MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 4216–4233 doi: 10.1111/mec.13750

Finding needles in a genomic haystack: targeted capture identifies clear signatures of selection in a nonmodel plant species

MATTHEW J. CHRISTMAS,* ED BIFFIN,† MARTIN F. BREED* and ANDREW J. LOWE*

*Environment Institute and School of Biological Sciences, The University of Adelaide, North Terrace, SA 5005, Australia, †State Herbarium of South Australia, Hackney Road, Adelaide, SA 5000, Australia

Abstract

Teasing apart neutral and adaptive genomic processes and identifying loci that are targets of selection can be difficult, particularly for nonmodel species that lack a reference genome. However, identifying such loci and the factors driving selection have the potential to greatly assist conservation and restoration practices, especially for the management of species in the face of contemporary and future climate change. Here, we focus on assessing adaptive genomic variation within a nonmodel plant species, the narrow-leaf hopbush (*Dodonaea viscosa* ssp. *angustissima*), commonly used for restoration in Australia. We used a hybrid-capture target enrichment approach to selectively sequence 970 genes across 17 populations along a latitudinal gradient from 30°S to 36°S. We analysed 8462 single-nucleotide polymorphisms (SNPs) for F_{ST} outliers as well as associations with environmental variables. Using three different methods, we found 55 SNPs with significant correlations to temperature and water availability, and 38 SNPs to elevation. Genes containing SNPs identified as under environmental selection were diverse, including aquaporin and abscisic acid genes, as well as genes with ontologies relating to responses to environmental stressors such as water deprivation and salt stress. Redundancy analysis demonstrated that only a small proportion of the total genetic variance was explained by environmental variables. We demonstrate that selection has led to clines in allele frequencies in a number of functional genes, including those linked to leaf shape and stomatal variation, which have been previously observed to vary along the sampled environmental cline. Using our approach, gene regions subject to environmental selection can be readily identified for nonmodel organisms.

Keywords: *Dodonaea viscosa*, F_{ST}, gene–environment associations, hybrid-capture, shrub

Received 31 January 2016; revision received 27 June 2016; accepted 6 July 2016

Introduction

Understanding the genetic basis of adaptation and uncovering the drivers of selection are two of the key pursuits in evolutionary and ecological genomics (Savolainen *et al.* 2013). Biologists are now better placed than ever to generate data to explore these phenomena as next-generation sequencing has enabled the production of genome-scale data, including for nonmodel organisms (Davey *et al.* 2011). New tools are regularly being developed to make best use of the vast quantities of sequence data that are now easily generated for addressing questions of selection and adaptation (Rellstab *et al.* 2015).

Revealing loci under selection as well as identifying the main environmental drivers of selection can have direct applications to conservation and restoration (Hoffmann *et al.* 2015; Shafer *et al.* 2015), particularly at a time when global climate change is shifting selection pressures on natural populations. The flow of adaptive alleles across landscapes may assist adaptation as local

Correspondence: Andrew J. Lowe, Fax: +61 8 8303 4364; E-mail: andrew.lowe@adelaide.edu.au

genotypes become more maladapted under a changing climate (Christmas *et al.* 2015a). Conservation and restoration practices, such as assisted gene flow and migration, and the establishment of corridors to connect fragmented landscapes, could use this information to identify which populations hold the most adaptive potential and which populations may benefit from the introduction of adaptive alleles.

Numerous recent studies have attempted to identify signatures of environment-based selection by scanning the genome (e.g. Bonin *et al.* 2006; Namroud *et al.* 2008; Keller *et al.* 2011; Prunier *et al.* 2011; Chen *et al.* 2012; Tsumura *et al.* 2012; Chavez-Galarza *et al.* 2013; Steane *et al.* 2014). Whilst this pursuit is not futile, the chances of discovering such signatures are slim as the majority of the genome is thought to be nonfunctional (for example, as little as 8.2% of the human genome is thought to be functional; Rands *et al.* 2014), the parts that are functional are not necessarily related to environmental variation (Meirmans 2015), and functional traits are thought to be mostly polygenic (i.e. affected by many genes of small effect; Rockman 2012). For example, in a genome-wide association study into climate adaptation in *Arabidopsis thaliana*, only 0.002% of 213 248 single-nucleotide polymorphisms (SNPs) were shown to associate with fitness (Fournier-Level *et al.* 2011). Despite this, genotype–environment associations are often found in genome scan studies such as those referenced above (although many of these may be false positives; Meirmans 2015), and the models used to discover them are continually improving. Environmental association analyses, which seek genetic variants that correlate with environmental factors, can reveal significant adaptive loci as well as the drivers of selection (Rellstab *et al.* 2015).

Population genetic structure arising from neutral processes such as mutation, genetic drift and gene flow needs to be taken into account in gene–environment association analyses as it can mimic patterns expected under non-neutral processes (Excoffier *et al.* 2009). In particular, range expansions and isolation by distance (IBD) generally play important roles in shaping population genetic structure, with increased distance between individuals correlating with increased genetic dissimilarity (Wright 1943). Over a latitudinal gradient, IBD and range expansions can result in neutral genetic variation exhibiting similar patterns to those expected for loci under selection (Rellstab *et al.* 2015). When testing for signatures of selection, such patterns could result in large numbers of false positives (Meirmans 2012). Outlier detection methods that look for loci with values of F_{ST} that are higher or lower than neutral expectations do not take this spatial aspect into account (Narum & Hess 2011). However, more recent environment association analysis methods can incorporate population genetic structure and/or spatial factors into their models to account for neutral genetic structure (Coop *et al.* 2010; Frichot *et al.* 2013; Guillot *et al.* 2014; Rellstab *et al.* 2015).

Many methods are currently available to generate genome-wide data to investigate selection (Ekblom & Galindo 2010; Mamanova *et al.* 2010; Stapley *et al.* 2010; Davey *et al.* 2011; Peterson *et al.* 2012). Ideally, a high-quality, annotated reference genome would be available. Reference genomes assist in locating and assigning function to candidate loci, reducing false-positive rates and improving statistical power (Manel *et al.* 2016). However, the cost of generating a quality reference genome is still prohibitive for most studies of nonmodel organisms. A common alternative has been to call SNPs from restriction enzyme-based reduced-representation libraries and then look for F_{ST} outliers and/or correlations between alleles and environmental factors (e.g. Steane *et al.* 2014; Guo *et al.* 2015). This approach holds promise for identifying regions of the genome that are under selection as it results in the generation of millions of sequences distributed throughout the genome and so, even if the majority of the genome is nonfunctional and not under selection, such genome-wide screening often does identify signatures of selection. However, in the absence of a reference genome, the location and potential functional importance of any significant variant is unknown and so progression from these initial screening studies to studies that demonstrate the adaptive importance of genetic variation is difficult.

An alternative genome-partitioning method that is useful with or without a reference genome is the targeted capture and sequencing of specific genomic regions using hybrid-capture probes (Mamanova *et al.* 2010). Advantages of this over alternative genome-partitioning methods, such as those involving restriction enzymes, are a reduction in variance in target coverage, increased accuracy of SNP calls and greater reproducibility across samples (Jones & Good 2016). Also, if capture probes are designed on annotated transcriptome sequences, for example, then functional information of sequences can be known even for species without a reference genome.

Here, we used a targeted approach to identify signatures of selection in the nonmodel plant *Dodonaea viscosa* ssp. *angustissima* (hereafter *D. v. angustissima*) in the absence of a reference genome. *Dodonaea viscosa* ssp. *angustissima* is a woody shrub that grows up to 4 m, is widely distributed throughout central and southern Australia and inhabits a diversity of habitats and environments from sandy loams to rocky outcrops on hillsides and from arid to temperate areas. It is therefore an interesting species to study in terms of adaptation to its environment and the processes by which it has

achieved this adaptability. It is also widely used in restoration projects around Australia yet little is known about, for example, levels of local adaptation and the potential consequences of moving seed across the landscape. This study provides insight into the adaptive processes that have acted on this species and the findings we present could be used to inform its use in restoration.

Previous work on *D. v. angustissima* has demonstrated clinal leaf width variation over space and time, with a pattern of narrowing leaves associated with warmer climatic conditions (Guerin & Lowe 2012; Guerin *et al.* 2012). In this study, we used a recently published annotated *D. v. angustissima* transcriptome (Christmas *et al.* 2015b) to design hybrid-capture probes for the selective sequencing of a set of candidate genes (Mamanova *et al.* 2010). To increase the chances of uncovering signatures of selection, we chose candidate genes with putative functions and, therefore, we have the expectation that their products may be involved with traits under selection. For example, we targeted aquaporin and abscisic acid (ABA)-related genes, which have been shown to be involved in water use efficiency in plants (Tyerman *et al.* 2002; Kaldenhoff *et al.* 2008). We focussed on a clinal distribution of *D. v. angustissima* in South Australia and analysed targeted gene sequence data from 17 sampling sites using a combination of F_{ST} outlier analysis, genotype–environment association analysis and redundancy analysis (RDA) to provide evidence towards answering two main questions: (i) to what extent has environmental variation among our target populations driven selection, leading to gene–environment correlations in functional genes? and (ii) how much of the genetic variation can be accounted for by variation in environmental factors?

Materials and methods

Sampling

We sampled *D. v. angustissima* along a ~700-km environmental gradient throughout the Adelaide geosyncline region, with sampling effort stretching from Kangaroo Island in the south, through the Mount Lofty and Flinders Ranges to the Gammon Ranges in the north (Fig. 1). Leaf samples were collected from 5 to 8 individual plants from each of 17 naturally occurring populations (a total of 89 individuals) (Table 1). Samples were collected from plants at least 10 m apart that were not direct neighbours where possible to minimize the chances of sampling close relatives. Leaf samples were stored in teabags on silica gel prior to DNA extraction.

Environmental data

We sampled across a steep temperature and rainfall gradient, where mean annual maximum temperature ranged from 19.5 to 28.5 °C and mean annual precipitation from 467 to 134 mm in the most southern and most northern sites, respectively. To summarize and reduce redundancy among the many environmental variables that vary across the region, we ran a principle component analysis (PCA) on fifteen environmental parameters downloaded from the Atlas of Living Australia (http://www.ala.org.au, accessed 20 September 2015; Table S1, Supporting information) using the PCA function in the R package FACTOMINER, after having first standardized all parameters with the SCALE function. The first two principle components (PCs) accounted for 76% of the variance in the data and were selected for use in the genotype–environment analyses. In PC1 (57.7% of variance), maximum mean temperature, evaporation, vapour pressure deficit and radiation were strong positive correlates with loadings over 0.8, and organic carbon, water stress index, precipitation, aridity and humidity were strong negative correlates with loadings under −0.8. For PC2 (18.4% of variance), only site elevation showed a loading over 0.8. The contributions of each environmental variable to the reduced PCs are presented in Table 2. PC1 correlated positively with latitude (Pearson's $r = 0.67$, $P = 0.003$), whilst PC2 did not (Pearson's $r = -0.34$, $P = 0.19$).

Candidate genes and genotyping

We used the published *D. viscosa* transcriptome (Christmas *et al.* 2015b) to design MYbaits hybrid-capture baits (MYcroarray, Michigan, USA) for the capture of a set of 970 genes. 80mer baits with a 2× tiling density were designed and produced at MYcroarray, Michigan, USA. We targeted genes that would increase our chances of finding signatures of selection. For example, the products of aquaporin and ABA genes are involved in water use efficiency in plants and the presence of putatively adaptive variants have been demonstrated in previous studies (Tyerman *et al.* 2002; Kaldenhoff *et al.* 2008; Audigeos *et al.* 2010). This led to a priori expectations that these genes may be targets of selection along our aridity gradient. Within our target gene set, 45 genes were 'aquaporin' or 'ABA'-related genes. A further 308 targeted genes had been assigned the gene ontology (GO) term 'response to water deprivation' in the transcriptome annotation (Christmas *et al.* 2015b) and, again, we had a priori expectations that these genes may be under selection. The remaining 617 genes in the target gene set comprised genes shown to contain nonsynonymous mutations among two *D. viscosa*

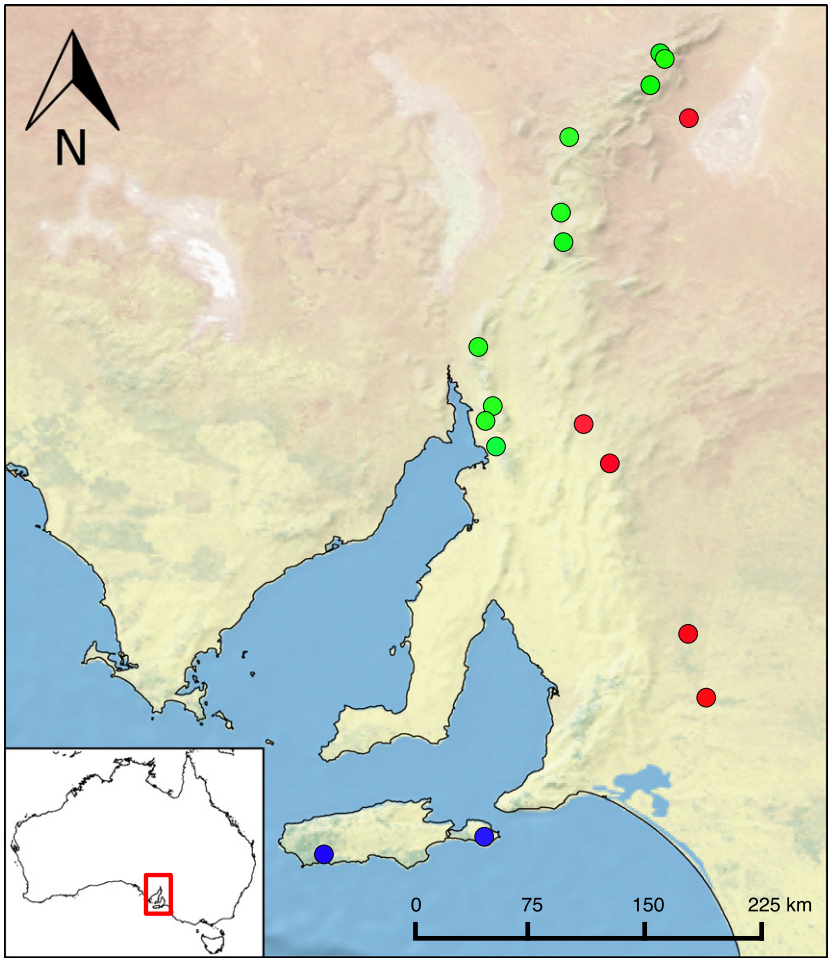

Fig. 1 Sampling locations along the Adelaide Geosyncline Region where leaf samples from 5 to 8 individuals were collected for DNA extraction from each of 17 sites. Colours of dots represent population assignment to one of three genetic clusters identified in both STRUCTURE and DAPC analysis (blue = Kangaroo Island cluster; red = Eastern cluster; green = Flinders Ranges cluster).

subspecies (ssp. *angustissima* and ssp. *spatulata*) in Christmas *et al.* (2015b). Mutations resulting in amino acid changes may lead to functional changes in the gene products and, if so, be likely targets of selection.

DNA was extracted using the Macherey-Nagel Nucleospin Plant II Kit at the Australian Genome Research Facility (AGRF, Adelaide, Australia) and then sonicated for random sheering. Illumina's TruSeq Nano DNA protocol was used for size selection, and sequencing adapter and barcode ligation. The hybrid-capture enrichment reactions were carried out following the MYbaits protocol (v.2.3.1; MYcroarray) with 12 cycles of postcapture PCR. One hundred base pair dual indexed paired-end reads were sequenced on the Illumina HiSeq 2000 platform at AGRF (Melbourne, Australia) and processed using the ILLUMINA CASAVA pipeline (version 1.8.2). Read quality was assessed using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and all high-quality reads were mapped to the reference transcriptome (Christmas *et al.* 2015b) using Burrows–Wheeler Aligner (BWA; Li & Durbin 2009). The indexed reference was created using default settings. Resulting SAM files were compressed and sorted by reference contig, and duplicated sequences were marked using PICARD tools (http://broadinstitute.github.io/picard/). Mapping characteristics were assessed using SAMTOOLS (Li *et al.*, 2009). The SAMTOOLS utility 'mpileup' was used to call variant sites per individual and variants were output as genotype probabilities in the VCF format. Output VCF files were then merged and a custom script was used to convert variant calls from genotype probabilities to genotype calls.

VCFTOOLS (Danecek *et al.* 2011) was used to filter variants with the following cut-offs: minimum depth of 10 reads, minor allele frequency >10%, missing data per SNP <25%. Contigs containing fewer than 10 base pairs per SNP were removed to control for mapping errors. Linkage disequilibrium (LD) among SNPs was accounted for using the LD pruning tool in PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/), where independent pairwise comparisons between each SNP were made and an r^2 value was calculated. A cut-off of $r^2 > 0.5$ was used, whereby one of a pair of SNPs was removed from the data set if the coefficient of

Table 1 *Dodonaea viscosa* ssp. *angustissima* sample coordinates at 17 sites along a latitudinal gradient in South Australia. Site values for the first two principal components of a principal component analysis of 15 environmental variables are also shown

Collection site	Individuals per population	Latitude	Longitude	ENVPC1	ENVPC2
Kangaroo Island East	5	−35.85	138.02	−5.687	1.478
Kangaroo Island West	5	−35.98	136.86	−6.486	0.172
Peterborough	5	−33.15	138.93	−1.145	−0.696
Orroroo	5	−32.86	138.74	−1.119	0.502
Brookfield	8	−34.38	139.49	−0.606	1.727
Monarto	5	−34.84	139.62	−1.138	1.011
Brachina Gorge	5	−31.33	138.57	2.555	−0.043
Wilpena Pound	5	−31.55	138.59	0.551	−3.035
Dutchmans Stern	5	−32.30	137.97	−0.166	−0.778
Mambray Creek	5	−32.84	138.03	−0.341	1.616
Telowie Gorge	5	−33.02	138.10	−1.448	0.529
Alligator Gorge	5	−32.73	138.08	−2.217	−2.130
Gammon Ranges 1	5	−30.22	139.32	3.071	−1.444
Gammon Ranges 2	5	−30.79	138.63	3.039	−0.505
Gammon Ranges 3	6	−30.18	139.29	3.415	−0.795
Gammon Ranges 4	5	−30.65	139.50	5.310	4.366
Gammon Ranges 5	5	−30.41	139.22	2.411	−1.975

Table 2 Loadings for each of the 15 environmental variables included in the principal component analysis for both the first and second principal components (PC1 and PC2). The correlation between a component and a variable estimates the information they share. Values >0.8 and <−0.8 show a strong relationship between the variable and the component and are highlighted grey

Environmental variable	Loading (PC1)	Loading (PC2)
Elevation (m)	0.37	−0.87
Soil depth (m)	−0.57	0.57
Organic carbon	−0.89	−0.18
Nutrient status	−0.23	0.47
Water stress index, annual mean	−0.97	−0.01
Moisture variability	0.68	0.46
Precipitation, annual mean (cm)	−0.86	−0.21
Temperature, annual max mean (°C)	0.89	0.30
Temperature, annual min mean (°C)	0.07	0.61
Evaporation, annual mean (mm)	0.89	−0.27
Aridity index, annual mean	−0.96	−0.03
Vapour pressure deficit, annual mean (KPa)	0.98	0.06
Humidity, annual relative mean	−0.97	0.21
Radiation, annual mean ($MJ/m^2/day$)	0.96	−0.18
Runoff, average, (megalitres/ 5 × 5 km/year)	0.02	−0.73

determination between the pair was >0.5, thus removing SNPs showing strong signals of LD. Following these filtering steps, 8462 SNPs remained for downstream analysis.

Neutral genetic structure

Neutral population genetic structure among our samples was assessed using a putatively neutral subset of the main SNP data set (described above). This was generated by removing outlier SNPs identified by BAYESCAN v.2.0 (Foll & Gaggiotti 2008) with prior odds = 10 and a false discovery rate (FDR) of 0.05, resulting in a conservative output of neutral SNPs. The resultant SNP set was filtered to reduce LD using the same method described above and was thinned to ensure at least 100 bp between each SNP. One SNP per transcriptome-reference contig (after Christmas *et al.* 2015b) was selected at random to further reduce linkage within the data set, giving a final set of 659 neutral SNPs. This SNP set was analysed for population genetic structure using the model-based clustering program STRUCTURE v.2.3.4 (Pritchard *et al.* 2000) as well as a nonmodel-based multivariate approach, which seeks discriminating functions between groups of individuals whilst minimizing variation within clusters, DAPC (Jombart *et al.* 2010).

We ran STRUCTURE using the admixture and correlated allele frequency models, the model parameter α was inferred from the data, a burn-in of 200 000 followed by 1 000 000 iterations, for K values of 1–10 with ten replicates per K value. STRUCTURE HARVESTER v.0.6.94 (Earl & vonHoldt 2012) was used to calculate ΔK to assess the most likely value of K and then results from replicate runs for this K value were combined using CLUMPP (Jakobsson & Rosenberg 2007) with default settings.

For DAPC, genetic data were first transformed into uncorrelated components using PCA. The number of genetic clusters was then defined using k-means, a clustering algorithm that looks for the value of K that maximizes the variation between groups. The Bayesian Information Criterion (BIC) was calculated for K = 1–10 and the K value with the lowest BIC was selected as the optimal number of clusters. A discriminant analysis was then performed on the first 40 principal components using the function DAPC, implemented in R, to efficiently describe the genetic clusters.

Detection of adaptive genetic variation

We employed three methods to identify signatures of selection in candidate genes that may reflect adaptive variation among populations of *D. v. angustissima* across the study region. Such methods are susceptible to high FDRs (De Mita *et al.* 2013; Meirmans 2015). For example, IBD may confound genotype–environment associations as neutral genetic variation is often spatially autocorrelated (Meirmans 2012). In addition, geographically proximate populations often share similar environments, potentially leading to autocorrelation of genetic and environmental variation (Dillon *et al.* 2014). A general recommendation in the literature is to run multiple methods and compare outputs (De Mita *et al.* 2013; Villemereuil *et al.* 2014; Lotterhos & Whitlock 2015; Rellstab *et al.* 2015), although this may lead to loci under weak selection being overlooked (i.e. false negatives; Lotterhos & Whitlock 2015). Villemereuil *et al.* (2014) found that error rates could be greatly reduced by considering the outputs of BAYESCAN, LFMM and BAYENV2 together when identifying loci displaying selection signatures. We therefore compared the outputs of these three methods, two of which (LFMM and BAYENV2) take neutral population structure into account. We focussed on SNPs that demonstrated significant relationships across at least two of the three methods to conservatively identify those SNPs that consistently showed signatures of selection.

Outlier detection. The 8462 SNPs were analysed in the F_{ST}-based outlier analysis software BAYESCAN v. 2.0 (Foll & Gaggiotti 2008) to identify outlier SNPs, that is SNPs that demonstrated stronger or weaker differentiation among populations than would be expected under neutral expectations. Such SNPs may be demonstrating signatures of diversifying or stabilizing selection, respectively. This approach can result in the detection of many false positives, particularly in species that exhibit IBD or have undergone range expansion (Lotterhos & Whitlock 2014), which is a possibility for our samples. We therefore ran BAYESCAN multiple times with prior odds of neutrality of 10, 100, 1000 and 10 000 to assess which was most appropriate for the number of SNPs, giving sufficient control of the false-positive rate whilst not being too conservative as to result in a high false-negative rate (Lotterhos & Whitlock 2014). A FDR of 0.05 was used.

Genotype–environment analysis. We used two Bayesian methods that seek associations between allelic variation and environment whilst taking into account potentially confounding population genetic structure: BAYENV2, which implements a generalized linear regression model (http://gcbias.org/bayenv/; Coop *et al.* 2010), and LFMM, which implements a linear mixed model (http://membres-timc.imag.fr/Olivier.Francois/lfmm/index.htm; Frichot *et al.* 2013).

In BAYENV2, the set of 659 putatively neutral SNPs used for neutral population genetic analysis were used to estimate the empirical pattern of covariance in allele frequencies among the populations, given a covariance matrix. 100 000 MCMC steps were used and the covariance matrix was estimated independently five times to ensure convergence. The output matrix was then used as a null model against which each SNP from the full data set was tested. For each SNP, Bayes factors were provided as a measure of support for a model where the environmental variable (PC1 or PC2) has a linear effect on the transformed allele frequencies compared with a model given by the covariance matrix alone (Coop *et al.* 2010). As the program implements MCMC algorithms, stochastic error can lead to large run-to-run variation and therefore Bayes factors should be averaged over multiple runs (Rellstab *et al.* 2015). Bayes factors were calculated for each SNP using 100 000 steps in each of eight independent runs against both environmental factors and then averaged across runs for each SNP. Strength of evidence for significant associations was based on the value of the $\log_{10}$ Bayes factor ($\log_{10}$BF), with the following $\log_{10}$BF cut-offs: 0.5–1 = substantial evidence; 1–2 = strong evidence; >2 = decisive (Kass & Raftery 1995). The linear model underlying the Bayes factor might not be correct or outliers within our data might misguide the model (BAYENV2.0 manual, https://bitbucket.org/tguenther/bayenv2_public/src). To deal with this, BAYENV2 also calculates the nonparametric Spearman's rank correlation coefficient, ρ. SNPs with a $\log_{10}$BF >0.5 as well as an absolute value of $\rho > 0.3$ (where ρ ranges from −1 to 1) were therefore considered as robust candidates demonstrating signatures of selection.

Latent factor mixed models (LFMMS) were used to test for associations between loci and environmental variables. We implemented an MCMC algorithm for regression analysis whereby potentially confounding

population structure is modelled with unobserved (latent) factors (Frichot *et al.* 2013). The number of latent factors was set at $K = 3$ based on the identification of three distinct genetic clusters in the population genetics analysis. The MCMC algorithm was implemented for each of the two environmental variables (i.e. PC1 and PC2), using 50 000 steps for burn-in and 100 000 additional steps to compute LFMM parameters (z-scores) for all loci. Due to run-to-run variation, the analysis was repeated over ten independent runs and z-scores across runs were then combined in R using the LEA package (Frichot & François 2015). The LEA package was also used for the adjustment of *P*-values for multiple testing using the Benjamini–Hochberg procedure and the calculation of the genomic inflation factor to modify z-scores and allow for the control of the FDR, as described in Frichot & François (2015). A list of candidate loci with an FDR of 1% and adjusted *P*-values of <0.001 was then generated for each environmental variable. Histograms of adjusted *P*-values are included in the supporting information (Figs S1 and S2, Supporting information).

Redundancy analysis. We used a redundancy analysis (RDA) approach to investigate which variables were most important in explaining the genetic variation in our data, to complement the identification of specific loci under selection. Here, RDA is used to effectively estimate the genetic variance components associated with PC1 and PC2, as well as with a spatial component. RDA was performed in R using the VEGAN package (Oksanen *et al.* 2009), using a modified version of the R script provided in the supplementary material of Meirmans (2015). A forward selection procedure using the STEP function in VEGAN and an alpha value of 0.01 was used to determine which spatial variables to include in the RDA. One spatial variable was retained: y^3. Among population variation was partitioned into pure spatial, pure PC1, pure PC2 and overlapping components using the VARPART function. The significance of the partitioning was tested using the ANOVA.CCA function with 999 permutations. As RDA is a combination of PCA and multiple regression, and the fact that total variation of a PCA on allele frequencies is equivalent to F_{ST} (McVean 2009), the percentage of the total genetic variation that is explained by the spatial and environmental variables combined was calculated by multiplying the proportion of constrained variation by the value of global F_{ST} across all samples (0.102).

Results

Outlier detection

With prior odds of 10, 100, 1000 and 10 000, BAYESCAN identified 880, 200, 74 and 24 outlier SNPs, respectively. To control for false positives whilst also taking into account the fact that our targeted sequencing approach had potentially enriched for genes under selection, we proceeded with the results for when prior odds were set at 100. This gave a total of 200 outliers from the 8462 input SNPs (2.36%), with F_{ST} values ranging from 0.265 to 0.518 (Fig. S3, Supporting information).

Neutral genetic structure

Both STRUCTURE and DAPC outputs agreed that the most likely value of *K* was three (Fig. S4, Supporting information), with all but one of the individuals assigned to a cluster with >90% assignment across both methods. The identified neutral genetic structure was taken into account in the subsequent genotype–environment association analyses to reduce the chances of falsely associating genetic differences due to neutral processes with environmental variables (i.e. false positives).

Genotype–environment analysis

BAYENV2 identified 170 (2.01%) and 101 (1.19%) SNPs with significant correlations ($\log_{10}(BF) > 0.5$, absolute $\rho > 0.3$) to PC1 and PC2, respectively (Fig. 2). LFMM identified 1011 (11.95%) and 919 (10.86%) SNPs with significant correlations (corrected $P < 0.001$) to PC1 and PC2, respectively. For PC1, seven SNPs were identified as significant outliers by all three methods, 27 SNPs correlated with PC1 in both BAYENV2 and LFMM analyses but not identified as outliers by BAYESCAN, 15 SNPs were common between LFMM and BAYESCAN but not BAYENV2, and six SNPs were identified by both BAYENV2 and BAYESCAN but not LFMM (Fig. 3a). This gave a total of 55 SNPs distributed over 47 genes that were identified by at least two of the three methods (from now on termed 'significant PC1 SNPs'; Table 3).

For PC2, five SNPs were identified as outliers by all three methods, 10 showed significant correlations in both the BAYENV2 and LFMM analyses but were not identified as outliers by BAYESCAN, 19 were identified as outliers by both LFMM and BAYESCAN but not BAYENV2 and four were common among BAYENV2 and BAYESCAN but not LFMM (Fig. 3b). These 38 SNPs were distributed over 34 genes (from now on termed 'significant PC2 SNPs'; Table 4).

Significant SNP gene functions

BLAST searches of assembled transcriptome contigs to the NCBI nonredundant protein database followed by GO annotation in Christmas *et al.* (2015b) meant that all targeted sequences in the current study had putative gene product names and functions associated with them. All

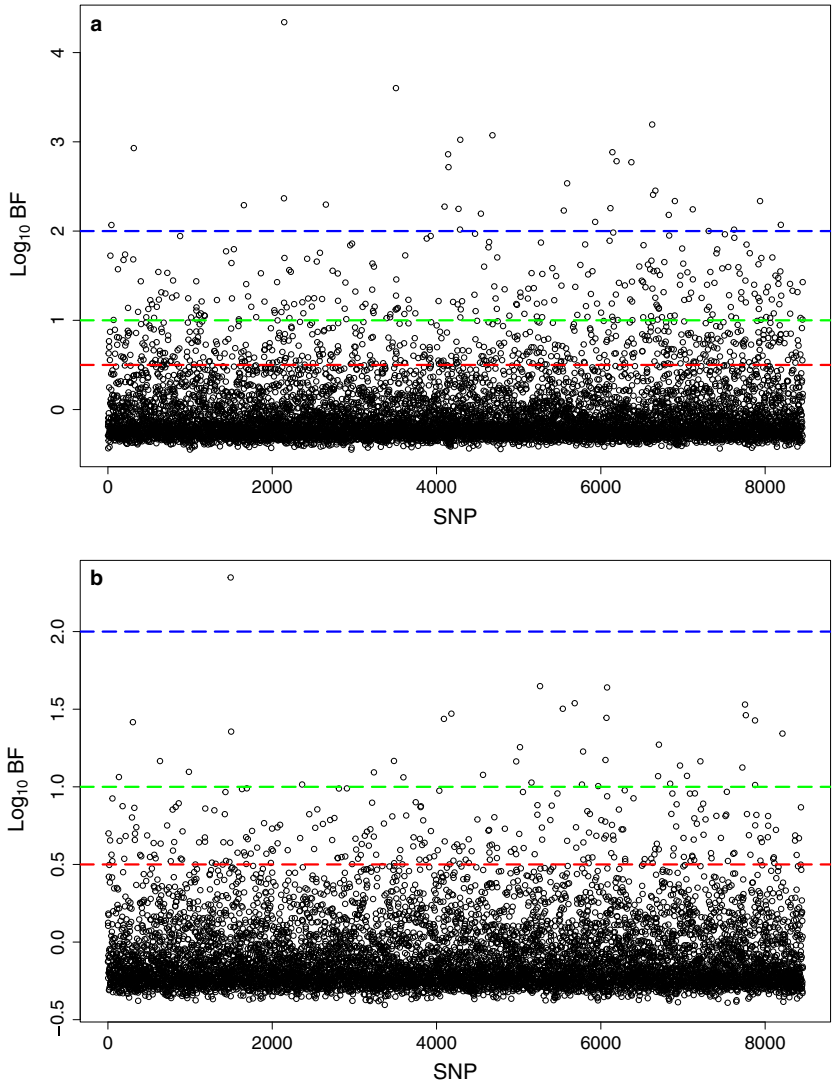

Fig. 2 Manhattan plots of genetic differentiation associated with environmental parameters. Log_{10} Bayes Factors (BF) from BAYENV2 output represent associations between single-nucleotide polymorphism (SNP) allele frequency variation with variation in (a) PC1 and (b) PC2, the first two principal components from a principal component analysis of environmental variables. Red dashed lines represent the lower threshold of $log_{10}(BF) = 0.5$ (substantial evidence), green dashed lines represent the threshold of $log_{10}(BF) = 1$ (strong evidence), and blue dotted lines represent the higher threshold of $log_{10}(BF) = 2$ (decisive evidence). SNP number is arbitrary as genome and/or chromosome locations are unknown.

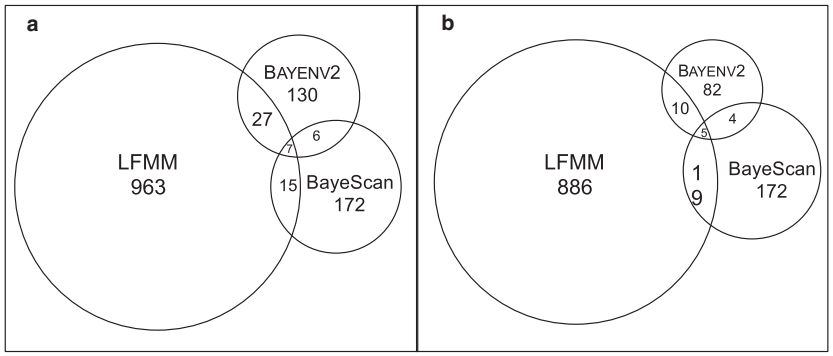

Fig. 3 Venn diagrams comparing the significant outputs from an outlier detection method (BAYESCAN) and two genotype–environment analysis methods (LFMM and BAYENV2), which look for significant correlations between single-nucleotide polymorphisms (SNPs) and (a) PC1 and (b) PC2 from a principal component analysis of environmental variables. Numbers indicate the number of significant SNPs identified by each method and their overlap with each of the other methods.

significant SNPs could therefore be related back to the gene in which they were found, in order for their putative adaptive function to be assessed. Details of all genes containing significant SNPs along with their product names and related GO terms can be found in Table S2 (Supporting information).

Of the 45 targeted aquaporin and ABA genes, two aquaporin and two ABA-related genes were found to

Table 3 SNPs identified as significant outliers by at least two of the three methods used: LFMM, BAYENV2 and BAYESCAN, along with the output statistics for each method. SNPs identified by LFMM and BAYENV2 show significant correlations with the first principal component (PC1) from a PCA of environmental variables. The column 'Methods' states by which of the three methods the SNP was found to be a significant outlier; BS = BAYESCAN, BE = BAYENV2. BF = Bayes Factor

		LFMM		BAYENV2		BAYESCAN		
SNP ID	Sequence description	Z-scores	*P*	$\log_{10}$(BF)	Spearman's ρ	$\log_{10}$(PO)	F_{ST}	Methods
Contig_30496_547	S-adenosyl-l-methionine-dependent methyltransferases superfamily protein isoform 1	−2.106	1.10E-04	3.602	−0.375	1.277	0.328	BS, BE, LFMM
Contig_15281_310	Syntaxin-121-like	1.847	6.96E-04	2.783	−0.351	1.962	0.342	BS, BE, LFMM
Contig_3074_517	Calcium-dependent protein kinase 13-like	3.024	2.81E-08	2.535	−0.402	1.708	0.332	BS, BE, LFMM
Contig_11820_850	Cytochrome p450 76a2-like	2.152	7.76E-05	1.699	−0.341	2.062	0.334	BS, BE, LFMM
Contig_6523_1180	Probable protein phosphatase 2c 27-like	1.779	1.09E-03	1.033	−0.405	2.795	0.378	BS, BE, LFMM
Contig_1257_1000	Alpha-glucan h isozyme-like	2.066	1.48E-04	0.780	−0.379	3.699	0.383	BS, BE, LFMM
Contig_72674_282	Serine–threonine protein plant-	−6.086	5.38E-29	0.621	−0.375	0.647	0.310	BS, BE, LFMM
Contig_11820_647	Cytochrome p450 76a2-like	2.436	7.73E-06	4.341	−0.428	−1.235	0.131	BE, LFMM
Contig_8437_1550	Programmed cell death protein 2-like	4.617	2.30E-17	3.073	−0.533	−0.644	0.158	BE, LFMM
Contig_6523_850	Probable protein phosphatase 2c 27-like	−1.914	4.40E-04	3.023	−0.358	0.004	0.218	BE, LFMM
Contig_51946_182	Probable carboxylesterase 18-like	3.322	1.06E-09	2.861	−0.565	−1.591	0.126	BE, LFMM
Contig_51946_405	Probable carboxylesterase 18-like	2.028	1.96E-04	2.716	−0.413	−1.916	0.123	BE, LFMM
Contig_11820_539	Cytochrome p450 76a2-like	−2.705	6.83E-07	2.367	−0.378	−1.290	0.130	BE, LFMM
Contig_36495_436	Ubx domain-containing protein	−3.672	1.56E-11	2.336	−0.339	−2.062	0.123	BE, LFMM
Contig_20553_1476	Pre-rRNA-processing protein tsr2 homolog	−2.019	2.09E-04	2.297	0.412	−1.602	0.125	BE, LFMM
Contig_14218_2104	Ethylene insensitive 3-like 3	−1.922	4.17E-04	2.256	−0.334	−2.083	0.123	BE, LFMM
Contig_23479_38	Nucleotide-diphospho-sugar transferase family protein	2.127	9.42E-05	1.704	−0.311	−1.528	0.120	BE, LFMM
Contig_57823_266	Ethylene-responsive transcription factor win1-like	2.413	9.39E-06	1.542	−0.430	−1.930	0.123	BE, LFMM
Contig_29525_169	F-box family protein	3.824	2.20E-12	1.508	0.303	−0.100	0.080	BE, LFMM
Contig_34790_123	Cytochrome p450	2.538	3.16E-06	1.501	−0.322	−1.848	0.122	BE, LFMM
Contig_9579_170	Protein	1.887	5.31E-04	1.475	−0.360	−2.116	0.123	BE, LFMM
Contig_35330_278	Protein	3.999	2.09E-13	1.437	−0.353	−2.004	0.122	BE, LFMM
Contig_25203_676	Rust resistance kinase lr10	2.104	1.12E-04	1.437	−0.376	−2.083	0.123	BE, LFMM
Contig_83_327	(-)-germacrene d synthase	3.400	4.32E-10	1.244	−0.412	−1.029	0.143	BE, LFMM
Contig_57823_398	Ethylene-responsive transcription factor win1-like	5.267	4.00E-22	1.232	−0.343	−1.773	0.124	BE, LFMM
Contig_43658_217	Ankyrin repeat family	1.887	5.31E-04	0.810	−0.367	−0.340	0.179	BE, LFMM
Contig_35330_1309	Protein	5.644	3.61E-25	0.765	−0.304	−1.778	0.121	BE, LFMM
Contig_20305_129	Disease resistance protein rpp13-like	−2.601	1.79E-06	0.727	−0.371	−1.954	0.123	BE, LFMM
Contig_36175_358	E3 ubiquitin–protein ligase chip-like	−2.472	5.63E-06	0.674	−0.318	−2.128	0.122	BE, LFMM
Contig_21273_330	Protein	1.910	4.53E-04	0.619	−0.349	−2.178	0.123	BE, LFMM
Contig_79879_574	Serine carboxypeptidase-like 51-like	5.360	7.41E-23	0.604	−0.328	−2.234	0.123	BE, LFMM
Contig_5640_75	Probable peptide nitrate transporter at1g22540-like	−1.972	2.95E-04	0.595	−0.344	−1.690	0.121	BE, LFMM
Contig_5183_755	Predicted protein	−2.450	6.83E-06	0.549	−0.358	−1.946	0.122	BE, LFMM
Contig_25237_199	Carnitine racemase like protein	2.619	1.52E-06	0.507	−0.331	−1.987	0.123	BE, LFMM
Contig_8825_830	ABC transporter g family member 22-like	−2.613	1.60E-06	1.705	−0.091	0.695	0.296	BS, LFMM
Contig_10542_747	DNA polymerase zeta subunit	−1.915	4.39E-04	1.113	−0.246	1000.000	0.359	BS, LFMM
Contig_11322_346	Aquaporin nip1-2	2.996	3.76E-08	0.910	−0.269	2.317	0.367	BS, LFMM
Contig_8314_143	Tir-nbs-lrr resistance protein	−2.692	7.70E-07	0.644	−0.024	3.699	0.365	BS, LFMM

Table 3 *Continued*

		LFMM		BAYENV2		BAYESCAN		
SNP ID	Sequence description	*Z*-scores	*P*	$\log_{10}$(BF)	Spearman's ρ	$\log_{10}$(PO)	F_{ST}	Methods
Contig_2046_388	Abscisic acid receptor pyl9-like	1.898	4.92E-04	0.481	−0.397	2.522	0.364	BS, LFMM
Contig_6331_1135	Peroxisome biogenesis protein 19-1-like	2.562	2.54E-06	0.437	−0.142	1000.000	0.364	BS, LFMM
Contig_11734_331	Mate efflux family protein chloroplastic-like	3.004	3.48E-08	0.431	−0.161	1000.000	0.410	BS, LFMM
Contig_20725_122	PREDICTED: uncharacterized protein LOC100815781	−2.257	3.41E-05	0.175	0.009	1000.000	0.394	BS, LFMM
Contig_13800_1822	NAD-binding Rossmann-fold superfamily protein	2.217	4.70E-05	0.053	−0.220	0.946	0.308	BS, LFMM
Contig_79879_387	Serine carboxypeptidase-like 51-like	−1.794	9.89E-04	0.022	−0.187	1.209	0.337	BS, LFMM
Contig_42439_1172	Monoglyceride lipase-like	4.024	1.47E-13	−0.024	−0.189	0.648	0.310	BS, LFMM
Contig_7758_543	Probable inactive leucine-rich repeat receptor-like protein kinase at3g03770-like	2.412	9.44E-06	−0.102	−0.101	2.317	0.374	BS, LFMM
Contig_35584_728	Gibberellin 20 oxidase 1-like	−1.800	9.51E-04	−0.149	−0.157	0.729	0.311	BS, LFMM
Contig_32936_1517	Zinc finger an1 domain-containing stress-associated protein 12-like	2.638	1.28E-06	−0.296	0.046	1.014	0.314	BS, LFMM
Contig_7535_3225	Respiratory burst oxidase homolog protein a-like	3.002	3.56E-08	−0.305	−0.228	1.167	0.327	BS, LFMM
Contig_6331_350	Peroxisome biogenesis protein 19-1-like	−1.539	4.71E-03	2.249	−0.312	1000.000	0.403	BS, BE
Contig_6523_69	Probable protein phosphatase 2c 27-like	1.241	2.27E-02	2.017	−0.390	1.331	0.297	BS, BE
Contig_19119_1922	Pentatricopeptide repeat-containing protein chloroplastic-like	1.028	5.90E-02	1.757	−0.313	1.209	0.318	BS, BE
Contig_21612_169	Aquaporin nip2-1-like	1.722	1.56E-03	1.525	−0.443	2.853	0.390	BS, BE
Contig_606_1275	Abscisic acid-responsive element-binding factor 2	1.698	1.82E-03	0.908	-0.353	3.398	0.385	BS, BE
Contig_1842_785	E3 ubiquitin–protein ligase at1g12760-like	0.536	3.25E-01	0.570	−0.346	0.970	0.374	BS, BE

PCA, principle components analysis; SNP, single-nucleotide polymorphism.

Table 4 SNPs identified as significant outliers by at least two of the three methods used: LFMM, BAYENV2 and BAYESCAN, along with the output statistics for each method. SNPs identified by LFMM and BAYENV2 show significant correlations with the second principal component (PC2) from a PCA of environmental variables. The column 'Methods' states by which of the three methods the SNP was found to be a significant outlier; BS = BAYESCAN, BE = BAYENV2. BF = Bayes Factor

		LFMM		BAYENV2		BAYESCAN		
SNP ID	Sequence description	*Z*-scores	*P*-values	$\log_{10}$(BF)	Spearman's ρ	$\log_{10}$(PO)	F_{ST}	Methods
Contig_2166_2584	Protein	2.497	3.00E-06	1.503	0.398	1000.000	0.369	BS, BE, LFFM
Contig_20725_122	PREDICTED: uncharacterized protein LOC100815781	1.880	4.35E-04	1.417	0.310	1000.000	0.394	BS, BE, LFFM
Contig_606_1275	Abscisic acid-responsive element-binding factor 2	−1.798	7.69E-04	1.356	0.386	3.398	0.385	BS, BE, LFFM
Contig_13834_222	12-oxophytodienoate reductase 3-like	2.587	1.29E-06	0.686	0.325	1000.000	0.376	BS, BE, LFFM
Contig_23186_242	Protein fez-like	1.837	5.89E-04	0.666	0.303	2.249	0.335	BS, BE, LFFM
Contig_13925_454	GTP binding protein	3.099	6.67E-09	1.640	−0.441	−1.660	0.125	BE, LFMM
Contig_44461_50	Mitogen-activated protein kinase kinase 2	−2.588	1.28E-06	1.164	−0.344	−2.032	0.123	BE, LFMM
Contig_293_1473	ABC transporter c family member 4-like	2.516	2.51E-06	0.966	0.317	−1.806	0.124	BE, LFMM
Contig_19715_73	Leucine-rich repeat transmembrane protein kinase	−4.755	5.80E-19	0.926	0.319	−1.894	0.124	BE, LFMM
Contig_17195_61	Probable sodium metabolite cotransporter chloroplastic-like	2.933	4.07E-08	0.854	−0.322	−2.139	0.122	BE, LFMM
Contig_69440_1541	Retrotransposon ty1-copia subclass	1.889	4.08E-04	0.849	−0.316	−2.052	0.123	BE, LFMM
Contig_8156_939	Cryptochrome 2	2.191	4.14E-05	0.723	0.361	−2.062	0.123	BE, LFMM
Contig_5953_250	ABC transporter c family member 4-like	4.367	3.05E-16	0.638	0.314	−2.152	0.123	BE, LFMM
Contig_8314_263	Tir-nbs-lrr resistance protein	1.853	5.28E-04	0.622	−0.307	−2.014	0.122	BE, LFMM
Contig_8314_464	Tir-nbs-lrr resistance protein	2.788	1.82E-07	0.531	−0.308	−2.191	0.123	BE, LFMM
Contig_43658_216	Ankyrin repeat family	2.514	2.56E-06	0.577	−0.249	0.554	0.274	BS, LFMM
Contig_36850_60	Ammonium transporter 3 member 1-like	−2.514	2.55E-06	0.354	0.152	2.299	0.342	BS, LFMM
Contig_1648_277	Carnitine racemase like protein	−2.509	2.68E-06	0.199	0.194	0.946	0.307	BS, LFMM
Contig_11820_305	Cytochrome p450 76a2-like	−2.118	7.42E-05	0.197	−0.180	0.851	0.322	BS, LFMM
Contig_35584_728	Gibberellin 20 oxidase 1-like	3.079	8.32E-09	0.059	0.103	0.729	0.311	BS, LFMM
Contig_8314_143	Tir-nbs-lrr resistance protein	3.161	3.34E-09	−0.013	0.149	3.699	0.365	BS, LFMM
Contig_8825_830	ABC transporter g family member 22-like	2.131	6.66E-05	−0.021	0.076	0.695	0.296	BS, LFMM
Contig_11272_2066	Dual specificity protein kinase spla-like	−3.818	9.08E-13	−0.032	0.111	1.829	0.399	BS, LFMM
Contig_17783_564	E3 ubiquitin–protein ligase sdir1-like	−4.050	3.54E-14	−0.078	0.150	1.024	0.300	BS, LFMM
Contig_35394_55	Low quality protein: uncharacterized loc101207585	4.492	4.32E-17	−0.094	0.092	1000.000	0.427	BS, LFMM
Contig_36175_123	E3 ubiquitin–protein ligase chip-like	−2.932	4.11E-08	−0.153	0.096	2.093	0.346	BS, LFMM
Contig_36495_15	UBX domain-containing protein	−4.707	1.30E-18	−0.153	−0.058	1.908	0.338	BS, LFMM
Contig_37898_1096	Probable inactive purple acid phosphatase 16-like	−2.702	4.28E-07	−0.160	−0.181	1000.000	0.400	BS, LFMM
Contig_42439_218	Monoglyceride lipase-like	−4.692	1.64E-18	−0.166	0.120	0.773	0.292	BS, LFMM
Contig_42439_513	Monoglyceride lipase-like	5.293	4.07E-23	−0.171	−0.097	1000.000	0.518	BS, LFMM
Contig_7130_49	Calcium-independent aba-activated protein kinase	−4.006	6.60E-14	−0.204	0.104	1.686	0.335	BS, LFMM
Contig_7535_2983	Respiratory burst oxidase homolog protein a-like	1.827	6.31E-04	−0.223	−0.062	1.954	0.330	BS, LFMM
Contig_7535_3225	Respiratory burst oxidase homolog protein a-like	−1.758	1.01E-03	−0.281	−0.019	1.167	0.327	BS, LFMM
Contig_20162_46	Las1-like family protein	−2.777	2.04E-07	−0.286	−0.046	0.686	0.289	BS, LFMM
Contig_37591_348	Malate glyoxysomal-like	0.438	4.12E-01	1.137	0.358	2.104	0.344	BS, BE
Contig_4517_54	Histone h4	−1.021	5.61E-02	1.063	0.363	1.105	0.312	BS, BE
Contig_4391_4421	Histidine kinase 3-like	−1.243	2.01E-02	1.015	0.313	1.343	0.346	BS, BE
Contig_34461_1104	Transmembrane protein 53-like	−1.422	7.81E-03	0.616	0.335	0.740	0.277	BS, BE

PCA, principle components analysis; SNP, single-nucleotide polymorphism.

contain significant PC1 SNPs, and one ABA-related gene (no aquaporin genes) contained a significant PC2 SNPs. Thirteen genes labelled with the GO term 'response to water deprivation' were found to contain significant PC1 SNPs, whilst 19 genes with the same GO term contained significant PC2 SNPs. In total, 56 genes with GO terms related to an environmental response (namely 'response to water deprivation', 'response to ABA stimulus', 'response to cold', 'response to salt stress' and 'photoperiodism, flowering') were found to contain significant SNPs (Fig. 4).

The most significant SNPs, in terms of $\log_{10}$BF values, correlating with PC1 were found in genes with the following products: s-adenosyl-l-methionine-dependent methyltransferases superfamily protein isoform 1, syntaxin-121-like, calcium-dependent protein kinase 13-like and cytochrome p450 76a2-like. Population allele frequencies for these genes display clear correlations when plotted against PC1 values (Fig. 5).

Redundancy analysis

Redundancy analysis showed that 12.3% of the variation across all 8462 SNPs was explained by either PC1 (0.3%), PC2 (2.5%), space (7.6%), a combination of PC1 and PC2 (0.02%), or PC2 and space (1.9%), leaving 87.7% of the variance unexplained (Fig. 6). The global F_{ST} was 0.102 and the 12.3% of explained variation is equivalent to an F_{ST} of 0.013. The relatively high percentage contribution of 'space' compared with PC1 and PC2 indicates IBD is present within the data, which may confound the identification of signatures of selection.

Discussion

Our study identifies signatures of selection as well as potential drivers of selection along an environmental gradient for an ecologically important Australian species that is widely used in restoration. Using a targeted sequencing approach, we identified 93 SNPs distributed over 81 genes that displayed evidence of directional selection and strong correlations with environmental factors, namely moisture and temperature variation as well as elevation.

As expected, the RDA showed that only a small fraction of the variation (<2.82%) across the 8462 SNPs could be explained purely by environmental variation, suggesting that local selection due to the environmental variables considered here in the 17 populations, and/or selection in the ancestral populations from which they were recently derived, has left its mark on only a small number of the genes considered in this study. This is further supported by the equivalent F_{ST} value of 0.013 for the explained variation. For most, if not all species, only a small percentage of the genome is expected to be under selection, and neutral processes (e.g. genetic drift) have determined the majority of genetic variation (Rockman 2012; Rands *et al.* 2014; Meirmans 2015). Of the 970 targeted genes and the 8462 SNPs present within them, only 93 SNPs (1.1%) distributed over 81 genes (8.2%) exhibited strong evidence of selection. It is likely that we have underestimated the total number of SNPs under selection due to high stringency on the LFMM and BAYENV2 outputs, plus the fact that we only included SNPs that demonstrated significant signatures of selection by at least two of the methods (LFMM, BAYENV2, and BAYESCAN). This conservative approach may therefore have resulted in the rejection of true positives, particularly for SNPs showing signs of weak selection (Rockman 2012; Lotterhos & Whitlock 2015). The low proportion of SNPs showing signs of selection was therefore not surprising despite the fact that we selected putatively functional genes with a priori expectations that they might be

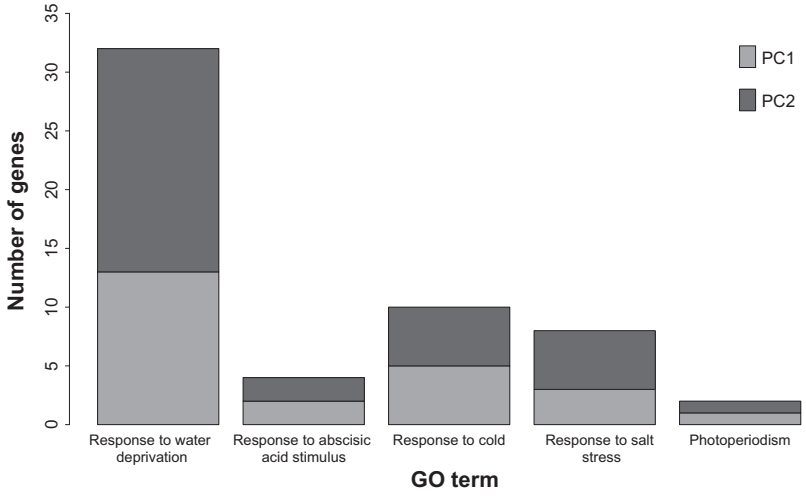

Fig. 4 Number of genes (with gene ontology term relating to an environmental response) containing single-nucleotide polymorphisms that correlate significantly with either PC1 or PC2.

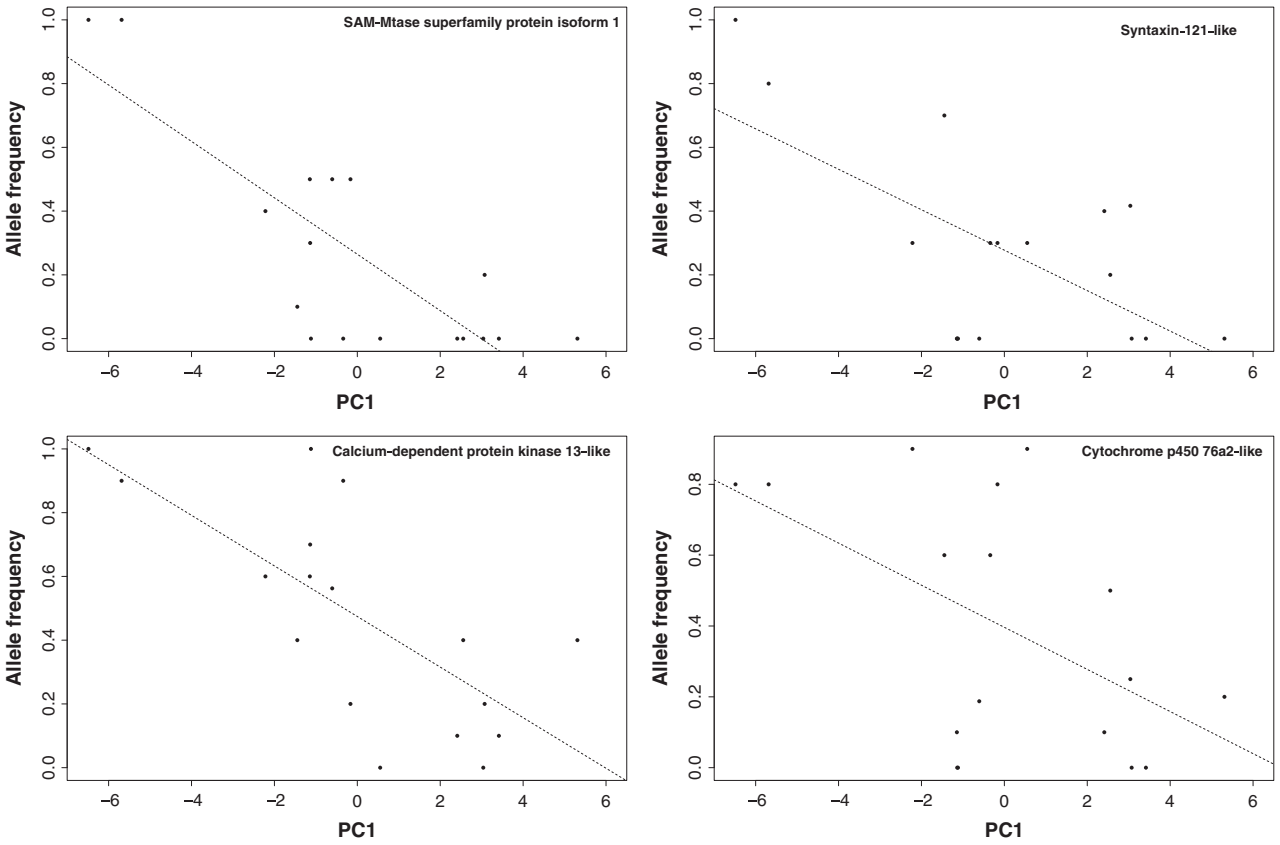

Fig. 5 Example correlations between allele frequency and environmental factors summarized by PC1 (temperature and water related) for the four single-nucleotide polymorphism markers with the greatest Bayes Factor score that were identified as being significant outliers by all methods (BAYESCAN, LFMM, BAYENV2). Linear regression lines are shown, all of which are significant at $P < 0.05$.

targets of selection. It is also likely that there are other factors driving selection, which were not included in and did not correlate with the 15 environmental factors included in the PCA (e.g. biotic interactions), and so any SNPs under selection from other factors would not have been detected.

The high stringency employed in the genotype–environment association analyses in this study was necessary as currently unpublished work on the same populations suggests that their current distributions may be the result of post-Pleistocene range expansions. Range expansions have been shown to lead to high false-positive rates in genotype–environment association analyses as they can result in allele frequency changes that mimic those resulting from adaptive processes (Lotterhos & Whitlock 2014). The use of BAYENV2, which uses a set of neutral SNPs as a null model to which all other SNPs are compared, and LFMM, which uses the number of population clusters as latent factors in the model, helped to ensure that the population genetic structure in our data was accounted for, giving some control over false-positive rates.

Selection and adaptation

The first principal component of the PCA accounted for 57.7% of the variance among the 15 environmental variables. This PC also correlated positively with latitude, which is consistent with decreases in temperature, evaporation, water stress, aridity and radiation and increases in precipitation, humidity and organic carbon with an increase in latitude (Table 2). The genes containing the 55 SNPs that showed strong associations with PC1 may therefore be under selection from temperature, water availability, radiation and/or available organic carbon pressures. However, as these factors all covary, it is difficult to ascertain the exact agent of selection responsible for the patterns observed in this study.

Our study is effective in identifying potential targets of selection within the genome, but the actual agents of selection as well as the functional consequences of selection cannot be concluded without further investigation. Gene product functions of outlier genes may provide hints towards these ends, but as gene products can have several functions and be involved in a range of

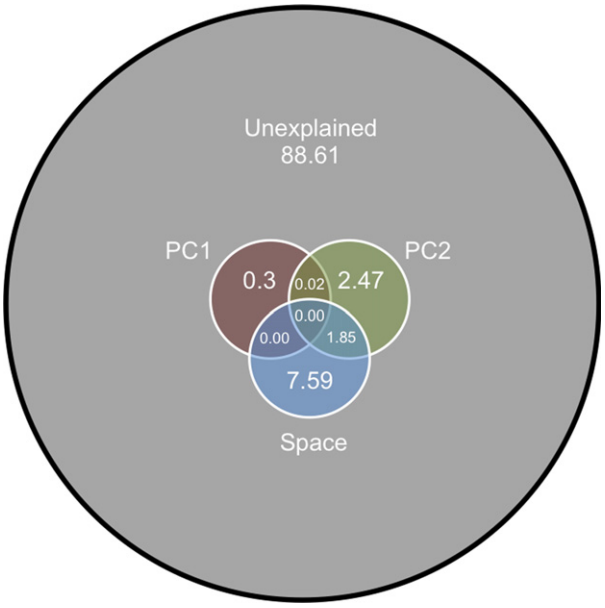

Fig. 6 Decomposition of among-population variation of 17 populations of *Dodonaea viscosa* ssp. *angustissima* distributed along an environmental gradient using redundancy analysis. Variation is decomposed into environmental components (PC1: red, and PC2: green), a spatial component (blue), overlap between these components and an unexplained component (grey) for all 8462 single-nucleotide polymorphisms.

pathways, it is difficult to come to any firm conclusions (Pavlidis *et al.* 2012). Common garden trials with manipulation of environmental factors could provide more evidence towards the actual agents of selection. For example, gene expression analysis before and after a drought treatment may shed light on whether any of the outlier genes identified in this study are differentially expressed among populations in response to drought.

The second PC accounted for 18.4% of the variance in environmental variables and mostly summarized elevational differences between populations, as well as annual mean minimum temperatures (which correlated with elevation). Elevation ranged from 27 m at the Gammon Ranges 4 site to 750 m at the Wilpena Pound site. Our findings suggest that genes correlating with PC2 may be under selection from the abiotic changes that occur with increased elevation (e.g. decreased temperatures).

Genes with a wide variety of functions were found to contain SNPs demonstrating significant correlations with either PC1 or PC2, including the specifically targeted genes coding for aquaporin and ABA-related proteins, both of which have been shown to play roles in response to environmental (particularly water) stress (Zhu 2002; Kaldenhoff *et al.* 2007). Aquaporins are a class of integral membrane proteins that form pores in cell membranes for water permeation. The function of aquaporins is interconnected with signal transduction by the plant hormone ABA (Tyerman *et al.* 2002) and transcellular water flow through aquaporins is influenced by ABA activity (Kaldenhoff *et al.* 1996; Morillon & Chrispeels 2001). Aquaporins and ABA are thought to be involved in the maintenance of plant homeostasis and water balance under water-stressed conditions (Tyerman *et al.* 2002; Galmes *et al.* 2007). For the populations studied here, it is easy to envisage that more efficient water uptake by as well as greater control over water movement throughout plants in the more arid Gammon Ranges would be a selective advantage. The observed changes in allele frequencies in aquaporin and ABA-related genes may well be a reflection of this selection.

Recent work on *D. v. angustissima* along the same environmental gradient has identified clines in leaf width with latitude and altitude (Guerin & Lowe 2012; Guerin *et al.* 2012) and a positive correlation between stomatal density and mean summer maximum temperature (Hill *et al.* 2015). Both of these clines are believed to be adaptive responses to climate (specifically temperature and water availability) but whether they have a genetic basis or are plastic responses is unknown. In this study, we identified SNPs that are potentially involved in clinal phenotypic variation. Three genes containing significant SNPs correlating with PC1 had GO terms relating to stomata: a syntaxin-121-like gene (contig 15281) and a respiratory burst oxidase gene (contig 7535) both had the GO term 'regulation of stomatal movement' associated with them, and an NAD-binding Rossmann-fold superfamily gene (contig 13800) with the GO term 'stomatal complex morphogenesis'. Syntaxins and respiratory burst oxidase proteins have both been shown to play roles in the control of stomatal opening as a response to abiotic stress in plants (Zhu *et al.* 2002; Torres & Dangl 2005). Abiotic stress resulting from higher temperatures and lower water availability will be higher in the more northern populations and may act as a selecting agent. The genetic variants identified here may therefore have been selected for, enabling northern populations more efficient stomatal control to better respond to this stress. Seven further genes with GO terms relating to stomatal movement and/or genesis were found to contain SNPs correlating with PC2.

Significant SNPs were also found in two genes with GO terms relating to leaf development and structure: a calcium-independent ABA-activated protein kinase gene (contig 7130), and a protein fez-like gene (contig 23186) both showed significant correlations with PC2. The genotypic clines identified in this study may hint at the genetic underpinnings of the phenotypic clines identified in Guerin & Lowe (2012), Guerin *et al.* (2012) and Hill *et al.* (2015). Although we cannot confirm in this

study that the identified genes are the direct targets of selection, our results do suggest that the environment has shaped the distribution of alleles across a number of functional genes in this species. Hypotheses around the importance of the identified clinal variation based on the function of the products of the variable genes will require further testing. Common garden or reciprocal transplant experiments using seed collected from populations throughout the gradient should be undertaken to determine how much of the phenotypic variation identified in the field is genetically determined and to provide stronger evidence for the link between phenotype and genotype.

Nonmodel species and genomic resources

If a reference genome is available for a study species, then loci identified as demonstrating signatures of selection can be traced back to their exact location in the genome (Bragg *et al.* 2015). For example, when seeking genomic regions under selection in the Grey Wolf, Schweizer *et al.* (2016) identified variants associating strongly with environmental variables, which they traced to genes relating to immunity, coat pigmentation and metabolism. However, the cost and complexity of generating a high-quality fully assembled and annotated genome is still prohibitive in most cases. Functional information of outlier loci can still be achieved via alternative, cheaper genomic methods as demonstrated in this study. The generation of transcriptome data via RNA-seq is possible at a fraction of the cost of whole genome sequencing (Martin & Wang 2011; Garg & Jain 2013; Christmas *et al.* 2015b). Although a transcriptome does not give any information on the location of genes in a genome, it does provide the researcher with an abundance of data on which genes are being transcribed and, therefore, potentially of functional importance. BLAST searches and GO annotations can then provide information on the putative functions of transcribed sequences (De Wit *et al.* 2012) and targeted capture of genes of interest can provide population-level measures of variation within those genes (Mamanova *et al.* 2010; Jones & Good 2016), as demonstrated here. From this relatively low cost approach, we have generated information on putatively functional genetic variation in a suite of targeted genes with known function for populations distributed along a 700-km latitudinal gradient, establishing an invaluable baseline for further experimental studies into climate adaptation in this species.

Conservation and restoration implications

In this study, we identified genomic signatures of selection across an environmental gradient in a species commonly used in revegetation and restoration. This information can be used to develop seed-sourcing guidelines. If the signals in the genomic data truly reflect adaptive variation to climate, then careful selection of seed in terms of source population location could greatly assist the movement of adaptive genotypes across the landscape (Aitken & Whitlock 2013; Steane *et al.* 2014; Prober *et al.* 2015). A southward shift in climate conditions is expected under climate change in the study region, and so the flow of adaptive alleles from more northern (e.g. Gammon Ranges) into more southern (e.g. southern Flinders Ranges) populations could assist adaptation to climate in the southern populations.

More long-term strategies of sourcing seed from areas with high adaptive potential under future climate scenarios are needed (Breed *et al.* 2012; Prober *et al.* 2015). Predictive or climate-adjusted provenancing, whereby local germplasm is combined with seed from increasingly warmer and drier sites across a gradient, should be particularly useful for *D. v. angustissima* across our study region based on the discovery of genes under environmental selection. We therefore propose that restoration efforts in the south of the region should begin to incorporate seed from northern, seemingly more arid-tolerant populations of *D. v. angustissima* in order to assist this required flow of adaptive alleles under climate change. These should be established in a way that allows for the long-term monitoring of establishment and fitness of the different source provenances to assess the success of the seed-sourcing strategy. The potential benefits of this type of strategy should be weighed up against the risks of outbreeding depression (Breed *et al.* 2012; Prober *et al.* 2015). The movement of seed between the three genetic clusters identified here may carry such risks.

The putatively adaptive variation identified here needs further confirmation of its importance to adaptation to climate through, for example common garden or reciprocal transplant experiments: whether these gene variants translate into different phenotypes, which in turn result in fitness differences, is yet to be revealed (Storz & Wheat 2010). Embedding experiments such as reciprocal transplants into restoration projects holds real promise for testing the effectiveness of moving germplasm across the landscape and to more fully assess the success of provenancing strategies (Breed *et al.* 2012).

Acknowledgements

The authors wish to thank the Australian Research Council for funding support (LP110100721 awarded to AJL; DE150100542 awarded to MFB; DP150103414 awarded to AJL and MFB), the

South Australian Premier's Science and Research Fund awarded to AJL, and the Field Naturalist Society of South Australia Lirabenda Endowment Fund and the Australian Wildlife Society Student Grant awarded to MJC.

References

Aitken SN, Whitlock MC (2013) Assisted gene flow to facilitate local adaptation to climate change. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 367–388.

Audigeos D, Buonamici A, Belkadi L *et al.* (2010) Aquaporins in the wild: natural genetic diversity and selective pressure in the PIP gene family in five Neotropical tree species. *BMC Evolutionary Biology*, **10**, 202.

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution*, **23**, 773–783.

Bragg JG, Supple MA, Andrew RL, Borevitz JO (2015) Genomic variation across landscapes: insights and applications. *New Phytologist*, **207**, 953–967.

Breed MF, Stead MG, Ottewell KM, Gardner MG, Lowe AJ (2012) Which provenance and where? Seed sourcing strategies for revegetation in a changing environment. *Conservation Genetics*, **14**, 1–10.

Chavez-Galarza J, Henriques D, Johnston JS *et al.* (2013) Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular Ecology*, **22**, 5890–5907.

Chen J, Källman T, Ma X *et al.* (2012) Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (*Picea abies*). *Genetics*, **191**, 865–881.

Christmas MJ, Breed MF, Lowe AJ (2015a) Constraints to and conservation implications for climate change adaptation in plants. *Conservation Genetics*, **17**, 305–320.

Christmas MJ, Biffin E, Lowe AJ (2015b) Transcriptome sequencing, annotation and polymorphism detection in the hop bush, *Dodonaea viscosa*. *BMC Genomics*, **16**, 803.

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

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

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

De Mita S, Thuillet AC, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

De Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.

Dillon S, McEvoy R, Baldwin DS *et al.* (2014) Characterisation of adaptive genetic diversity in environmentally contrasted populations of *Eucalyptus camaldulensis* Dehnh.(River Red Gum). *PLoS ONE*, e103515. DOI: org/10.1371/journal.-pone.0103515.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetic Resources*, **4**, 359–361.

Ekblom R, Galindo J (2010) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

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

Fournier-Level A, Korte A, Cooper MD *et al.* (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science*, **334**, 86–89.

Frichot E, François O (2015) LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, **6**, 925–929.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Galmes J, Pou A, Alsina MM, Tomas M, Medrano H, Flexas J (2007) Aquaporin expression in response to different water stress intensities and recovery in Richter-110 (*Vitis* sp.): relationship with ecophysiological status. *Planta*, **226**, 671–681.

Garg R, Jain M (2013) RNA-Seq for transcriptome analysis in non-model plants. In: *Legume Genomics Methods and Protocols* (ed. Rose RJ), pp. 43–58. Humana Press, New York.

Guerin GR, Lowe AJ (2012) Leaf morpholgy shift: new data and analysis support climate link. *Biology Letters*. DOI: 10.1098/rsbl.2012.0860.

Guerin GR, Wen H, Lowe AJ (2012) Leaf morphology shift linked to climate change. *Biology Letters*, **8**, 882–886.

Guillot G, Vitalis R, le Rouzic A, Gautier M (2014) Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. *Spatial Statistics*, **8**, 145–155.

Guo B, DeFaveri J, Sotelo G, Nair A, Merilä J (2015) Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. *BMC Biology*, **13**, 19.

Hill KE, Guerin GR, Hill RS, Watling JR (2015) Temperature influences stomatal density and maximum potential water loss through stomata of *Dodonaea viscosa* subsp. *angustissima* along a latitude gradient in southern Australia. *Australian Journal of Botany*, **62**, 657–665.

Hoffmann A, Griffin P, Dillon S *et al.* (2015) A framework for incorporating evolutionary genomics into biodiversity conservation and management. *Climate Change Responses*, **2**, 1–24.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.

Kaldenhoff R, Kolling A, Richter G (1996) Regulation of the *Arabidopsis thaliana* aquaporin gene AthH2 (PIP1b). *Journal of Photochemistry and Photobiology*, **36**, 351–354.

Kaldenhoff R, Bertl A, Otto B, Moshelion M, Uehlein N (2007) Characterization of plant aquaporins. *Methods in Enzymology*, **428**, 505–531.

Kaldenhoff R, Ribas-Carbo M, Sans JF *et al.* (2008) Aquaporins and plant water balance. *Plant, Cell & Environment*, **31**, 658–666.

Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Keller I, Taverna A, Seehausen O (2011) Evidence of neutral and adaptive genetic divergence between European trout populations sampled along altitudinal gradients. *Molecular Ecology*, **20**, 1888–1904.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

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

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Manel S, Perrier C, Pratlong M *et al.* (2016) Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular Ecology*, **25**, 170–184.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*, **12**, 671–682.

McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics*, **5**, e1000686.

Meirmans PG (2012) The trouble with isolation by distance. *Molecular Ecology*, **21**, 2839–2846.

Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them. *Molecular Ecology*, **24**, 3223–3231.

Morillon R, Chrispeels MJ (2001) The role of ABA and the transpiration stream in the regulation of the osmotic water permeability of leaf cells. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 14138–14143.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

Narum SR, Hess J (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

Oksanen J, Kindt R, Legendre P (2009) *Vegan: Community ecology package*. R package version 1.15-3. Available from http://CRAN.R-project.org/package=vegan

Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012) A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, **29**, 3237–3248.

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

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Prober SM, Byrne M, McLean EH *et al.* (2015) Climate-adjusted provenancing: a strategy for climate-resilient ecological restoration. *Frontiers in Ecology and Evolution*, **3**, 65.

Prunier J, Laroche J, Beaulieu J, Bousquet J (2011) Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Molecular Ecology*, **20**, 1702–1716.

Rands CM, Meader S, Ponting CP, Lunter G (2014) 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genetics*, **10**, e1004525.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348–4370.

Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*, **66**, 1–17.

Savolainen O, Lascoux M, Merila J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.

Schweizer RM, Robinson J, Harrigan R *et al.* (2016) Targeted capture and resequencing of 1040 genes reveal environmentally driven functional variation in grey wolves. *Molecular Ecology*, **25**, 357–379.

Shafer AB, Wolf JB, Alves PC *et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, **30**, 78–87.

Stapley J, Reger J, Feulner PG *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology and Evolution*, **25**, 705–712.

Steane DA, Potts BM, McLean E *et al.* (2014) Genome–wide scans detect adaptation to aridity in a widespread forest tree species. *Molecular Ecology*, **23**, 2500–2513.

Storz JF, Wheat CW (2010) Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution*, **64**, 2489–2509.

Torres MA, Dangl JL (2005) Functions of the respiratory burst oxidase in biotic interactions, abiotic stress and development. *Current Opinion in Plant Biology*, **8**, 397–403.

Tsumura Y, Uchiyama K, Moriguchi Y, Ueno S, Ihara-Ujino T (2012) Genome scanning for detecting adaptive genes along environmental gradients in the Japanese conifer, *Cryptomeria japonica*. *Heredity*, **109**, 349–360.

Tyerman SD, Niemietz CM, Bramley H (2002) Plant aquaporins: multifunctional water and solute channels with expanding roles. *Plant, Cell & Environment*, **25**, 173–194.

Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114.
Zhu J-K (2002) Salt and drought stress signal transduction in plants. *Annual Review of Plant Biology*, **53**, 247.
Zhu J, Gong Z, Zhang C *et al*. (2002) OSM1/SYP61: a syntaxin protein in *Arabidopsis* controls abscisic acid–mediated and non-abscisic acid–mediated responses to abiotic stress. *The Plant Cell*, **14**, 3009–3028.

M.J.C., E.B., M.F.B. and A.J.L. designed the research. M.J.C. and E.B. performed field collections and laboratory work. M.J.C. analysed the data. M.J.C. wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

Data accessibility

Sequencing reads are archived at the NCBI SRA under the Accession no. SRP077342, associated with BioProject PRJNA326697. The variants file is available as a supplementary file (Appendix S1, Supporting information). Details of transcriptome data archiving can be found in Christmas *et al*. (2015b).

Supporting information

Additional supporting information may be found in the online version of this article.

Figs. S1 and S2 Histograms of corrected *P*-values from the Latent Factor Mixed Model (LFMM) analysis for PC1 and PC2 respectively.

Fig. S3 Global outlier detection among 8462 SNPs in 17 *Dodonaea viscosa* ssp. *angustissima* populations from South Australia in BAYESCAN with prior odds = 100.

Fig. S4 Graphs showing outputs from neutral population genetic analyses in STRUCTURE and DAPC.

Table S1 Environmental data downloaded from the Atlas of Living Australia for each of the sampling locations.

Table S2 Gene Ontologies of genes containing SNPs identified as outliers in BAYESCAN as well as showing significant correlations with the first (Table 1) or second (Table 2) principal component from a PCA of environmental variables in a latent factor mixed model (LFMM) analysis and/or a mixed effect model (BAYENV2).

Appendix S1 A .vcf formatted file containing the SNP data analysed in this study.