intercept	1	3.595	0.262	187.690	<.001	sSPcDFD040B03103-274	PCA3	G/G	1	-0.913	0.240	14.497	<.001
sSPcDFE049E11411-220	A/G	Intercept	1	55.480	14.105	15.471	<.001	sSPcDFE049E11411-220	Longitude	A/G	1	0.459	0.115	15.951	<.001
sSPcDFE049E11411-220	G/G	Intercept	1	-16.023	25.156	0.406	0.524	sSPcDFE049E11411-220	Longitude	G/G	1	-0.112	0.205	0.301	0.583
tbe-1259	G/G	Intercept	1	-45.414	13.729	10.943	<.001	tbe-1259	Longitude	G/G	1	-0.376	0.112	11.323	<.001

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Identification of cold-hardiness related genes in coastal Douglas fir (<i>Pseudotsuga menziesii</i>)

Richting: **Master of Statistics-Bioinformatics**
Jaar: **2014**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Vangestel, Carl

Datum: **5/02/2014**