Assessing genomic variation in the hopbush, Dodonaea viscosa, to investigate micro-evolution and adaptation

By Matthew John Christmas

School of Biological Sciences

The University of Adelaide

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in December 2015

"One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die."

Darwin, 1859

Table of Contents

Summary

In this thesis I use a range of genomic tools in order to investigate aspects of the evolutionary history of the hopbush, *Dodonaea viscosa* (L.) Jacq. (hopbush). Understanding the past evolutionary processes that have led to the distribution and adaptation of contemporary populations can help to inform on how populations may continue to adapt into the future. A changing climate is altering selection pressures and populations will need to adapt to these if they are to persist. I pay consideration to how plant populations may adapt to a changing climate, as well as what constrains adaptive responses and how an understanding of these constraints can inform conservation management, in a review article (Chapter 2).

Dodonaea viscosa is globally widespread and distributed across the Australian continent where it originated. Within Australia there are seven taxonomically identified subspecies, but subspecies boundaries are not clear. Here, I analysed genome-wide single nucleotide polymorphic (SNP) markers for each of the seven subspecies using genetic structure analysis as well as network-based and Bayesian-based methods to assess phylogenetic relationships and show that defining subspecies for *D. viscosa* based on morphology is not wholly accurate or useful (chapter three). A consideration of a population's geographic location, environment and Pleistocene history appears to better explain the distribution of genetic variation across this species, rather than the presence of distinct gene pools corresponding to morphologically defined subspecies.

The thesis then moves onto an examination of the population history and signatures of selection in the narrow-leaf hopbush, *Dodonaea viscosa* ssp. *angustissima* along an environmental gradient within South Australia. Firstly, I generated a transcriptome reference genome by sequencing RNA from several individuals and used a range of bioinformatic tools to assemble and annotate it (chapter four). This transcriptome was

then used to selectively target a subset of 970 genes via 'hybrid-capture target enrichment' in order to develop a set of SNP markers present within putatively functional genes distributed throughout the genome. The SNP markers were used to characterise population genomic metrics of the target populations, such as genetic diversity and structure (chapter five). A range of analytical methods was used, including Principal Component Analysis, Mantel tests, AMOVA, and an F_{ST} outlier analysis algorithm. The data demonstrates there to be three distinct genetic clusters across the study region with low gene flow among them. The potential origins of these clusters are discussed, along with the implications for restoration practice using germplasm derived from these distinct gene pools.

SNP markers within the sequenced target genes were also analysed for signatures of selection, providing evidence for local adaptation driven by climatic factors (chapter six). A combination of F_{ST} outlier analysis and genotype-environment association analyses found 74 SNPs showing strong evidence for selection. Genes containing these SNPs under environmental selection were diverse, including aquaporin and abscisic acid (ABA) genes, as well as genes with ontologies relating to environmental responses, such as 'response to water deprivation'. Selection acting on these populations has led to clines in allele frequencies in a number of functional genes, including some genes associated with leaf shape and stomatal characteristics, the phenotypes for which have been observed previously to vary along this environmental gradient. The implications for conservation and restoration practice using such data are further discussed.

The thesis concludes with a consideration of the future directions of research that can be informed by and further strengthen the findings of this thesis. Confirmation of adaptive significance of the genomic variation identified here is required. The use of common garden and reciprocal transplant experiments in order to provide evidence for

links between genotype and phenotype and to asses the relative roles of genetic adaptation and plasticity in this species are discussed.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: Date:

Acknowledgements

I feel extremely grateful for the wealth of experiences I have gained during the past three and a half years and, although undertaking a PhD can be seen as quite an individual pursuit, it would not have been possible without the large number of people who have helped me, supported me, and kept me (mostly) sane and grounded along the way. Thank you to everyone who I have met during this time, you have all helped to make this one of my most fulfilling experiences.

To Andy, without your trust and confidence in taking me on as a student none of what I have achieved over the past three and a half years would have been possible and I am forever thankful to you. At the start I was not really sure what I was letting myself in for or whether I was really up to the task, so thank you for taking the risk and giving me a chance. You have been nothing but supportive throughout, providing invaluable advice, critique, and direction. You have also ensured the financial support was available for my project, as well as enabling me to attend conferences and workshops. Thank you for your unwavering support and I look forward to continuing to work with you into the future.

To Ed, I was definitely out of my depth when I first walked in to the lab, not knowing my PCR primers from my sequencing adaptors. I had a whole new set of skills to learn, as well as the vocabulary to go with it, and I owe my survival through these early stages largely to you. Your tutorship, patience, endless knowledge, and effective yet unorthodox lab skills helped me to learn a whole new set of skills in the lab, in front of the computer and in the field. Your continued support and advice has helped to shape this final thesis into form and I am extremely grateful for all the hours you have given me.

To Martin, your energy and enthusiasm is infectious and has really helped me to bring all of this together over the past 12 months. Thank you for your continued support and advice, our discussion sessions really paid off and helped to get from 'large dataset' to something that is manageable, interpretable and, of course, publishable! Beyond the office, thanks for the surf sessions that have helped to keep me sane and for convincing me that taking a couple of weeks away from thesis writing to hike the South Coast Track was a good idea!

To all the members of the Lowe lab group, it has been a pleasure working with you all, and drinking with you all, over this time. Wherever I end up in my career, I am sure the Lowe lab group retreats will not be surpassed, in injury rates at least! Particular thanks to Duncan, my office and Friday beers buddy, cheers for being a true friend through it all, you've been there for me through the good and the bad, and I am extremely grateful for your friendship.

To Angela, I will always be grateful to you for helping to convince me that taking on a PhD is something I should do and am capable of achieving, and for providing me with support throughout. Thank you.

To my family, I know being millions of miles away is not easy but knowing you are all there for me gives me a lot of strength. Thank you for all the unconditional love and support you give. Thank you to mum and dad who have always encouraged my love for nature. The infinite chain of pets as a child, the trips to the Natural History and Science museums in London, the endless summers spent outside, all inspired and fed my interest in the natural world. Without this encouragement I would not be where I am today.

To my aunty Kim, I never knew where I would end up, and I am still not sure, but I am following my nose, driven by my passions and interests and a love for fun and I thank you for helping to instil this attitude into me. I miss you.

To Monique, thank you for your no-nonsense attitude when I needed it the most, for helping to kick me into shape to get over the line with this, for all your help and support through the 'broken arm' saga, and for always keeping my spirits high. Here's to the start of a very bright future.

Thank you to the University of Adelaide, the Environment Institute and the School of Biological Sciences for providing me with the facilities, infrastructure and support required for getting through a postgraduate degree.

Thank you to the funding bodies who have supported my research, the Australian Research Council, the Field Naturalist Society of South Australia, and the Australian Wildlife Society. Without your support this work would not have been possible.

Finally, a thank you to Charles Darwin. He has provided an answer to one of the greatest mysteries of how the biotic world has come to be. Without him and his revelations, the field that inspires me most and that I hope to build a research career in would be nowhere near as advanced as it is today. I pay homage to Darwin through the quotes included throughout this thesis.

Chapter 1 - Introduction

"There is grandeur in this view of life, with its several powers, having been originally breathed by the Creator into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been and are being evolved."

Darwin, 1859

Standing on the shoulders of giants

"As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be naturally selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form." *Darwin (1859).*

When Darwin first wrote this statement for *The origin of species*, exactly what was being inherited from one generation to the next was unknown to him, making his newly developed theory to the workings of nature even more profound. Similarly, the breeding experiments of Gregor Mendel and his exploration of inheritance patterns in pea plants during the 1860s were blind to the genetics behind inheritance (Mendel 1866). The discovery of DNA in 1869 by Friedrich Miescher was the first in a series of revelations regarding the 'object' of inheritance, culminating in the description of the structure of DNA presented by James Watson and Francis Crick in 1953 (Watson & Crick 1953). Alongside these discoveries, work by eminent scientists Ronald Fisher, JBS Haldane and Sewall Wright established the field of population genetics, integrating Mendelian genetics with the theory of natural selection in a statistical framework, giving rise to the modern evolutionary synthesis.

The development of one of the first DNA sequencing technologies by Frederick Sanger in 1977 (Sanger et al. 1977b) meant that biologists suddenly had access to the underlying genetic sequences that gave rise to the proteins and phenotypes they had been studying. The development of the chain-termination or Sanger sequencing method led to the sequencing of whole genomes: first, that of a bacteriophage (Sanger et al. 1977a), in 1995 the first genome of a free-living organism (Fleischmann et al. 1995),

then the first eukaryote genome (Goffeau et al. 1996), and in 2003 the human genome project was completed (Consortium 2004).

Sanger sequencing was, however, slow and costly, leading to a high demand for and large investment in developing (relatively) low cost, high-throughput sequencing methods. This led to the next wave of sequencing technologies, termed "next-generation sequencing", which became widely available in the mid 2000s and are the predominant sequencing technologies used today. These include Roche 454 pyrosequencing, Illumina sequencing, SOLiD sequencing, and Ion semiconductor sequencing. These "next-gen" technologies are capable of generating millions of short sequence reads in one sequencing run, providing biologists with orders of magnitude more data than previously possible, opening up the study and comparison of genomes of all organisms and not just classic model species.

In evolutionary biology, the accessibility of genome-wide data has revolutionised the field. Researchers are no longer restricted to a small number of molecular markers such as microsatellites or AFLPs which do not give a complete picture of the effects of evolutionary processes across the genome. Instead, markers distributed across the whole genome, in introns, exons and promoter regions, regions of functional importance and neutral regions, can now be sequenced to better address questions of evolution relating to neutral and adaptive processes. The field of population genomics, as opposed to population genetics, has grown alongside the new sequencing technologies. Population genomics can be defined as the study of numerous loci simultaneously to better understand the roles of evolutionary processes that influence variation across genomes and populations (Luikart et al. 2003). Population genomics has the advantage over population genetics in that it can be used to identify and separate locus-specific effects (such as selection and mutation) from genome-wide effects (such as drift, gene flow and inbreeding) (Black et al. 2001; Luikart et al. 2003).

Multiple methods for rapidly and cheaply generating genome-wide markers for use in a population genomics context have been developed and include restriction-digest based methods such as RADseq (Davey & Blaxter 2010), as well as more targeted approaches such as RNAseq (Van Verk et al. 2013), exome capture and target-enrichment (Mamanova et al. 2010). These new methods allow for the reliable genotyping of hundreds if not thousands of genetic markers (usually single nucleotide polymorphisms, or SNPs) for multiple individuals across multiple populations. Genome-wide SNP markers can then be analysed to address questions in an evolutionary and ecological context (Bonin 2008; Rokas & Abbot 2009; Allendorf et al. 2010; Ouborg et al. 2010; Stapley et al. 2010; Savolainen et al. 2013).

In this thesis, I have utilised the advancement in sequencing technology to develop reduced-representations of the genome of a non-model plant species in a population genomics context in order to address questions of phylogeography, population demography and connectivity, and adaptive processes through a series of four empirical chapters. I focus on the hopbush, *Dodonaea viscosa* (Sapindaceae), a dioecious woody shrub that is widespread throughout Australia and also has a global distribution. In Australia it inhabits a wide range of habitats and displays a diversity of ecotypes suggesting that evolutionary processes have had a large role to play in establishing the patterns we see today. *Dodonaea viscosa* is also commonly used in restoration projects as it is a very hardy species that survives well in most habitats.

A precedent has already been set for studying adaptive processes in *D. viscosa.* Previous work on this species has demonstrated the presence of phenotypic clines (namely leaf width and stomatal density) along a latitudinal gradient in South Australia (Guerin & Lowe 2012; Guerin et al. 2012; Hill et al. 2015). Until now, no genetic analysis had been carried out on these populations and genetic resources for the species as a whole were greatly lacking. I made use of the same study region of Guerin & Lowe

(2012), Guerin et al. (2012) and Hill et al. (2015) in order to develop our understanding of the neutral processes as well as the adaptive responses to environment that have, in part, shaped the genome of this species along this latitudinal gradient.

Through the following chapters, I sought to answer a number of questions relating to the evolutionary history of this species through the analysis of neutral and adaptive genetic variation:

1. Does the distribution of genetic variation in this species reflect morphological variation across its Australian range? (Chapter three).
2. What is the diversity and distribution of genetic variation found within and among populations of *D. viscosa* ssp. *angustissima* distributed along a North-South environmental gradient within the Adelaide geosyncline? Can we infer population demographic history from these measures? (Chapters 4 and 5).
3. Can we detect signatures of environmental selection among these same populations in a set of targeted genes with *a priori* expectations they may be involved in responses to environment, such as water use and temperature resistance? (Chapter 6).

Thesis outline

Chapter 2: Constraints to and Implications for climate change adaptation in plants

This chapter is a review paper published in *Conservation Genetics*. It considers in what ways plant populations could respond to contemporary climate change and, in particular, what constrains these possible responses. Population responses are categorised into one of three types: migration, *in situ* adaptation, or extirpation. Constraints to each potential response, such as available space, rate of change, and presence of adaptive variation, are considered. The use of different methods to assess

adaptive responses is discussed with a focus on the use of genome scans and measuring populations distributed over environmental gradients. Implications for conservation are also considered, whereby different strategies for assisting plant populations to overcome adaptive constraints are suggested.

This chapter acts as a broad overview of a number of the issues considered in finer detail in the empirical chapters of the thesis, including: measuring population genetic diversity and adaptive capacity as a way of assessing the potential to persist under climate change; using environmental gradients as a surrogate for time, where warmer and drier regions of the gradient may reflect future conditions for currently cooler and wetter regions; assessing best conservation and restoration practice by incorporating information gained from genomic studies such as measures of gene flow, diversity and adaptive potential.

Chapter 3: Geography, not morphological variation, defines genetic partitioning in a genome-wide SNP screen of Dodonaea viscosa *(Sapindaceae)*

In Australia, *D. viscosa* displays a wide range of trait variation, particularly in leaf morphology. As a result, West (1984) provided a revision of the taxonomy of the species and divided it into seven subspecies, based mainly on differences in leaf morphology. Since this work, Harrington and Gadek (2009) attempted to determine the evolutionary history of the species through Bayesian analysis of nuclear ribosomal transcribed spacers and they divided the species into two distinct, geographically based, intraspecific lineages. In this chapter, I set out to build on this work and further test the subspecies classifications of West (1984) via the analysis of genome-wide SNP markers. Through a combination of genetic structure analysis and Bayesian-based and network-based phylogenetic analyses, I assessed relationships among a continent-wide sample of the species including representatives of all seven proposed subspecies.

This chapter serves two roles to the thesis. Firstly, previous taxonomic work on the species has not clearly defined the division (if any) of the different morphological varieties/subspecies across the continent from a genomics perspective. The use of genome-wide SNP markers may provide greater resolution than has previously been achieved when looking for phylogeographic signals. Secondly, consideration of continent-wide phylogeography helps to determine the appropriateness of the subspecies classifications across the study region of chapters four, five and six. Chapter four of the thesis only considers two of the subspecies (ssp. *angustissima* and *spatulata*) and chapters five and six focus on one subspecies (ssp. *angustissima*) distributed throughout the Adelaide geosyncline region within South Australia.

Chapter 4: Transcriptome sequencing, annotation and polymorphism detection in the hop bush, Dodonaea viscosa

One of the main aims of this thesis was to investigate genomic signatures of selection across an environmental gradient. In order to achieve this, it is ideal practice to have a reference genome in order to identify the specific location and function of genes under selection. The genome of *D. viscosa* is yet to be sequenced and the undertaking of genome sequencing, assembly and annotation was beyond the scope and budget of this project. Whilst the identification of random sections of the genome that display signatures of selection is possible and relatively straightforward, the absence of functional information reduces the utility of such a study. Therefore, I decided to take an RNAseq approach, whereby I sequenced RNA extracted from several individuals and assembled this into a transcriptome, providing a functional representation of the genome.

This chapter has been published in *BMC Genomics* and outlines the steps taken from RNA extraction to transcriptome assembly and annotation. The use of BLAST searches, whereby the resultant contiguous sequences (contigs) produced from the assembly

were matched to sequence data in the NCBI non-redundant database, meant that functional annotations were obtained for a large proportion of the transcriptome. This chapter was essential to the final two empirical chapters of the thesis in that it provided the reference genomic resources required for the targeted gene sequencing approach taken in these chapters.

Chapter 5: Determining the level and structure of population genomic variation for the narrow-leaf hopbush across a continental biodiversity refugium - the Adelaide Geosyncline

A targeted sequencing approach was taken to provide the raw data for chapters five and six. This involved the design of hybrid-capture baits based on the transcriptome sequences of ~1000 genes identified in chapter four. These baits were then used to selectively sequence these target genes in at least five individuals of each of 17 populations. By calling SNPs among the population samples I was able to generate a SNP dataset derived from both neutral and putatively adaptive loci distributed across the targeted genes. In this chapter I present the analysis of the neutral loci. Neutral loci were first identified using an F_{ST}-based outlier test and then used to derive measures of genetic diversity and population structure. This study was informative in its own right, as it allowed me to address hypotheses regarding dispersal to and from refugia during and since past climatic changes as well as to further consider the hypothesis that the Adelaide geosyncline has acted as a Pleistocene refugium. As well as this, the analysis performed in this chapter was essential to chapter six in that, in order to identify signatures of selection across the landscape, population genetic structure needed to be defined. If such structure is not taken into consideration then any signatures of selection identified may be confounded by similar signals resulting from neutral processes, leading to type I error (false positives).

Chapter 6: Finding needles in a genomic haystack: targeted sequencing to identify signatures of selection in a non-model species

This chapter had the overall aim of identifying signatures of selection driven by environmental factors. The development of the transcriptome (chapter four) and measurements of genetic structure and diversity (chapter five) allowed for a thorough assessment of functional genetic diversity along the gradient. I used three methods in order to look for signatures of selection throughout the sequenced genes, including the F_{ST}-outlier method of chapter five alongside two model-based regression methods that take geography and genetic structure into account. The availability of the published transcriptome (chapter three) meant that any SNPs displaying signatures of selection could be traced back to the gene they were present in and, therefore, potential functional consequences could be discussed.

This chapter demonstrates the use of a novel genomic method to generate functional genetic data for a non-model species in the absence of a reference genome. The data were then used to investigate whether environmental/climatic factors have acted as selection pressures on these populations in the past, leading to local adaptation. The implications of these findings for adaptation to future change and to restoration practice are discussed. It has been accepted for publication in the journal Molecular Ecology.

Chapter 7: Conclusion and future directions

The final chapter is an overall conclusion of the main findings in the thesis as well as a consideration of the issues faced and lessons learned, particularly from a sampling design and methodology perspective. The next steps or future directions that can be informed by and further confirm the findings of this thesis are also discussed.

Chapter 2 - Constraints to and conservation implications for climate change adaptation in plants

"Look at a plant in the midst of its range! Why does it not double or quadruple its numbers? We know that it can perfectly well withstand a little more heat or cold, dampness or dryness, for elsewhere it ranges into slightly hotter or colder, damper or drier districts. In this case we can clearly see that if we wish in imagination to give the plant the power of increasing in numbers, we should have to give it some advantage"

Darwin, 1859

Statement of Authorship

Title of Paper	Constraints to and conservation implications for climate change adaptation in plants
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published in Conservation Genetics, 2015, DOI: 10.1007/s10592-015-0782-5

Principal Author

Name of Principal Author (Candidate)	Matthew J Christmas
Contribution to the Paper	Planning of paper theme; conducted literature review, wrote manuscript as principal author.
Overall percentage (%)	80%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 9.12.15.

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Martin F Breed
Contribution to the Paper	Planning of paper theme; advised on focus and structure; edited manuscript. 10%
Signature	Date 8/12/15

Name of Co-Author	Andrew J Lowe
Contribution to the Paper	Planning of paper theme; advised on focus and structure; edited manuscript. 10%
Signature	Date 8.17.15

Christmas, M.J., Breed, M.F. and Lowe, A.J. (2015). Constraints to and conservation implications for climate change adaptation in plants. *Conservation Genetics*, 17(2), 305–320.

Chapter 3 - Geography, not morphological variation, defines genetic partitioning in a genome-wide SNP screen of Dodonaea viscosa (Sapindaceae)

"I am fully convinced that species are not immutable; but that those belonging to what are called the same genera are lineal descendants of some other and generally extinct species, in the same manner as the acknowledged varieties of any one species are the descendants of that species. Furthermore, I am convinced that natural selection has been the most important, but not the exclusive, means of modification."

Darwin, 1859

Statement of Authorship

Title of Paper	Geography, not morphological variation, defines genetic partitioning in a genome-wide SNP screen of Dodonaea viscosa (Sapindaceae)
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	This chapter has been prepared in manuscript style for submission for publication

Principal Author

Name of Principal Author (Candidate)	Matthew Christmas
Contribution to the Paper	Designed the study; carried out field collections of samples and obtained samples from herbarium; performed all lab work; analysed sequencing data; wrote manuscript as principal author
Overall percentage (%)	85%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 9.12.15

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Ed Biffin
Contribution to the Paper	Assisted with study design; assisted with field collections of samples; assisted with lab work and data analysis; advised on and edited manuscript. 10%
Signature	Date 9.12.15

Name of Co-Author	Andrew Lowe
Contribution to the Paper	As principal supervisor, obtained funding for the research; advised on study design; advised on development of the manuscript. 5%
Signature	Date 8.12.15

Geography, not morphological variation, defines genetic partitioning in a genome-wide SNP screen of Dodonaea viscosa (Sapindaceae)

Matthew J. Christmas[1], Ed Biffin[2] and Andrew J. Lowe[1]

[1]Environment Institute and School of Biological Sciences, The University of Adelaide, North Terrace, SA 5005, Australia

[2]State Herbarium of South Australia, Hackney Road, Adelaide, SA 5000, Australia

matthew.christmas@adelaide.edu.au

ed.biffin@adelaide.edu.au

andrew.lowe@adelaide.edu.au

Author for correspondence:

Andrew J. Lowe, tel +61 8 8313 1140, fax: +61 8 8303 4364, Email: andrew.lowe@adelaide.edu.au

Manuscript prepared for submission to The Journal of Biogeography

Abstract

The division of *Dodonaea viscosa* into seven distinct subspecies within Australia was based mainly on variations in leaf morphology. Whether the subspecies divisions are reflected in the distribution of genetic variation within the species complex is yet to be fully tested. Here, we used a genotyping-by-sequencing approach to genotype 67 a large set of 941 single nucleotide polymorphisms (SNPs) distributed across the genome of each subspecies. We looked for genetic structure amongst the samples and used network-based and Bayesian-based methods to assess phylogenetic relationships. Structure analysis identified two genetic clusters amongst the samples, with seven further clusters identified in the substructure of one of these clusters. Cluster membership only partially aligned with subspecies classifications and introgression between clusters appears likely, where genetic patterns are better explained by geography. The arid-zone subspecies *mucronata* showed the strongest evidence for a distinct genetic group but it did not form a monophyletic group in the Bayesian phylogenetic analysis. A consideration of a population's geographic location, environment and Pleistocene history appears to better explain the distribution of genetic variation across this species, rather than the presence of distinct gene pools corresponding to morphologically defined subspecies. The implication is that the major growth form and leaf characteristics previously used to classify these subspecies are plastic and/or environmentally controlled. The study provides an ideal demonstration of the power of phylogenomics to examine subspecies relationships in problematic taxonomic entities.

INTRODUCTION

The rapid advancement in sequencing technologies over the past decade is providing opportunities to assess the variation and spatial distribution of within-species genetic diversity at scales finer than previously possible (Mardis 2008). This advance in genomic techniques has enabled the fast and cost-effective generation of multilocus sequence data for addressing phylogeographic and phylogenetic questions (McCormack, et al. 2011). The ability to sequence multiple loci for non-model species without the laborious tasks involved in the development and sequencing of more traditional molecular markers (e.g. microsatellites) is clearly attractive and provides the potential to genotype many more individuals for many more loci in a single study.

In this study, we apply a next-generation sequencing technique for marker discovery in order to investigate phylogeographic relationships in the hopbush, *Dodonaea viscosa* (L.) Jacq. *D. viscosa,* a member of the Sapindaceae family, is a widely distributed, highly variable dioecious woody shrub species found on six continents. Within Australia, where it is thought to have originated (Harrington and Gadek 2009), it is distributed throughout the continent inhabiting a wide range of habitat types and with a diversity of growth forms: in arid shrub lands, desert gullies, and at high elevations as a prostrate shrub; in temperate forests as a shrub or small tree; and in littoral regions next to mangroves. Its wide distribution is partly attributable to seed characteristics and dispersal mechanisms. Seeds are contained in winged capsules which assist wind dispersal as well as via overland water flow (West 1980). Seeds also have physical dormancy which can be overcome by heat (Ralph 2003; Baskin, et al. 2004) and this, coupled with their ability to float, would assist long distance dispersal over water (West 1980). Its diversity in form and habitat gives it high ecological significance and it is widely used in revegetation projects. Culturally, it has been valued for a wide range of uses (Ghisalberti 1998) and so current distributions may in part have been influenced by human-assisted dispersal. As a

result of its wide distribution and polymorphic nature it has undergone a number of taxonomic assessments (Bentham 1863; Radlkofer 1900, 1933; Sherff 1945, 1947; West 1980, 1984; Harrington and Gadek 2009).

In a revision of *Dodonaea*, West (1984) recognised seven *D. viscosa* subspecies based mainly on leaf characters, namely ssp. *viscosa*, ssp, *burmanniana*, ssp. *angustifolia*, ssp. *angustissima*, ssp. *cuneata*, ssp. *mucronata*, and ssp. *spatulata*. Significant disjunctions between subspecies and the presence of what appeared to be ecotypes were thought to be enough to justify formal recognition of the subspecies (West 1980). However, West (1984) did note that, in around 30% of the populations examined, there was evidence of full to partial intergradation between subspecies.

In their treatment of the *D. viscosa* complex based on phylogenetic analyses of nuclear ribosomal ITS and ETSf sequences, Harrington and Gadek (2009) identified two geographically based intraspecific lineages, which they estimated to have diverged from a common ancestor 1.1-2.1 mya, during the late Pliocene to early Pleistocene. Harrington and Gadek (2009) stated that the distinct entities identified by the genetic markers do not have consistent, distinct morphological characteristics, suggesting that it is not useful to try to define subspecies based on morphology. They suggested the term 'ochlospecies' (White 1962) be used when considering *D. viscosa* in order to take into account its variable nature across its range. However, sampling in their study was very limited and firm conclusions about the status of the subspecies could not be made.

In this study, we have generated a large dataset of genome-wide single nucleotide polymorphic (SNP) molecular markers using a genome complexity reduction method (Lischer, et al. 2013; Van Dijk, et al. 2014), to genotype 67 individuals collected across the Australian range. Phylogeographic relationships within the Australian *Dodonaea viscosa* complex were then investigated through the analysis of the SNP markers. We use Bayesian assignment tests and phylogenetic analyses of the SNP data in order to look for

evidence of genetic divergence and structure within the complex to test whether the phenotypically-based subspecies assignments are reflected in the genetic variation across the species. This study serves as an extension of the work by Harrington and Gadek (2009) and furthers our understanding of the demographic history and geographic connectivity in this widely dispersed, highly variable species across the Australian continent.

MATERIALS AND METHODS

Sampling

A combination of field collected (13) and herbarium (54) specimens was used (table 1). Herbarium specimens were sampled from collections at the State Herbarium of South Australia (AD). Specimens were selected with the aim of achieving a broad, representative sample, which encompassed all seven subspecies across their continental range (fig. 1).

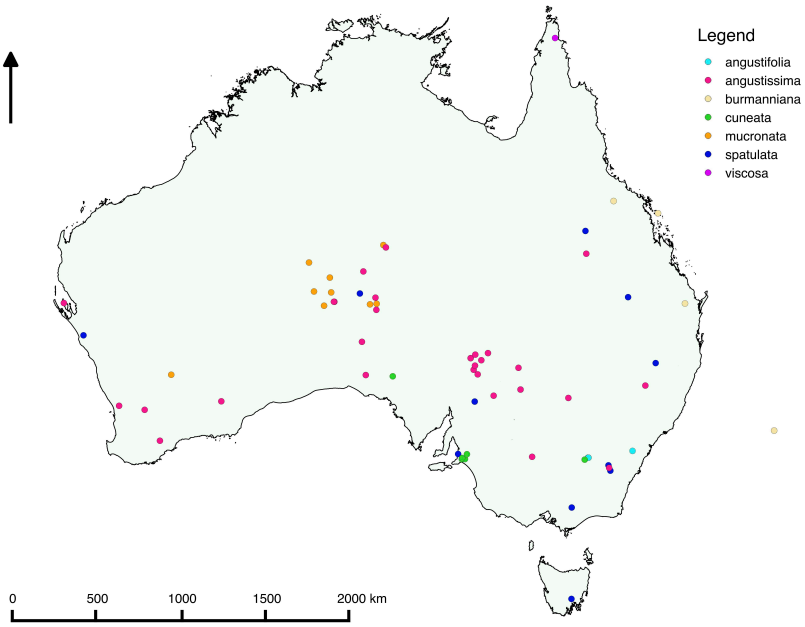

Figure 1. Sample locations of the 67 *Dodonaea viscosa* individuals used in the study.

Table 1. Details of 67 samples used in this study including subspecies assignments after West (1984). H = herbarium sample (AD); F = field collected sample

ID	Subspecies	Herbarium?	Accession number	Latitude	Longitude
Af1A	*angustifolia*	H	AD97822008	-34.67	147.25
Af1B	*angustifolia*	H	AD98514127	-33.98	150.12
Af1C	*angustifolia*	H	AD97745473	-29.15	151.67
A1D	*angustissima*	H	AD246036	-31.53	145.49
A1E	*angustissima*	H	AD154965	-30.14	142.09
A1F	*angustissima*	H	AD162684	-30.33	150.34
A1H	*angustissima*	H	AD99423236	-31.32	142.34
A2A	*angustissima*	H	AD119997	-34.95	143.48
A2B	*angustissima*	H	AD230708	-31.73	116.12
A2C	*angustissima*	H	AD220089	-33.90	118.53
A2D	*angustissima*	H	AD219732	-32.12	117.75
A2E	*angustissima*	H	AD140217	-32.04	122.80
A2F	*angustissima*	H	AD206102	-25.75	113.52
A2G	*angustissima*	H	AD113291	-26.63	132.86
A2H	*angustissima*	H	AD122436	-31.76	140.62
A3A	*angustissima*	H	AD98623029	-27.30	132.92
A3B	*angustissima*	H	AD183107	-29.05	132.01
A3C	*angustissima*	H	AD192402	-30.87	132.26
A3D	*angustissima*	H	AD97343170	-25.20	132.10
A3E	*angustissima*	H	AD99128113	-35.08	148.72
A3F	*angustissima*	H	AD99332276	-23.57	145.65
A3G	*angustissima*	H	AD199698	-29.45	140.07
A3H	*angustissima*	H	AD199697	-26.85	130.26
A4A	*angustissima*	H	AD99120046	-30.28	150.16
A8D	*angustissima*	F	Unassigned	-30.18	139.29
A8E	*angustissima*	F	Unassigned	-30.65	139.50
A8F	*angustissima*	F	Unassigned	-30.41	139.22
A8G	*angustissima*	F	Unassigned	-29.86	139.68
A8H	*angustissima*	F	Unassigned	-29.58	139.27
A9B	*angustissima*	F	Unassigned	-29.78	138.98
A9C	*angustissima*	F	Unassigned	-29.78	138.98
A9D	*angustissima*	F	Unassigned	-23.88	133.47
B4B	*burmanniana*	H	AD99611049	-20.88	149.62
B4C	*burmanniana*	H	AD99330408	-20.53	146.92
B4D	*burmanniana*	H	AD98440041	-25.57	152.03
B4E	*burmanniana*	H	AD97850134	-31.52	159.07
C4F	*cuneata*	H	AD173647	-30.93	134.01
C4G	*cuneata*	H	AD199853	-35.34	138.98
C4H	*cuneata*	H	AD155971	-35.06	139.09
C5A	*cuneata*	H	AD119758	-35.29	138.78
C5B	*cuneata*	H	AD98213004	-34.80	147.02
C5C	*cuneata*	H	AD119754	-35.29	138.78
C9F	*cuneata*	F	Unassigned	-35.36	138.77

M5D	*mucronata*	H	AD156074	-30.35	119.70
M5E	*mucronata*	H	AD199695	-24.68	128.77
M5F	*mucronata*	H	AD237868	-25.52	130.03
M5G	*mucronata*	H	AD99334050	-27.00	132.52
M5H	*mucronata*	H	AD99602542	-26.34	130.10
M6A	*mucronata*	H	AD110100	-26.97	132.93
M6B	*mucronata*	H	AD121946	-27.06	129.64
M6C	*mucronata*	H	AD109632	-26.27	129.04
M6D	*mucronata*	H	AD99423215	-31.32	142.34
M6E	*mucronata*	H	AD99507206	-26.05	130.33
M8A	*mucronata*	F	Unassigned	-23.75	133.31
S6F	*spatulata*	H	AD99308019	-22.33	145.45
S6G	*spatulata*	H	AD99206116	-25.67	148.50
S6H	*spatulata*	H	AD157725	-34.95	148.65
S7A	*spatulata*	H	AD106734	-35.23	148.82
S7B	*spatulata*	H	AD97744484	-27.65	114.45
S7C	*spatulata*	H	AD99431336	-29.03	150.78
S7D	*spatulata*	H	AD237612	-26.41	131.89
S7E	*spatulata*	H	AD98621287	-42.62	147.25
S7F	*spatulata*	H	AD99448235	-37.52	146.52
S8C	*spatulata*	F	Unassigned	-35.08	138.50
S9E	*spatulata*	F	Unassigned	-34.94	139.08
S9G	*spatulata*	F	Unassigned	-35.21	139.04
V7H	*viscosa*	H	AD99332162	-11.80	142.67

DNA extraction, library preparation and sequencing

A 1 cm^2 section of leaf tissue was cut from each individual. DNA was then extracted at the Australian Genome Research Facility (AGRF), Adelaide. Reduced-representation libraries of genomic DNA were prepared for each individual following a method similar to that employed for AFLP libraries (Vos, et al. 1995). This involved a restriction-digest step whereby two restriction enzymes (*Mse*I and *Eco*RI) digest the DNA. Double-stranded adapters (*Eco*RI and *Mse*I), along with T4 ligase, are then added and left over night to ligate. A pre-selective PCR, using the adapters as priming sites and PCR primers with one selective base, is then carried out. This reduces the genome to only those fragments that had both the *Eco*RI and *Mse*I adapters present as well as the selective base in the *Eco*RI priming site. The pre-selective PCR products were then diluted and amplified again using

fusion primers. A selective PCR was then carried out, using an *Mse*I primer with three selective bases alongside an *Eco*RI primer with 2 selective bases and individual barcodes. Libraries were then pooled and a clean-up step using 0.8x AMPure (Beckman Coulter) was carried out. Fragments of 250-300bp were selected using an E-Gel SizeSelect (Life Technologies). Selected fragments underwent 15 rounds of PCR followed by an AMPure clean-up. The prepared sequencing library was then quantified using a Qubit (Life Technologies) and a Tapestation (Agilent Technologies) prior to sequencing. Pooled libraries were sequenced on the Ion Torrent Proton (Life Technologies) at the Institue of medical and veterinary science (IMVS), Adelaide.

Raw sequencing data pre-processing

Sequencing reads were imported into CLC Genomics Workbench v6.5.2 (CLC) (http://www.clcbio.com). Reads were demultiplexed to separate out reads per individual. Reads were then trimmed to remove barcodes and adapter sequences and filtered on length (minimum length: 40 bp) and quality (cut off: 0.05).

Reference assembly and SNP calling

Reads that passed the pre-processing steps were combined to create a *de novo* assembly in CLC with the following settings: Mapping mode = map reads back to contigs and update contigs; Automatic bubble size; Min contig length = 50; Automatic word size; Perform scaffolding; Mismatch cost = 2; Deletion cost = 2; Length fraction = 0.5; Similarity fraction = 0.9. The assembled contigs were used as a reference for finding SNP loci. The reference, along with the cleaned and trimmed sequencing reads, were input into the GATK (Broad Institute) pipeline of programs for variant calling. Briefly, each individual's sequencing reads were mapped separately to the reference using BWA. Indel realignment and base recalibration steps were then carried out using the respective GATK tools. Variant calling via the UnifiedGenotyper tool was then used to produce a set of raw genotypes for each individual. Here, the standard minimum phred-scaled

confidence threshold for calling a variant (-stand_call_conf) was set at 30 and the minimum base quality score required to consider a base for calling (-mbq) was set at 20.

SNP Filtering

A SNP variant is called at any point where there is a mismatch between the reference and the mapped read. However, as mismatches can occur for a number of reasons, such as sequencing errors or assembly errors, the output SNP set must undergo rigorous filtering to ensure the resultant SNPs are valid and reliable for downstream analysis. A set of hard filters was employed to filter out the low quality SNPs using the Variant Filtration tool in the GATK. Full details of filter settings can be found in supplementary information S1. To minimise the influence of linkage between SNPs, one SNP per contig was selected at random from the filtered SNP set using a custom script. SNPs and individuals were also filtered on percentage of missing data, whereby SNPs with >50% missing data across all individuals and individuals with >50% missing data were removed. SNPs were then concatenated into one sequence per individual for downstream analysis.

Genetic structure analysis

Individual assignments to genetic clusters were conducted using STRUCTURE (Pritchard, et al. 2000) and discriminant analysis of principal components (DAPC; Jombart, et al. 2010). In STRUCTURE the admixture model was utilised to determine the number of clusters (K) from 1 through 10 using a burn-in of 200,000 followed by 1,000,000 iterations. The LOCPRIOR model was used for incorporating subspecies assignments and geographical locations on separate runs. Here, data were either coded by their subspecies or by their sampling location. Analyses were repeated 10 times for each K value. The average and standard deviation of the likelihood of each model were used to calculate ΔK (Evanno, et al. 2005) in Structure Harvester (Earl and vonHoldt 2012), and the K value with the highest ΔK was selected as the model with the most

support. This was repeated to look for substructure within clusters identified from the first structure run. Replicate analyses for each K value were assessed in CLUMPP (Jakobsson and Rosenberg 2007).

DAPC was implemented in the R package `adegenet` (Jombart 2008). DAPC is a non-model-based multivariate approach, which seeks discriminating functions between groups of individuals whilst minimising variation within clusters. Firstly, principal component analysis (PCA) was used to transform the genetic data into uncorrelated components. The number of genetic clusters was then defined using K-means, a clustering algorithm that looks for the value of K that maximises variation between groups. The Bayesian Information Criterion (BIC) was calculated for all values of K 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` in order to efficiently describe the genetic clusters. Membership probabilities of each individual to the identified clusters are also provided.

Phylogeny, divergence and biogeography

Genetic distances between individuals were estimated using the 'genpofad' algorithm implemented in POFAD (Joly, et al. 2014). Distances were standardised and a Jukes-Cantor correction for multiple hits was used. A phylogenetic network based on these distances was constructed in SplitsTree4 (Huson and Bryant 2006) using the Neighbor-Net algorithm (Bryant and Moulton 2004). This is a distance based method of constructing phylogenetic networks based on the Neighbor-Joining algorithm of Saitou and Nei (1987). Least squares fit index (LSfit) was computed to check how well the network represented the distances.

A Bayesian phylogenetic tree was generated in BEAST v.2.1.3 (Drummond and Rambaut 2007) with a GTR substitution model, lognormal relaxed clock model and Yule

priors on the branching pattern. A Markov chain length of 1×10^8 steps was used with an initial burn-in of 2×10^6 discarded. Convergence of parameters and across independent runs was assessed using Tracer version 1.6. Three runs were performed with the same settings and TreeAnnotator, part of the BEAST package (Drummond and Rambaut 2007), was used to combine outputs and visualise the maximum clade credibility phylogeny.

We ran SDIVA analysis using RASP v.3.1 (Yu, et al. 2015) to infer the biogeographic history of *D. viscosa* based on the phylogeny constructed in BEAST. Here, we defined eight distribution regions for the samples, based roughly on the areas of endemism used in Crisp, et al. (1995): the South East (SE; 6 samples), Eastern Queensland (EQ; 12 samples), Eastern desert (ED; 5 samples), Adelaide (AD; 16 samples), the South West (SW; 8 samples), Western Desert (WD; 18 samples), Tasmania (TA; 1 sample), and Cape Yorke (CY; 1 sample) (FIGURE?). We loaded 12,501 trees produced in BEAST, the condensed tree from TreeAnnotator, as well as a file defining the sample distribution regions. SDIVA was ran with the 'Allow Extinction' and 'Allow Reconstruction' options selected.

RESULTS

Sequencing and pre-processing

Proton sequencing of the pooled AFLP-seq libraries generated 75.6 million reads. The distribution of reads across the 67 samples was very uneven, with the maximum number of reads for one sample at 2.5 million and the minimum at 61, 284. The average number of reads per sample was 712, 136. The absence of distinguishing barcodes on 20.8 million reads (27%) meant these reads were excluded, as they could not be assigned to an individual. Following removal of barcode and adapter sequences and quality trimming, 54,834,835 reads remained for use in downstream analysis.

Reference assembly and SNP calling

De novo assembly of cleaned and filtered reads returned 4,585 contigs to act as a reference for variant calling. Average contig length was 97.14 bp with a maximum length of 495 bp, a minimum length of 30 bp and an N50 of 109 bp. The SNP calling pipeline implemented in the GATK plus subsequent filtering resulted in 941 SNPs called across 67 individuals. The percentage of missing data across all individuals was 25% following removal of loci with > 50% missing data and individuals with > 50% missing data (ten omitted).

Genetic structure analysis

Structure analysis of the 67 individuals in STRUCTURE using the ΔK method (Evanno, et al. 2005) revealed the most likely number of genetic clusters, *K*, to be 2 (ΔK = 304.08). This placed 17 ssp. *angustissima* individuals from mostly central regions into one cluster and the remaining 50 individuals into another cluster. The DAPC analysis also identified K=2 to be the most likely number of clusters, as assessed by BIC values (supplementary figure **X**). The same individuals were assigned to each of the two clusters as in the STRUCTURE analysis.

Substructure within the two clusters was also analysed. No substructure was revealed amongst the 18 individuals in cluster one, whereas for cluster two ΔK was highest at K = 7 (ΔK = 401.31) (fig. 2). When subspecies assignment was used in the LOCPRIOR model ΔK was also highest at K = 7 but ΔK was low (16.00) and clusters did not represent the subspecies assignments. With geographic location used as a prior, ΔK was highest at K = 2 (ΔK = 486.92), yet the clusters were not geographically separated, with much overlap. No further substructure was identified in the DAPC analysis.

Greatest likelihood was achieved with no LOCPRIOR and cluster assignment seemed to align with a combination of geographic distribution and subspecies classification, with southeastern sspp. *spatulata, cuneata,* and *angustifolia* samples forming a cluster, central ssp. *mucronata* forming a cluster, a cluster represented by southwestern sspp.

angustissima and *spatulata*, another of eastern ssp. *angustissima*, the four ssp. *burmanniana* samples along with the northeastern ssp. *viscosa* sample and a northeastern *angustissima* sample forming a cluster, a cluster comprised of eastern *spatulata* and *angustifolia* samples, and a south central cluster of sspp. *angustissima*, *cuneata*, and *spatulata* (fig. 2).

Phylogeny, divergence and biogeography

Average standardised distance between samples was 0.48, with a maximum pair-wise distance of 0.71. The neighbor-net network produced in SplitsTree4 showed several groupings, which generally align with the structure cluster definitions (fig. 3). The least squares fit value of 98.8% demonstrates that the network is a good fit to the data and most of the branches had strong bootstrap support values of 70-100% (data not shown). The length of the edge to the *D. viscosa* ssp. *viscosa* sample suggests that this sample is very strongly split from the rest of the network.

Bayesian phylogenetic analysis of the data using BEAST produced a tree that was very comparable to the neighbor-net network (fig. 2). On the whole, neither the tree nor the network provided much support for the subspecies classifications. Low posterior probabilities for a significant number of the branching events suggest that there is low clade support throughout the tree. This is as would be expected if the species is not highly genetically divergent, despite divergence in phenotype seen among subspecies. Homoplasy between isolated groups may also obscure signals within the data. However, a number of branching events are statistically well supported (particularly among shallow divergences), such as the split between southeastern and southwestern ssp. *angustissima* samples (posterior probability = 0.979). The evolutionary relationships suggested in the Bayesian tree again closely parallel the clusters identified in the structure analysis (fig 2).

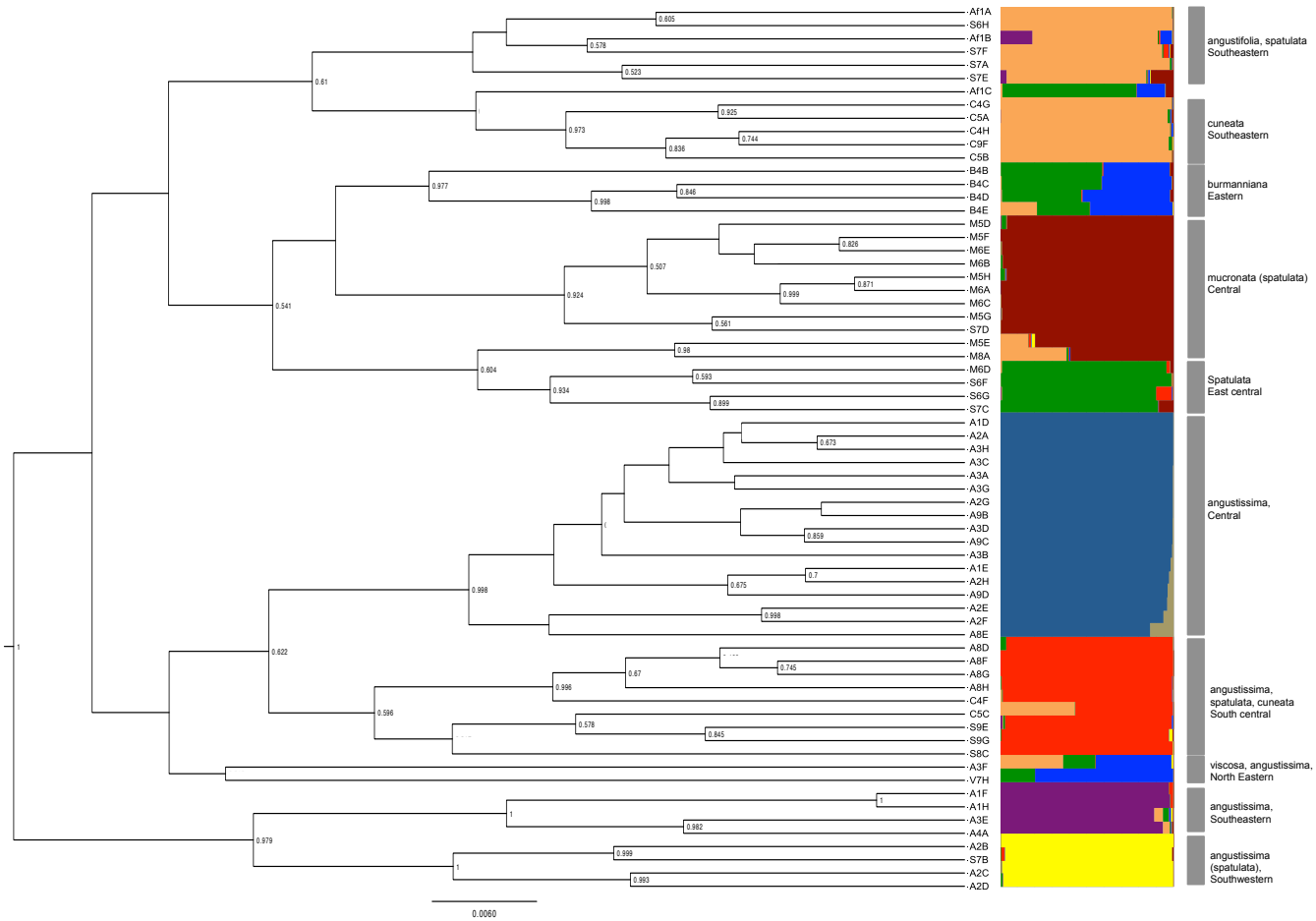

Figure 2. Bayesian phylogenetic tree outputted from BEAST. Posterior probabilities of greater than 0.5 are shown. STRUCTURE output for each individual shown alongside their branch. Labels indicate taxonomic subspecies classifications as well as geographic locations.

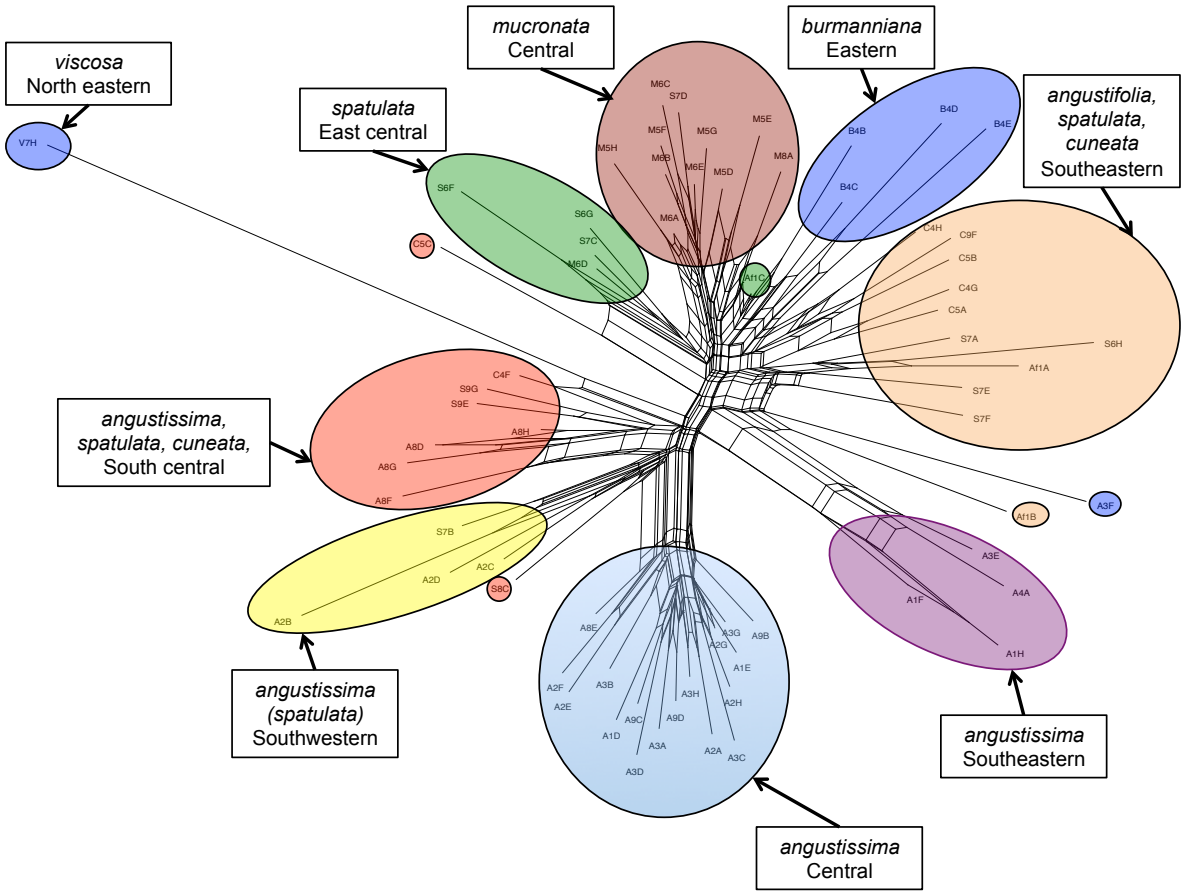

Figure 3. Neighbor-net network implemented in SplitsTree4. Branch lengths are proportional to the weight of the splits. The network has a least squares fit value of 98.8%, showing it to be a good fit of the data. Coloured circles represent assigned clusters (>50% assignment) in the STRUCTURE analysis.

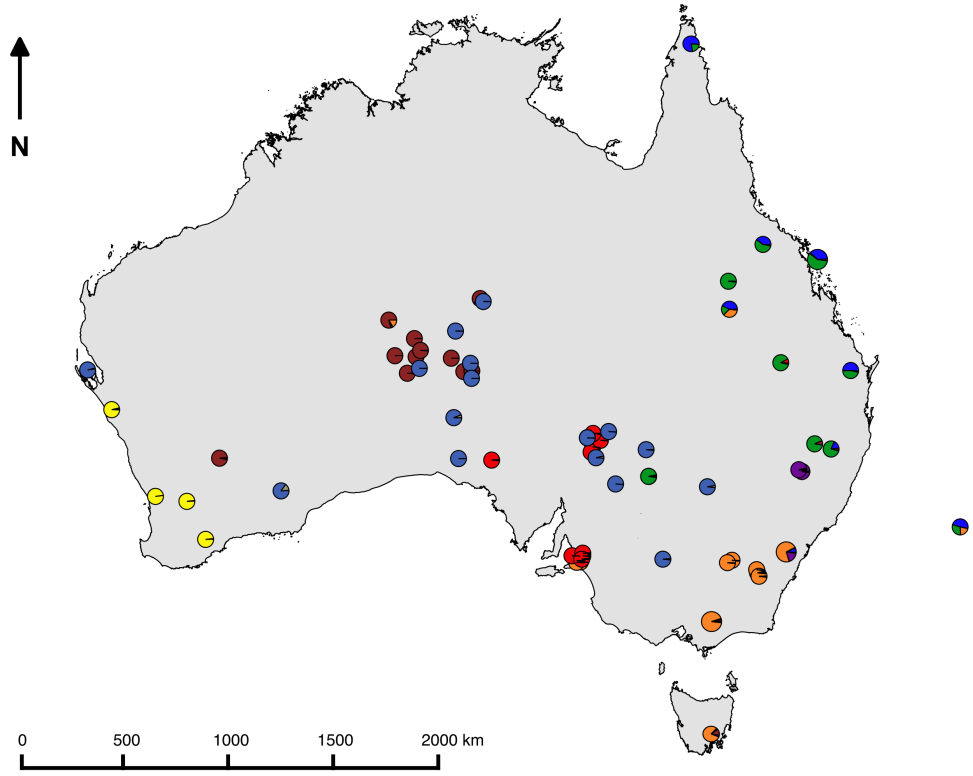

Figure 4. Distribution map of samples with pie charts representing cluster assignment from STRUCTURE output for each individual. Colouring matches that used in figures 2 and 3.

DISCUSSION

The morphological diversity and continental (and beyond) distribution of *D. viscosa* has made it a taxonomically difficult entity to study, undergoing several reclassifications over the years. Phylogenomics and the advent of next-generation sequencing provide researchers with more power than ever to address questions of taxonomic resolution and evolutionary history. A reduced representation library approach to marker discovery, such as the method used in this study, means that a large number of genetic markers dispersed throughout the genome can be screened rapidly, easily and relatively cheaply, in the absence of any prior genomic resources, making such data generation ideal for non-model organisms (McCormack, et al. 2011). Even with conservative filtering steps, we have generated a set of 941 SNP markers with high representation

across 67 individuals from one run on an Ion Proton sequencing machine. These types of data are becoming more common in phylogenetic and phylogeographic studies (e.g. Emerson, et al. 2010; Gohli, et al. 2014) and the sheer number of markers generated, when compared with more traditional molecular markers such as microsatellites or single/linked gene sequencing, helps to overcome sampling bias and potentially reduce the effect of loci under selection by providing a more representative sample of genomic variation (Spinks, et al. 2014). Using these new methods in *D. viscosa* has demonstrated that lineages defined by leaf morphology do not align well with patterns of genetic structure, and the separation of the species into distinct morphologically-defined subspecies is not supported.

Comparison with previous subspecies framework

Previous work on *D. viscosa* has used genetic markers to examine phylogeography and here we use new molecular techniques to delve further into the story of this diverse and widely dispersed species. In their study of the *D. viscosa* species complex through the analysis of nuclear ribosomal transcribed spacer sequences, Harrington and Gadek (2009) divided the species into two groups: group I contained all the extra-Australian samples (except for those from New Zealand) as well as the north eastern ssp. *viscosa* and ssp. *burmanniana*. They proposed that climate-driven vicariance during the Pleistocene would have restricted gene flow across the McPherson Range (an extensive montane region of wet forest along the central east coast) between northern tropical populations of *D. viscosa* and populations found in the southern temperate and more arid areas. Our results support this theory, to an extent, with the single ssp. *viscosa* sample showing high divergence from all other samples as demonstrated by the long branch in the network analysis, and samples of ssp. *burmanniana* clustering together in the structure analysis (fig. 2), splits network (fig. 3), and also forming a monophyletic group (albeit with low posterior support value) in the Bayesian tree (fig. 2). However,

the ssp. *burmanniana* samples show evidence of admixture with northern (ssp. *viscosa*) and southern (ssp. *spatulata*) groups, despite no previous evidence of intergraded morphologies (West 1984). This may suggest that the McPherson ranges are now a contact zone between previously isolated populations.

Harrington and Gadek (2009) grouped the remaining subspecies into one clade, with support for three separate evolutionary lineages: ssp. *angustissima* in a lineage with the New Zealand samples; ssp. *spataulata, cuneata,* and *angustifolia* in a lineage with *D. procumbens*; and ssp. *mucronata* in a lineage with *D. biloba*. Our findings provide some support for these previously identified lineages. In the structure analysis of all samples, a subset of 17 of the ssp. *angustissima* samples formed a distinct cluster. These samples were mostly located in central and eastern regions, except for two (ssp. *angustissima* 2E and 2F), which were from the west. The ssp. *mucronata* samples also formed a distinct cluster in the secondary structure analysis, as well as grouping together in the splits network (fig. 3). It is interesting to note that West (1984) found no evidence of integradation between subspecies in the arid zone, and the genetic signal from the ssp. *mucronata* samples is consistent with this hypothesis, with all ssp. *mucronata* samples forming a distinct group relative to closely located ssp. *angustissima* samples.

The south-western cluster of three ssp. *angustissima* and one ssp. *spatulata* samples represent the four most geographically isolated samples. Members of this cluster formed a single clade in the Bayesian analysis, with high posterior probability (0.979; fig. 2), as well as forming a cluster in the splits network (fig. 3). In the structure analysis, these four samples had between 40-50% assignment to the cluster of 17 ssp. *angustissima* samples. This makes sense in terms of their distribution: apart from the ssp. *mucronata* samples which appear to stand alone, the 17 ssp. *angustissima* samples are closest geographically to these more isolated south-western samples (fig. 4).

The south-eastern cluster of ssp. *angustifolia*, *spatulata*, and *cuneata* samples, supported by all three sets of analyses, align with Harrington and Gadek (2009) IIb lineage. There is, however, a separate cluster of south-central ssp. *cuneata* and *spatulata* samples, along with four *angustissima* samples, which overlap in distribution yet are genetically dissimilar to the south-eastern samples. This could be the result of expansion from separate refugia during Pleistocene climate oscillations, one from the east and one from the west. *D. viscosa* is also widely used in revegetation programs and has been culturally used for a range of purposes (Ghisalberti 1998), so there is the possibility that movement of plants across the landscape has resulted in mixed genetic signals.

Two further clusters, one of the four southeastern ssp. *angustissima* samples and one comprising three samples of ssp. *spatulata* plus a ssp. *mucronata* and a ssp. *angustifolia* sample, all with an eastern distribution, were also identified in the analysis.

Is geography the answer?

West (1984) noted that intergrades between subspecies were present, particularly in the higher-rainfall southern temperate areas. It was even observed that in some populations leaf morphology appeared continuous from ssp. *cuneata* through ssp. *spatulata* to ssp. *angustissima*. Such integrades make for particularly tricky field identification, especially when the only determining character is leaf shape. Leaf shape has been shown to vary in response to environment in this species (Guerin and Lowe 2012; Guerin, et al. 2012) and so may be an unreliable defining characteristic. This fact, along with the evidence we present here of genetic clustering which does not adhere to subspecies classifications, suggest that splitting of the species into subspecies based on leaf traits is not wholly informative. In fact, the labelling of samples to subspecies based on leaf morphology can be quite misleading particularly where intergrades between leaf forms are found. This may have had some influence on the findings presented here: some of the specimens used may have been difficult to place in a subspecies (this was

the case for a number of the herbarium samples) contributing to the mixture of subspecies found in a number of the genetic groups.

In a review of evidence for refugia during the Pleistocene in southern Australia, Byrne (2008) suggested two explanations for patterns of divergence within species resulting from climate oscillations during this period: during early to mid Pleistocene climatic cycles, major contractions to geographically isolated refugia followed by subsequent expansion during favourable conditions may explain the presence of geographically structured, highly divergent lineages. In contrast, during later Pleistocene cycles, multiple localised refugia throughout the species' range with limited migration and only localised contraction and expansion, as well as extinction, may explain highly localised haplotypes within lineages (Byrne 2008). These scenarios potentially offer a better explanation of the observed genetic variation within the *D. viscosa* complex in Australia. Strong cooling in the Southern Ocean starting during the Miocene led to increased seasonality and aridity in southern Australia (Gentilli 1971). During the LGM, aridity reached its peak in southern Australia around 16-18,000 years ago (Bowler 1982), leading to desert formation in central Australia. Coastal areas such as the southeast and southwest, which would have maintained higher humidity than the arid centre, may therefore have acted as refugia to mesic adapted species. The genetic differentiation among the western samples, the central cluster and the eastern samples of *D. viscosa* may therefore reflect separate contractions into geographically isolated southern refugia, followed by late Pleistocene expansion. A growing number of studies are providing evidence towards the effects of Pleistocene climate oscillations on contemporary species distributions (Wheeler and Byrne 2006; Edwards, et al. 2007; Toon, et al. 2007; Neaves, et al. 2012; Weber, et al. 2014) and similar patterns of contractions to and expansions from more mesic areas have been found for Australian magpies (Toon, et al. 2007) and Myobatrachid frogs (Edwards, et al. 2007). Late Pleistocene more-localised refugial contractions over heterogeneous landscapes

followed by subsequent range expansions may have resulted in further genetic structure identified in this study, such as the separation of the central arid-dwelling ssp. *mucronata* from populations in the more temperate regions of its distribution. A combination of molecular dating and climatic modelling could be used to test these hypotheses further.

CONCLUSIONS

The morphological diversity and continental (and beyond) distribution of *D. viscosa* has made taxonomy difficult and, as a result, the species has undergone several reclassifications over the years. Phylogenomics and the advent of next-generation sequencing are providing researchers with more power than ever to address questions of species evolutionary history. Using these new methods in *D. viscosa* has demonstrated that lineages defined by leaf morphology do not align well with the spatial partitioning of the neutral genetic variation presented here, and the separation of the species into distinct subspecies is not well supported. Instead, genomic variation appears to be more geographically structured and could potentially be interpreted to support distribution changes (range contraction and expansion) during Pleistocene climate oscillations. This would align with estimates for the divergence of *D. viscosa* from its most recent common ancestor, which it shared with *D. camfieldii*, during the Late Pliocene to Early Pleistocene (2.7-1.4 Ma; Harrington and Gadek 2009). Temperate populations exhibit a wide diversity of leaf morphologies within and among populations, suggesting that these characters, previously used to define subspecies, are plastic and/or environmentally defined (Guerin and Lowe 2012; Guerin, et al. 2012).

ACKNOWLEDGMENTS

We thank the State Herbarium of South Australia and particularly Helen Vonow for access to herbarium specimens. Thanks also goes to the Australian Research Council for funding support (LP110100721 awarded to AJL), the South Australian Premier's Science and Research Fund awarded to AJL, and the Field Naturalist Society of South Australia Lirabenda Endowment Fund awarded to MJC.

REFERENCES

Baskin JM, Davis BH, Baskin CC, Gleason SM, Cordell S. 2004. Physical dormancy in seeds of *Dodonaea viscosa* (Sapindales, Sapindaceae) from Hawaii. Seed Science Research 14:81-90.
Bentham G. 1863. 1878 Flora australiensis. In: London: Lovell Reeve.
Bowler J. 1982. Aridity in the late Tertiary and Quaternary of Australia. Evolution of the flora and fauna of arid Australia:35-45.
Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Molecular biology and evolution 21:255-265.
Byrne M. 2008. Evidence for multiple refugia at different time scales during Pleistocene climatic oscillations in southern Australia inferred from phylogeography. Quaternary Science Reviews 27:2576-2585.
Crisp MD, Linder HP, Weston PH. 1995. Cladistic biogeography of plants in Australia and New Guinea: congruent pattern reveals two endemic tropical tracks. Systematic biology 44:457-473.
Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology 7:214.
Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4:359-361.
Edwards DL, Roberts J, Keogh JS. 2007. Impact of Plio‐Pleistocene arid cycling on the population history of a southwestern Australian frog. Molecular Ecology 16:2782-2796.
Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proceedings of the National Academy of Sciences 107:16196-16200.
Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14:2611-2620.
Gentilli J. 1971. Dynamics of the Australian troposphere. World Survey of Climatology: Climates of Australia and New Zealand:53-117.
Ghisalberti E. 1998. Ethnopharmacology and phytochemistry of Dodonaea species. Fitoterapia 69:99-113.
Gohli J, Leder E, Garcia‐del‐Rey E, Johannessen LE, Johnsen A, Laskemoen T, Popp M, Lifjeld JT. 2014. The evolutionary history of Afrocanarian blue tits inferred from genome‐wide SNPs. Molecular Ecology.
Guerin GR, Lowe AJ. 2012. Leaf morpholgy shift: new data and analysis support climate link. Biology Letters:rsbl20120860.
Guerin GR, Wen H, Lowe AJ. 2012. Leaf morphology shift linked to climate change. Biology Letters 8:882-886.

Harrington MG, Gadek PA. 2009. A species well travelled–the *Dodonaea viscosa* (Sapindaceae) complex based on phylogenetic analyses of nuclear ribosomal ITS and ETSf sequences. Journal of Biogeography 36:2313-2323.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Molecular biology and evolution 23:254-267.
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.
Joly S, Bryant DJ, Lockhart PJ. 2014. Flexible methods for estimating genetic distances from nucleotide data. bioRxiv:004184.
Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405.
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.
Lischer HE, Excoffier L, Heckel G. 2013. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of Microtus voles. Molecular biology and evolution:mst271.
Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends in Genetics 24:133-141.
McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2011. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular phylogenetics and evolution.
Neaves LE, Zenger KR, Prince RI, Eldridge MD. 2012. Impact of Pleistocene aridity oscillations on the population history of a widespread, vagile Australian mammal, Macropus fuliginosus. Journal of Biogeography 39:1545-1563.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.
Radlkofer L. 1900. Sapindaceae. Flora Brasiliensis 13:639-645.
Radlkofer L. 1933. Sapindaceae. Leipzig: Verlag von Wilhelm Engelmann.
Ralph M. 2003. Growing Australian native plants from seed.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4:406-425.
Sherff EE. 1947. Further studies in the genus Dodonaea: na.
Sherff EE. 1945. Some Additions to the Genus dodonaea L.(Fam. Sapindaceae). American Journal of Botany:202-214.
Spinks PQ, Thomson RC, Shaffer HB. 2014. The advantages of going large: genome‐wide SNPs clarify the complex population history and systematics of the threatened western pond turtle. Molecular Ecology 23:2228-2241.
Toon A, Mather PB, Baker AM, Durrant KL, Hughes JM. 2007. Pleistocene refugia in an arid landscape: analysis of a widely distributed Australian passerine. Molecular Ecology 16:2525-2541.
Van Dijk K, Waycott M, Biffin E, Cross H. 2014. Population genetic analysis of *Eucalyptus paludicola.* Final report to Department of Environment, Water and Natural Resources (DEWNR). In. Adelaide: University of Adelaide.
Vos P, Hogers R, Bleeker M, Reijans M, Lee Tvd, Hornes M, Friters A, Pot J, Paleman J, Kuiper M. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic acids research 23:4407-4414.
Weber LC, VanDerWal J, Schmidt S, McDonald WJ, Shoo LP. 2014. Patterns of rain forest plant endemism in subtropical Australia relate to stable mesic refugia and species dispersal limitations. Journal of Biogeography 41:222-238.
West JG. 1984. A revision of *Dodonaea* Miller (Sapindaceae) in Australia. Brunonia 7:1-194.
West JG. 1980. A taxonomic revision of *Dodonaea* (Sapindaceae) in Australia. [Ph.D. thesis]: University of Adelaide.

Wheeler M, Byrne M. 2006. Congruence between phylogeographic patterns in cpDNA variation in Eucalyptus marginata (Myrtaceae) and geomorphology of the Darling Plateau, south-west of Western Australia. Australian Journal of Botany 54:17-26.
White F. 1962. Geographic variation and speciation in Africa with particular reference to Diospyros. Syst. Assoc. Publ 4:71-103.
Yu Y, Harris AJ, Blair C, He X. 2015. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. Molecular phylogenetics and evolution 87:46-49.

Chapter 4 - Transcriptome sequencing, annotation and polymorphism detection in the hop bush, Dodonaea viscosa

"But a plant on the edge of a deserts is said to struggle for life against the drought, though more properly it should be said to be dependent upon the moisture."

Darwin, 1859

Statement of Authorship

Title of Paper	Transcriptome sequencing, annotation and polymorphism detection in the hop bush, Dodonaea viscosa
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Published in BMC Genomics, 2015, 16:803, DOI 10.1186/s12864-015-1987-1

Principal Author

Name of Principal Author (Candidate)	Matthew J Christmas
Contribution to the Paper	Designed the study; carried out field collections of samples; performed all lab work; analysed sequencing data; wrote manuscript as principal author
Overall percentage (%)	85%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 9.12.15

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Ed Biffin
Contribution to the Paper	Assisted with study design; assisted with field collections of samples; assisted with lab work; and data analysis; advised on and edited manuscript. 10%
Signature	Date 9.12.15.

Name of Co-Author	Andrew J Lowe
Contribution to the Paper	As principal supervisor, obtained funding for the research; advised on study design; advised on development of the manuscript. 5%
Signature	Date 8.12.15

Christmas et al. BMC Genomics (2015) 16:803
DOI 10.1186/s12864-015-1987-1

BMC Genomics

RESEARCH ARTICLE Open Access

Transcriptome sequencing, annotation and polymorphism detection in the hop bush, *Dodonaea viscosa*

Matthew J. Christmas, Ed Biffin and Andrew J. Lowe*

Abstract

Background: The hop bush, *Dodonaea viscosa*, is a trans-oceanic species distributed oversix continents. It evolved in Australia where it is found over a wide range of habitat types and is an ecologically important species. Limited genomic resources are currently available for this species, thus our understanding of its evolutionary history and ecological adaptation is restricted. Here, we present a comprehensive transcriptome dataset for future genomic studies into this species.

Methods: We performed Illumina sequencing of cDNA prepared from leaf tissue collected from seven populations of *D. viscosa* ssp. *angustissima* and *spatulata* distributed along an environmental gradient in South Australia. Sequenced reads were assembled to provide a transcriptome resource. Contiguous sequences (contigs) were annotated using BLAST searches against the NCBI non-redundant database and gene ontology definitions were assigned. Single nucleotide polymorphisms were detected for the establishment of a genetic marker set. A comparison between the two subspecies was also carried out.

Results: Illumina sequencing returned 268,672,818 sequence reads, which were *de novo*assembled into 105,125 contigs. Contigs with significant BLAST alignments (E value < $1e^{-5}$)numbered at 44,191, with 38,311 of these having their most significant hits to sequences from land plant species. Gene Ontology terms were assigned to 28,440 contigs and KEGG analysis identified 146 pathways that the gene products from 5,070 contigs are potentially involved in. The subspecies comparison identified 8,494 fixed SNP differences across 3,979 contiguous sequences, indicating a level of genetic differentiation between them. Across all samples, 248,235 SNPs were detected.

Conclusions: We have established a significant genomic data resource for *D. viscosa*,providing a comprehensive transcriptomic reference. Genetic differences among morphologically distinct subspecies were found. A wide range of putative gene regions were identified along with a large set of variable SNP markers, providing a basis for studies into the evolution and ecological adaptation of *D. viscosa*.

Keywords: *Dodonaea viscosa*, *de novo* assembly, RNA-seq, SNP, Gene ontology

Background

Dodonaea viscosa (L.) Jacq. (hop bush) is a dioecious woody shrub with a worldwide distribution across six continents, spanning from 33°N (in California) to 44°S (in New Zealand's South Island). The species evolved on mainland Australia [1] and its wind-dispersed seeds are thought to be capable of traversing long distances over oceans due to their high physical dormancy [2]. In a flotation experiment, West showed that 30 % of seeds were still afloat after 100 days, and high germination rates were still found after immersion in seawater for 6 months [3]. Hop bush is very widely distributed across a broad range of ecosystems and exhibits high levels of phenotypic variation. As a result, *D. viscosa* has been split into seven subspecies, as characterised by JG West [4], based mainly on leaf characteristics and capsule morphology. Three of the subspecies, *viscosa, burmanniana,* and *angustifolia,* have extra-Australian distributions, whereas ssp. *angustissima, cuneata, mucronata,* and *spatulata* are all found only within Australia.

* Correspondence: andrew.lowe@adelaide.edu.au
Environment Institute and School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide 5005SA, Australia

In Australia, *D. viscosa* is a common species found in a wide range of habitat types, from temperate woodlands to desert gullies and arid shrublands. It has high ecological significance and is widely used in revegetation projects. The high level of phenotypic variation in this species along with its ecological amplitude makes it an ideal species for investigating species divergence and adaptation to local conditions. Within South Australia, *D. viscosa* spans a steep rainfall and temperature gradient within the Adelaide Geosyncline, with ssp. *angustissma* and ssp. *spatulata* restricted mostly to the north and south of this region respectively. Average annual rainfall varies from 200 mm at the northern extreme of the Flinders Ranges to 700 mm on Kangaroo Island in the south, and mean maximum temperature of the warmest month ranges from 23 °C in the south to 36 °C in the north. Within this region it has been demonstrated that *D. viscosa* ssp. *angustissima* exhibits a cline in leaf width, with narrower leaves at lower latitudes [5, 6]. Through the use of herbarium specimens dating back 127 years, GR Guerin, H Wen and AJ Lowe [6] also showed that average leaf width had decreased over that time period. Narrower leaves are an adaptation to hotter, drier climates with broader leaves being more susceptible to extremes in temperature [7]. As a result, GR Guerin, H Wen and AJ Lowe [6] suggest that this shift in leaf width over space is an adaptation to climate, and temporal shifts in leaf width are a response to historical climate variation. However, with only morphological data from field and herbarium specimens used to demonstrate this correlation, whether the observed patterns are a result of a plastic or a genetic adaptation response to climate is unknown. The genomic data presented here will act as a starting point to addressing these questions.

Despite the rapid increase in sequencing capabilities over the last decade, there is currently a lack of genomic data available for this genus. A search on the NCBI nucleotide database with the search term "*Dodonaea*" returned 182 results for the genus, with 146 for *D. viscosa* specifically (March, 2015). The vast majority of these data are barcode markers (e.g. nuclear, ribosomal ITS, plastid matK and rbcL) for phylogenetic studies. In order to identify and investigate potential genes underlying the phenotypic clines demonstrated in this species, genomic resources must first be developed.

For non-model species, RNA sequencing (RNA-seq) is now regularly utilised as an effective method for generating a reduced representation of a species' genome, specifically targeting the transcribed portion of the genome [8–17]. An advantage of transcriptome data, particularly when looking to address questions of adaptation, is that the majority of the transcriptome sequences generated will be from coding regions and therefore of potential functional importance. A major hurdle with the use of this type of data for non-model organisms is that, in the absence of a reference genome, transcript sequences must be assembled *de novo*. With billions of short reads to work with this is no mean feat, and requires a large amount of computational power along with robust, reliable algorithms. As a result, a suite of *de novo* assemblers have now been developed for this purpose [18]. However, depending on the assembler used, results can vary in terms of number and length of contigs [11, 19]. In comparison to assembly against a reference genome, *de novo* assemblies require higher coverage in order to accurately assemble contigs and, as there is no reference, sequencing errors and the presence of chimeric molecules can have a much greater impact [20, 21]. Bearing these issues in mind, being stringent on base quality thresholds and only using contigs with highly significant BLAST hits (an e-value of $\leq 1e^{-4}$ is commonly used throughout the literature [22]) to previously published, putatively homologous sequences (e.g. those found in NCBI non-redundant and Swiss-Prot databases) will help to ensure high confidence in the resultant assembly. In addition, an approach that makes use of more than one assembler and then compares the resultant assemblies by looking for shared and therefore potentially more robust contig calls derives an assembly of high confidence in the absence of a reference genome [23].

In this study, we characterise the leaf transcriptome of *Dodonaea viscosa* with the aim to identify and functionally annotate a large number of expressed genes as well as identify single nucleotide polymorphisms between populations collected along a latitudinal gradient for development as molecular markers. The outcomes from this study will be used to set the stage for future studies into the population and adaptation genomics of this species, with the developed marker set being utilised to further our understanding of adaptive variation along an environmental gradient. Our study presents a valuable resource for on-going research into this ecologically variable and significant species.

Results and discussion

Sequencing and assembly

Illumina Hiseq sequencing of seven cDNA libraries prepared from leaf mRNA generated 268,672,818 sequencing reads, totalling 26.86 Gb. 147,494,172 reads were from ssp. *angustissima* and 121,178,646 were from ssp. *spatulata*. Following quality control steps of removal of duplicates and trimming of sequences on length and quality, a total of 227,376,588 reads remained. 72.3 % of the reads were assembled using CLC Genomics Workbench v. 6 (CLC; http://www.clcbio.com/products/clc-genomics-workbench/) into 105,125 unique contiguous sequences (contigs) of mean length 615 bases (N50 = 812) with a total of 65,390,455 bp. The smallest and

largest contigs were 201 and 15,009 nucleotides respectively. 133,969,990 of the reads included in the assembly remained in pairs, whereas 27,204,111 were broken pairs. All reads were then mapped back to the assembly, with 175,102,401 successfully mapping and 52,274,187 not mapping.

Of the 227,376,588 cleaned reads, only 72.3 % were incorporated into contigs in the *de novo* assembly. The remaining 27.7 % were therefore not included in any further analysis as, at only 100 bp long, they were too short to be considered on their own. There are a number of possible reasons as to why these sequences were not incorporated into contigs. For example, the presence of short microRNA, degradation of RNA during the extraction process, sequencing errors, low sequence coverage, contamination from other organisms, the assembly algorithm used (see below), and low level expression of certain transcripts could all lead to sequences being omitted during the assembly [18, 24].

An expanding number of programs exist for the *de novo* assembly of short-read transcriptome data [18]. Attempting to assemble a transcriptome from short sequences without a reference genome is not a simple task computationally and, as a result, none of these assemblers claim to be perfect. Some of the issues faced are that high levels of coverage are required (over 30×) thus excluding transcripts with low expression [21] and the assembly process is very sensitive to sequencing errors and chimeric molecules [20]. Comparisons of assembler performance have shown that assemblies will vary depending on the assembler used in terms of number of contigs generated, length of contigs, and resultant BLAST success [11, 19]. We elected to use the CLC *de novo* assembler as it has performed well in a number of previous Illumina-based transcriptome assemblies [25, 26] and is the assembler of choice in [22].

In an attempt to validate the contigs produced using CLC the reads were also assembled through the Trinity pipeline [27]. This generated 208,604 contigs. Following removal of duplicates 185,384 contigs remained. Trinity is very effective at identifying splice variants, however this results in a high proportion of redundancy in the data compared to the CLC assembly [23]. It has been demonstrated that CLC is one of the least redundant assemblers [19]. In order to reduce this redundancy in the Trinity assembly the longest contig per component or putative gene was selected. This gave a total of 94,758 contigs, which is comparable in number to the CLC total of 105,125. Reciprocal mappings of the two sets of contigs showed there to be 79.25 % overlap between them. This demonstrates that the two independent assembly algorithms can give highly similar outputs, thus providing more confidence in the resultant set of contigs. All CLC contigs were used in the downstream analysis, rather than just those in common with the Trinity assembly, so as not to lose any potentially useful information.

Sequence annotation

Of the 105,125 CLC-generated contigs, 44,191 (42 %) had significant alignments ($\leq 1e^{-5}$) to sequences in the NCBI's non-redundant database. As has been the case in previous transcriptome assemblies, contig length was a significant predictor of the presence or absence of a significant BLAST hit (logistic regression, slope = 0.104, intercept = −4.037, $P < 2e^{-16}$) [14, 26]. Mean length of contigs with significant BLAST hits was 888 bp (max = 15,009, min = 201, SD = 877) and for those without significant BLAST hits was 421 bp (max = 5330, min = 201, SD = 262). 58 % of the contigs had no homologous sequences within the nr database. This could be for a number of reasons, such as the presence of untranslated mRNA, chimeric sequences resulting from assembly errors, sequences from uncharacterised genes, and sequences from genes unique to *D. viscosa*.

Thirty-eight thousand three hundred eleven (86.7 %) of contigs with significant BLAST hits had their top hit to a species within the Viridiplantae (Fig. 1). The remaining 13.3 % contigs had top BLAST hits to fungi (10.3 %), arthropoda (2.8 %), bacteria (0.09 %), or viruses (0.01 %). As these were field-collected samples with high risk of contamination from endophytes, parasites, and symbionts it was not surprising that a proportion of contig sequences had significant BLAST hits to non-plant sequences (13.3 % of contigs with significant BLAST hits). Taking the contigs with significant blast hits to fungal species, it was interesting to note that the fungal species represented in the data were very similar across all samples. *Ceriporiopsis subvermispora* and the sac fungus *Baudoinia compniacensis* were the most prevalent in all samples, with 13.7 and 13.9 % of the contigs identified as fungal blasting to homologous sequences with these species respectively. The microbial communities associated with *Dodonaea viscosa* are largely unknown and these data could provide a starting point for future research into this.

All contigs without a significant blast hit to plant sequences were excluded from any further analysis, leaving 38,311 contigs remaining. Within these 38,311 contigs, 30,027 unique protein accessions were identified from the BLAST results. The most closely related species to *D. viscosa* with sequenced genomes are *Citrus clementina* and *Citrus sinensis*, also members of Sapindales. The *Citrus sinensis* genome contains an estimated 29,445 protein-coding genes [28], suggesting that our transcriptome is a good representation of the genes present within *D.viscosa*. The most frequent BLAST hit species was *Vitis vinifera* with 108,515 hits. *Theobroma*

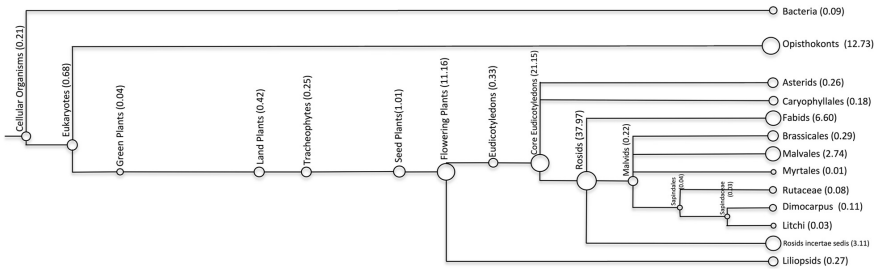

Fig. 1 Most recent common ancestors of assembled contigs. Output from MEGAN, where the output from BLASTX searches of assembled contigs against the NCBI nr database was used to assign contigs to a taxonomic node. The size of each circle reflects the number of contigs assigned to each node, with the percentage of assigned contigs indicated in parentheses

cacao had the second most frequent hits, with 60,718. In terms of the top BLAST hits with lowest e-values, sequences most closely matched those of *Theobroma cacao* (10,050 hits), with *Vitis vinifera* having the second highest number of top hits (6490) (Fig. 2). The higher similarity to *T. cacao* sequences is a reflection of the phylogenetic relatedness of these species to *D. viscosa* [29].

There is currently a lack of genomic information available for *Dodonaea*, as demonstrated by the lack of hits to *Dodonaea* sequences in the BLAST results. The majority of BLAST hits were to agriculturally relevant species, such as *Vitis vinifera*, *Theobroma cacao*, and *Glycine max,* that have had a plethora of genomic data generated for them and so hits to these species are much more likely, as found in other plant transcriptome characterisation studies [10, 15, 30].

This study acts as the starting point of exploration into adaptation across environmental gradients in *Dodonaea viscosa* and, as such, we are interested in identifying genes that may be under differential selection among contrasting climatic regimes. In terms of seeking out genes under selection over a latitudinal gradient where rainfall and temperature clines are steep, genes related to water balance and response to drought are obvious candidates. One such group of genes is the aquaporins, which are involved in membrane permeability, are ubiquitous amongst living organisms, and have been the subject of a number of in-depth functional studies in

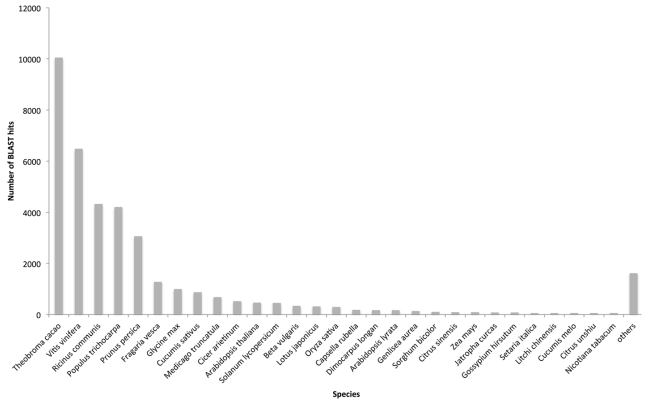

Fig. 2 Top BLAST hit species. Species from the NCBI nr database with the top BLAST hits (lowest e-value) to the assembled contigs. Species are listed on the x-axis in rank order according to their total number of top BLAST hits (y-axis)

plants [31, 32]. Within the transcriptome assembly presented here, 16 contigs had significant homologies with aquaporin or probable aquaporin genes (Table 1), potential targets for future genetic adaptation studies. Abscisic acid (ABA) has also been shown to be involved in water stress responses [33] and ABA signal transduction is interconnected with aquaporin function [34]. Genes involved in ABA production and function are therefore also good candidate genes for investigating adaptation to water-related stress. 29 such genes had homologous sequences within our assembled transcriptome, five of which had "response to water deprivation" gene ontologies assigned (Additional file 1).

Gene ontology (GO) annotation was performed on the 38,311 contigs using the CLC Blast2GOpro plugin. Of these contigs, 28,440 (74 %) were annotated, with a total of 85,444 GO terms assigned (Fig. 3). The number of GOs assigned per sequence ranged from 1 to 66. GO terms fit under three broad categories: biological processes (BP), cellular components (CC), and molecular function (MF). The number of assignments per category was as follows: Cellular components: 33,789 GOs; Biological processes: 23,615 GOs; Molecular function: 28,040 GOs (Fig. 4). It is interesting to note that, under the biological processes category, 8236 contigs were assigned GO terms relating to 'response to stimulus'

Table 1 Contigs with significant Blast hits to aquaporin genes

Contig number	Description	Minimum e-Value	Gene ontologies
11,322	Aquaporin nip1-2	$5.78e^{-103}$	P[1]:arsenite transport; P:response to hypoxia; C[2]:integral to membrane; F[3]:arsenite transmembrane transporter activity; F:lactate transmembrane transporter activity; P:hydrogen peroxide transmembrane transport; P:lactate transmembrane transport; P:response to arsenic-containing substance; C:endoplasmic reticulum; C:plasma membrane
17,499	Probable aquaporin sip2-1-like	$1.17e^{-128}$	C:integral to membrane; C:endoplasmic reticulum; P:transport; C:plasma membrane; F:transporter activity; P:response to arsenic-containing substance
21,612	Aquaporin nip2-1-like	$7.68e^{-45}$	P:silicate transport; C:Casparian strip; C:integral to membrane; F:silicate transmembrane transporter activity
30,905	Aquaporin nip2-1-like	$9.79e^{-109}$	C:Casparian strip; C:integral to membrane; F:silicate transmembrane transporter activity; P:silicate transport
37,229	Aquaporin nip6-1	$3.60e^{-28}$	F:borate transmembrane transporter activity; F:urea transmembrane transporter activity; F:glycerol transmembrane transporter activity; P:myo-inositol hexakisphosphate biosynthetic process; C:integral to membrane; P:borate transmembrane transport; F:water channel activity; P:urea transmembrane transport; P:glycerol transport; P:cellular response to boron-containing substance levels; C:plasma membrane
50,450	Aquaporin nip2-1-like	$1.03e^{-28}$	C:Casparian strip; C:integral to membrane; F:silicate transmembrane transporter activity; P:silicate transport
6446	Aquaporin tip4-1-like	$5.03e^{-131}$	P:water transport; C:plant-type vacuole membrane; C:central vacuole; P:transmembrane transport; C:integral to membrane; F:water channel activity; P:cytokinin mediated signaling pathway
59,060	Aquaporin	$7.44e^{-45}$	P:transport; C:integral to membrane; F:transporter activity
60,372	Aquaporin pip	$3.12e^{-32}$	P:transport; C:integral to membrane; F:transporter activity
64,509	Probable aquaporin nip7-1-like	$2.16e^{-20}$	P:borate transmembrane transport; F:borate transmembrane transporter activity; P:purine nucleobase transport; C:integral to membrane; F:water channel activity
68,886	Aquaporin pip1 1	$1.11e^{-47}$	P:response to water deprivation; C:chloroplast envelope; C:vacuole; C:anchored to plasma membrane; P:water transport; C:integral to membrane; P:response to salt stress; C:mitochondrion; F:transporter activity
72,268	Probable aquaporin pip1-2-like	$1.34e^{-35}$	P:brassinosteroid biosynthetic process; P:response to water deprivation; C:mitochondrion; P:response to salt stress; C:integral to membrane; P:acetyl-CoA metabolic process; F:water channel activity; C:chloroplast envelope; P:cellular response to iron ion starvation; P:iron ion transport; F:protein binding; P:water transport; C:anchored to plasma membrane; C:vacuole; P:sterol biosynthetic process
74,405	Aquaporin pip1 1	$3.98e^{-38}$	P:transport; C:integral to membrane; F:transporter activity
90,132	Aquaporin nip1-2-like	$8.07e^{-28}$	C:membrane
97,487	Probable aquaporin nip-type-like	$3.39e^{-60}$	C:membrane; P:transport; F:transporter activity
1354	Aquaporin	$2.55e^{-134}$	P:response to water deprivation; C:chloroplast envelope; C:vacuole; C:anchored to plasma membrane; P:water transport; C:integral to membrane; P:response to salt stress; C:mitochondrion; F:transporter activity

[1]*P* biological process, [2]*C* cellular component, [3]*F* molecular function

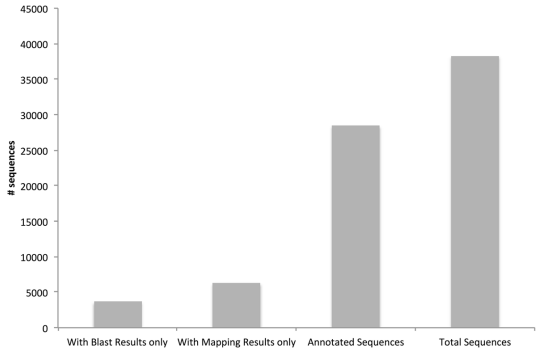

Fig. 3 Gene Ontology (GO) annotation outcome. Blast2GO was used to attempt to annotate the 38,311 contigs that had significant BLAST hits ($\leq 1e^{-5}$) to plant species. The number of contigs that had no GOs assigned (with Blast results only), mapping results only (GO terms retrieved but reliable functions not selected), and that were fully annotated are shown

(Fig. 4b). Of these, a number of GO terms assigned relate to a response to an environmental stressor. For example, response to salt stress (assigned 689 times), response to cold (417), defence response (414), response to water deprivation (308), response to oxidative stress (291), and response to high light intensity (255) were all abundantly assigned. Polymorphisms in and/or differential expression of these genes between populations along an environmental gradient could be an indication of adaptation to local conditions and so should inform future studies into such adaptation.

Nine hundred two enzyme codes were assigned to 5070 contig sequences, which were included in 146 different KEGG pathways (Additional file 2). The most highly represented pathways were "Purine metabolism", "Starch and sucrose metabolism" and "Phenylalanine metabolism" with 561, 548, and 316 assigned contigs respectively (Table 2). This mirrors the findings of a recent annotation of the Aleppo pine (*Pinus halepensis* Mill.) transcriptome, where 419 of their assembled contigs were assigned to the "Purine metabolism" pathway and 399 contigs were assigned to the "Starch and sucrose metabolism" pathway [26]. Given that our transcripts were from leaf tissue, it is not surprising to find that so many are involved in the metabolism of compounds such as amino acids, starch, and sugars. "Phenylpropanoid biosynthesis" was also a well-represented pathway, with 290 sequences assigned to it. Phenylpropanoids, a diverse family of secondary metabolites synthesised from phenylalanine, play vital roles in a wide range of responses to environmental stimuli in plants, such as UV photoprotection, attraction of insect pollinators, and defence against infection and herbivory, as well as being involved in reproduction and the internal regulation of cell physiology and signalling [35]. Again, as leaf tissue requires protection from UV light and is the main site of herbivory, a high prevalence of gene transcripts involved in phenylpropanoid synthesis is as expected.

Of the 38,311 contig sequences, 28,165 had InterPro protein annotations. The most commonly occurring protein region was a pentatricopeptide repeat (PPR), with 13,692 occurrences (Table 3). PPR is a 35-amino-acid motif, first identified in *Arabidopsis thaliana*, that occurs in tandem arrays [36]. Proteins containing PPRs are particularly prevalent in the plant kingdom and they are mainly involved in organelle gene expression through RNA binding, editing [37], splicing, and stability (reviewed in [38]). Such a high prevalence of transcripts coding for PPRs in the leaf transcriptome of *D. viscosa* is therefore unsurprising, as a high concentration of chloroplasts would determine the presence of such proteins. Other commonly occurring domains and sites identified in our contigs include the highly conserved protein kinase domain (6014 occurrences), which contains the catalytic function of protein kinases; proteins belonging to the cytochrome P450 family (5079 occurrences), a diverse group of enzymes involved in the oxidation of organic substances; and leucine-rich repeat regions (2952 occurrences), a repeating stretch of 20–29 amino acids that form an α/β horseshoe fold [39] and are involved in the formation of protein-protein interactions [40] (Table 3).

In this study, only leaf tissue was collected for RNA extraction. This was for two reasons: 1) leaves in this species demonstrate a morphological cline [5, 6] and so are a tissue of interest when looking to answer questions

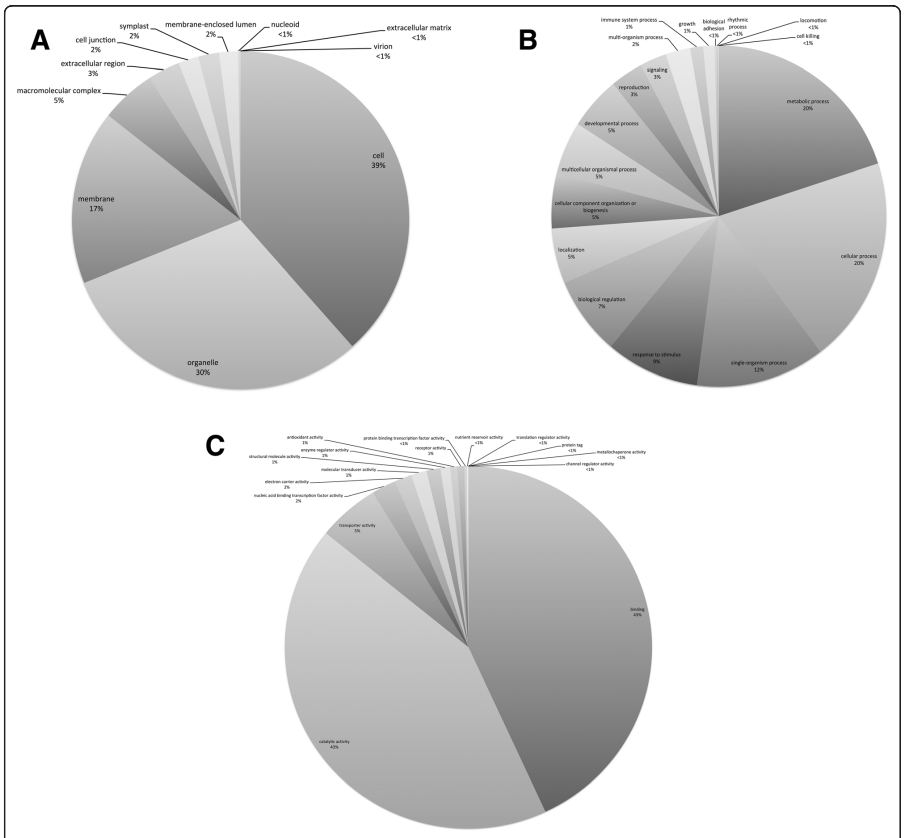

Fig. 4 Distribution of assembled *Dodonaea viscosa* ssp. *angustissima* and ssp. *spatulata* contigs across three main GO categories. Proportions of contigs assigned various GO terms under the cellular component (CC) category (**a**), the biological process (BP) category (**b**), and the molecular function (MF) category (**c**). The number of contigs with annotations in the CC, BP, and MF categories were 33,789, 23,615 and 28,040 respectively

about adaptation, and 2) it simplified the collection and extraction process. Gene activity varies between tissue type as well as varying with the time of day and across seasons. As each tissue type will have a different set of active genes, the transcriptome we have characterised is specific to the leaf tissue in this species. To gain a more holistic view and to include genes that are not transcribed in leaf tissue, other tissue types would also need to be sampled. Despite this, a broad variety of genes with diverse functions is represented in our contigs, as demonstrated by the BLAST results, gene ontology annotation, and KEGG analysis.

Subspecies comparison

The two subspecies within our sample, *D. viscosa* ssp. *angustissma* and ssp. *spatulata*, are distinguishable morphologically mainly by leaf shape [4]. However, this distinction appears not to be absolute and intergrades between forms can be found [4 and personal observation]. Their ranges within the study region overlap slightly, with ssp. *angustissima* restricted mostly to the hotter, drier north and ssp. *spatulata* in the cooler, wetter south, with the exception of sympatric populations on Kangaroo Island (Fig. 5). The degree of genetic differentiation between the subspecies is unknown. Here, we

Table 2 Top 10 KEGG pathways represented by contig sequences

Pathway	Number of sequences
Purine metabolism	561
Starch and sucrose metabolism	548
Phenylalanine metabolism	316
Phenylpropanoid biosynthesis	290
T cell receptor signalling pathway	262
Pyrimidine metabolism	212
Glycolysis/Gluconeogenesis	210
Flavonoid biosynthesis	196
Amino sugar and nucleotide sugar metabolism	189
Glycerolipid metabolism	187

sought to find fixed genomic differences between the transcriptomes of the two subspecies. Mapping the ssp. *spatulata* reads onto a set of consensus ssp. *angustissima* sequences and then looking for polymorphisms between the reference sequences and the mapped reads resulted in the identification of 8494 fixed SNP differences over 3979 shared contigs. The transition/transversion ratio of these SNPs was 1.34, showing a transition bias. This is as expected in the sense that transitions occur more readily due to the molecular mechanisms underlying them and is comparable to the transition/transversion ratio of 1.65 found by [10] in their comparison of big sagebrush (*Artemisia tridentata*) subspecies transcriptomes.

A wide diversity of GO terms were assigned to the subset of shared contigs across the broad categories of cellular components, biological processes and molecular functions (Additional file 3), suggesting that diversification of the subspecies has occurred over a wide range of genes and for a number of traits. Further investigation into fixed differences by, for example, large scale screening for these genes within populations of the two subspecies (as well as other *D.viscosa* subspecies) using targeted gene sequencing technologies such as hybrid capture may provide an insight into the genes involved in adaptation and possibly the mechanisms involved in reproductive isolation leading to speciation.

Variant detection and marker development

All cleaned reads from both subspecies were mapped onto the set of 38,311 contigs with significant blast hits to Viridiplantae gene sequences. Of the 227,376,588 reads, 165,172,256 (72.64 %) mapped to the reference sequences. 248,235 SNPs were identified across 25,270 contigs, with a transition/transversion ratio of 1.38. Stringent quality score settings ensured a high average SNP quality score of 44.65 and average coverage per SNP was 362.04.

Table 3 Top 20 InterPro scan annotations

IPR accession	Description	Type	Occurrence
IPR002885	Pentatricopeptide repeat	Repeat	13,692
IPR000719	Protein kinase domain	Domain	6014
IPR001128	Cytochrome P450	Family	5079
IPR001611	Leucine-rich repeat	Repeat	2952
IPR027417	P-loop containing nucleoside triphosphate hydrolase	Domain	2150
IPR000477	Reverse transcriptase domain	Domain	2084
IPR011009	Protein kinase-like domain	Domain	2058
IPR002290	Serine/threonine- /dual specificity protein kinase, catalytic domain	Domain	1744
IPR012337	Ribonuclease H-like domain	Domain	1652
IPR001680	WD40 repeat	Repeat	1504
IPR008271	Serine/threonine-protein kinase, active site	Active site	1491
IPR001841	Zinc finger, RING-type	Domain	1314
IPR001878	Zinc finger, CCHC-type	Domain	1233
IPR001650	Helicase, C-terminal	Domain	1153
IPR011990	Tetratricopeptide-like helical	Domain	1104
IPR001752	Kinesin, motor domain	Domain	1040
IPR020683	Ankyrin repeat-containing domain	Domain	1036
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic domain	Domain	915
IPR003657	DNA-binding WRKY	Domain	904
IPR001584	Integrase, catalytic core	Domain	894

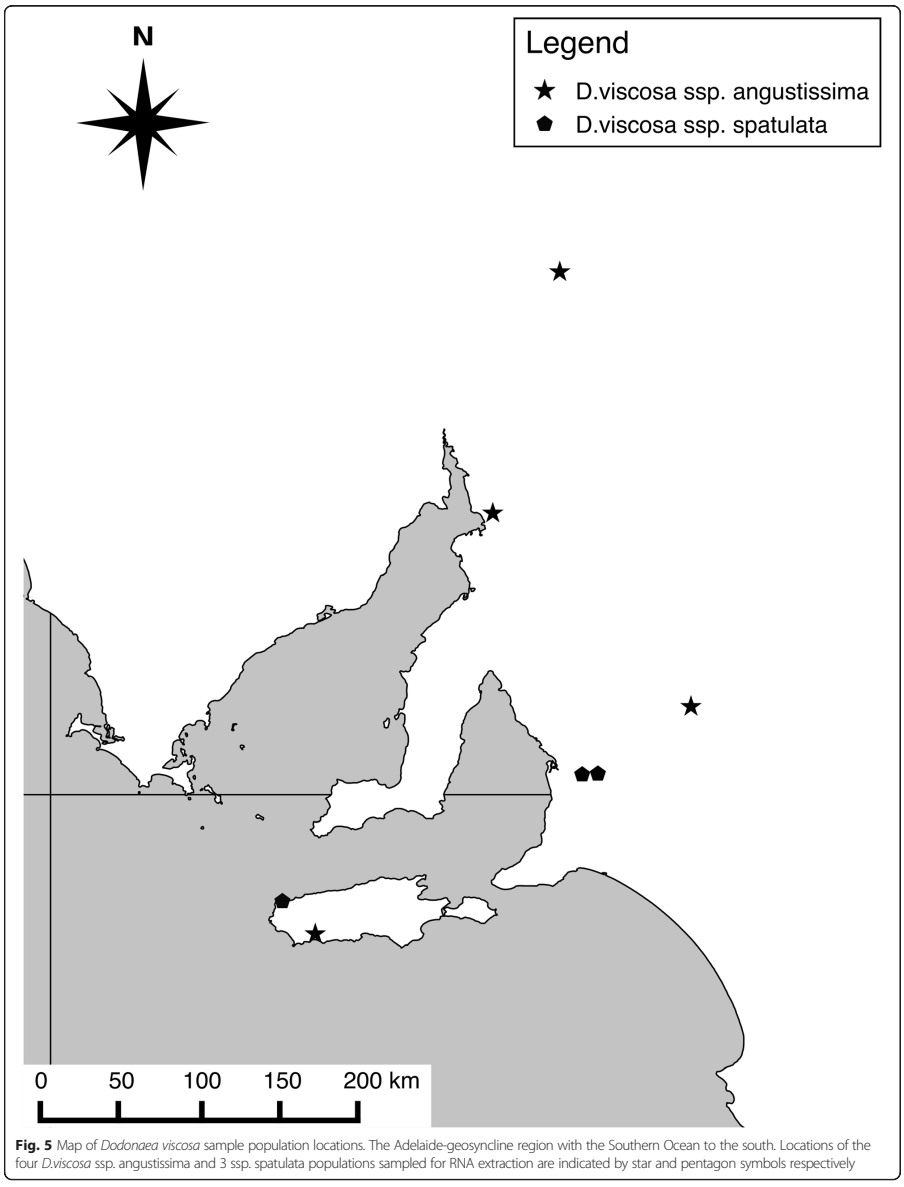

Fig. 5 Map of *Dodonaea viscosa* sample population locations. The Adelaide-geosyncline region with the Southern Ocean to the south. Locations of the four *D.viscosa* ssp. angustissima and 3 ssp. spatulata populations sampled for RNA extraction are indicated by star and pentagon symbols respectively

In the 19,210 contigs that were input into ESTScan, 17,388 coding regions were identified across 16,982 contigs (a number of longer contigs were identified as containing multiple coding regions). After mapping the reads from all populations onto these contigs, 140,246 SNPs were identified across 9400 contigs. Of these, 56,903 SNPs distributed over 2343 contigs were identified as non-synonymous. Non-synonymous SNPs result in amino acid changes in the translated polypeptide and, as such, can be targets of selection. Therefore, this identified set of 2343 contigs may be of particular interest in future studies when looking for signatures of selection.

SNPs are yet to be validated as markers suitable for investigating *D. viscosa* ecology and adaptation and future work is planned to this end. In this study, only two individuals per population were sampled making it difficult to assess allelic variation between populations. However, previous studies have relied on fewer individuals than this when identifying variants from transcriptome data [14, 16]. A more extensive population screening of these potential markers is required, with several individuals per population, in order to measure allele frequencies and fixed differences between populations. A population study using targeted hybrid-capture of the polymorphic loci identified here is in progress in an attempt to uncover signals of selection along an environmental gradient as well as demographic history in this species. Beyond this, the resource could be utilised in linkage mapping and gene-based association studies within *D. viscosa*, as well as for comparative genomics. This transcriptome could also serve as a relevant genetic resource more widely for the currently under-represented *Dodonaea* genus.

Conclusions

Using a single lane of an Illumina HiSeq^{TM} 2000 sequencing run we have generated an extensive genomic resource dataset, providing a broad characterisation of expressed genes in the leaves of *D. viscosa*, as well as identifying a large set of genetic SNP markers for future population genetic analysis of this species. Our results, along with several other recent studies, indicate that short reads from Illumina sequencing can be effectively assembled to provide a characterisation of the transcribed genes within a non-model species [8, 9, 11, 13, 15, 17, 25, 30]. The diversity of genes identified through BLAST searches, gene ontological annotations, KEGG pathways, and protein domain screening demonstrates that we have successfully assembled a robust set of transcriptome sequences representing a wide range of genes of functional significance that are expressed within the leaf tissue of *D. viscosa*. This study provides an extensive genetic resource that can be utilised to delve deep into the ecology and evolution of a species/genus currently lacking in such resources.

Methods

Source populations and leaf material collection

Leaves were field collected from seven populations in South Australia; Three *D. viscosa* ssp. *spatulata* populations and four *D. viscosa* ssp. *angustissima* populations. Sampling sites traversed a latitudinal gradient in South Australia (Fig. 5). Along this gradient average annual rainfall and mean maximum temperature ranges from >700 mm and 19 °C in the south to <200 mm and 25 °C in the north, respectively. Populations varied in size at each site from fewer than 10 individuals to more than 50 individuals. For each population, young leaves were harvested from two individuals, packed into falcon tubes and stored in liquid N_2 until extraction. Sampling was restricted by accessibility as we had to be able to drive as close to the collection sites as possible in order to minimise time between picking the leaves and getting them into the liquid N_2 in order to keep RNA degradation to a minimum. We also selectively picked young, actively growing leaves to increase RNA yield.

RNA extraction

RNA extraction was performed on 2 g of frozen leaf tissue per population sample using a modified CTAB extraction method [41]. The extracted RNA was analysed for yield and quality using Agilent 2200 TapeStation (Agilent Technologies, Santa Clara, CA) prior to sending to the Australian Genome Research Facility (AGRF) for cDNA synthesis and sequencing.

cDNA library preparation and Illumina sequencing

Library preparation and sequencing were performed by AGRF, Melbourne, Australia. Illumina's TruSeq RNA sample preparation protocol was followed for the preparation of mRNA-seq libraries. Briefly, polyadenylated mRNA was isolated from total RNA using oligo dT magnetic beads. mRNA is then fragmented and synthesised using SuperScript II Reverse Transcriptase (Invitrogen). 3′ adenylation of the resulting cDNA fragments then allows for ligation of the sequencing adapters. The resultant libraries were amplified via 12 cycles of PCR. Libraries were assessed using Agilent's Bioanalyser DNA 1000 chip and qPCR was used to quantify the libraries prior to normalising to 2 nM and pooling. Cluster generation for paired end sequencing was performed on an Illumina cBot following the manufacturer's protocol. Paired end sequencing was performed on one lane of an Illimina HiSeq with 208 cycles (101, 6 and 101 cycles for read 1, index read, and read 2 respectively) according to the manufacturer's protocol. Post run processing, including demultiplexing and generation of Fastq files, was performed using Illumina's CASAVA pipeline 1.8.2.

Quality control

Following sequencing, quality control of the sequencing reads was performed in CLC. Duplicate reads were removed using the 'remove duplicates' plugin followed by trimming using the 'trim sequences' tool with setting as follows: removal of low quality sequences: limit = 0.05; removal of ambiguous nucleotides: maximum 2 nucleotides allowed; removal of sequences on length: minimum length 80 nucleotides.

De novo assembly

Following the QC steps, *de novo* assembly of high quality reads was carried out using the *de novo* assembly algorithm within the CLC genomics workbench. This algorithm utilises *de Bruijn* graphs to represent overlapping reads. The following settings were used: automatic bubble size (50), minimum contig length of 201, automatic word size (23), perform scaffolding, and autodetect paired distances. The mapping mode was also used, whereby all reads were mapped back to the assembled contigs and contigs were then updated as a result of the mapping. The mapping settings were as follows: mismatch cost 2, insertion cost 3, deletion cost 3, length fraction 0.9, similarity fraction 0.8. Following *de novo* assembly any duplicate sequences were removed using the 'remove duplicates' CLC plugin.

To validate the CLC assembly, reads were also passed through the Trinity *de novo* assembly pipeline using default parameters [28]. The output contigs were then imported into CLC where duplicates were removed using the 'remove duplicates' plugin. As Trinity generates putative splice variants only the longest contig per component (equivalent to gene) was selected using a custom script. Reads were then mapped back to these contigs in CLC using the same settings as the mapping to CLC contigs. Reciprocal mappings between the two sets of contigs were then performed in CLC in order to compare the two assemblies.

Sequence annotation

BLAST

Assembled contigs were compared to the public NCBI NR protein database using BLASTx in order to identify putative protein homologies. Default parameters were used. An *e*-value cut-off of $\leq 1e^{-5}$ was used in order to restrict results to the most significant matches. As our samples were field collected there was a high chance of the presence of non-plant RNA from e.g. parasites and endophytes. Therefore, MEGAN 4 (http://ab.inf.uni-tuebingen.de/software/megan4/) (Huson et al. [42]) was used to create a list of contigs that were assigned to non-plant species. These could then be excluded from further analyses.

BLAST2GO

The BLASTx output was imported into CLC and gene ontology (GO) terms were assigned using the "mapping" and "annotation" tools in the Blast2GO plugin.

KEGG

GO terms were directly mapped to their enzyme code equivalents in the BLAST2GO Java application (http://www.blast2go.com/b2ghome) in order to generate enzyme code and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway annotations thus identifying which metabolic pathways the gene products are involved in.

InterPro scan

Functional analysis of the translated protein sequences was carried out using InterProScan via the BLAST2GO Java application. InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites (http://www.ebi.ac.uk/interpro/).

Comparison of subspecies

In order to look for fixed differences between the sampled subspecies, a number of steps were taken. The reads from ssp. *angustissima* populations only were mapped against the set of contigs with significant BLAST hits to land plant species in CLC with the following settings: mismatch cost: 2, insertion cost: 3, deletion cost: 3, length fraction: 0.7, similarity fraction: 0.8. Consensus sequences were then extracted from this mapping, using ambiguity codes where alternate alleles were present. These consensus sequences were then used as the reference for the ssp. *spatulata* reads to be mapped to with the same settings as above. This enabled us to specifically call variants between the ssp. *angustissima* consensus sequences and the mapped ssp. *spatulata* reads making the identification of fixed differences between the subspecies more straightforward. Quality-based variant detection in CLC was then run on this mapping in order to identify variants between the two subspecies with the following settings: neighbourhood radius: 10, maximum gap and mismatch count: 5, minimum neighbourhood quality: 20, minimum central quality: 30, ignore non-specific matches, ignore broken pairs.

Single Nucleotide Polymorphism (SNP) discovery

In order to discover SNPs between our samples, all reads that made it through the QC step were mapped on to the contig sequences with significant BLAST hits to members of the *Viridiplantae* clade using the 'map to reference' tool in CLC with the following settings: mismatch cost 2, insertion cost 3, deletion cost 3, length fraction 0.7, similarity fraction 0.8, global

alignment, auto-detect paired distances, ignore non-specific matches. The quality-based variant detection tool in CLC was then employed to identify variants using the following settings: neighbourhood radius: 10, maximum gap and mismatch count: 5, minimum neighbourhood quality: 20, minimum central quality: 30, ignore non-specific matches, ignore broken pairs, minimum coverage: 20, minimum variant frequency: 20 %, maximum expected alleles: 2.

In order to assess whether identified SNPs were synonymous or non-synonymous, ESTScan (http://estscan.sourceforge.net) was used to identify coding regions within the assembled contigs. This program employs a hidden Markov model in order to detect and extract coding regions from sequence data [43]. A subset of 19,210 contigs identified as having significant BLAST hits ($\leq 1e^{-5}$) to land plant species using MEGAN with a low complexity threshold of 0.44 (to ensure low complexity sequences were excluded) was used in this analysis. ESTScan relies upon score matrices specific to the study species. We used the *Arabidopsis thaliana* score matrix provided with the software [10]. Annotations indicating the predicted coding sequence were then added to the contigs using the ESTScan output. Raw reads from all samples were then mapped on to these annotated contigs in CLC with the same settings as in previous mappings. Quality-based variant detection was then run, again with the same settings as in previous variant detection. Resultant SNPs were identified as either synonymous or non-synonymous in the CLC output.

Ethics statement

No ethics approval was required for this study. All relevant permits and approvals were obtained for the field collections carried out in this study. Collection sites were within national parks and conservation parks managed by the South Australian Department of Environment, Water and Natural Resources. No protected species were sampled.

Availability of supporting data

The raw sequencing data sets supporting the results of this article are available in the NCBI SRA repository [accession numbers: SRR1914329, SRR1914332, SRR1914333, SRR1914334, SRR1914335, SRR1914337, SRR1914338, http://www.ncbi.nlm.nih.gov/bioproject/?term=dodonaea%20viscosa].

The assembled transcriptome contigs have been deposited at www.labarchives.com, DOI: 10.6070/H4NS0RW1.

Additional files

Additional file 1: Lists of contigs with significant BLAST hits (e-value < 1e-5) to abscisic acid andaquaporin related genes. (XLSX 49 kb)

Additional file 2: Output from KEGG analysis of contigs, listing identified pathways that gene products areinvolved in. (XLSX 120 kb)

Additional file 3: Gene Ontology terms assigned to the contigs with significant BLAST hits to plant genes,listed under the categories 'biological processes', 'cellular components', and 'molecular functions'. (XLSX 13 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AL and MC planned and coordinated the project. MC and EB carried out the field collections and laboratory work. MC performed the transcriptome assembly, annotation and analyses with assistance from EB. MC drafted the manuscript with input from AL and EB. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank Maurizio Rossetto, Marlien Van Der Merwe and Jeulian Siow at the Royal Botanic Gardens, Sydney for their advice and assistance with bioinformatic issues. Thanks also goes to the two anonymous reviewers who provided constructive feedback on the manuscript. MC, EB, and AL used the Australian Research Council Linkage Grant (LP110100721) to fund the project.

Received: 20 March 2015 Accepted: 6 October 2015
Published online: 16 October 2015

References

1. Harrington MG, Gadek PA. A species well travelled–the *Dodonaea viscosa* (Sapindaceae) complex based on phylogenetic analyses of nuclear ribosomal ITS and ETSf sequences. J Biogeogr. 2009;36(12):2313–23.
2. Baskin JM, Davis BH, Baskin CC, Gleason SM, Cordell S. Physical dormancy in seeds of *Dodonaea viscosa* (Sapindales, Sapindaceae) from Hawaii. Seed Sci Res. 2004;14(1):81–90.
3. West JG. A taxonomic revision of Dodonaea (Sapindaceae) in Australia. Ph.D. thesis. University of Adelaide; 1980.
4. West JG. A revision of *Dodonaea* Miller (Sapindaceae) in Australia. Brunonia. 1984;7:1–194.
5. Guerin GR, Lowe AJ. Leaf morpholgy shift: new data and analysis support climate link. Biol Lett. 2012. doi:10.1098/rsbl.2012.0860.
6. Guerin GR, Wen H, Lowe AJ. Leaf morphology shift linked to climate change. Biol Lett. 2012. doi:10.1098/rsbl.2012.0458.
7. Ackerly DD, Knight CA, Weiss SB, Barton K, Starmer KP. Leaf size, specific leaf area and microhabitat distribution of chaparral woody plants: contrasting patterns in species level and community level analyses. Oecologia. 2002;130(3):449–57.
8. Liu M, Qiao G, Jiang J, Yang H, Xie L, Xie J, et al. Transcriptome sequencing and de novo analysis for ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. PLoS One. 2012;7(10):e46766.
9. Xu D-L, Long H, Liang J-J, Zhang J, Chen X, Li J-L, et al. De novo assembly and characterization of the root transcriptome of *Aegilops variabilis* during an interaction with the cereal cyst nematode. BMC Genomics. 2012;13(1):133.
10. Bajgain P, Richardson B, Price J, Cronn R, Udall J. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). BMC Genomics. 2011;12(1):370.
11. Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M. Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics. 2011;12(1):317.
12. Jung H, Lyons RE, Dinh H, Hurwood DA, McWilliam S, Mather PB. Transcriptomics of a giant freshwater prawn (*Macrobrachium rosenbergii*): de novo assembly, annotation and marker discovery. PLoS One. 2011;6(12):e27938.
13. Lulin H, Xiao Y, Pei S, Wen T, Shangqin H. The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers. PLoS One. 2012;7(6):e38653.

14. Parchman T, Geist K, Grahnen J, Benkman C, Buerkle CA. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics. 2010;11(1):180.
15. Shi Y, Yan X, Zhao P, Yin H, Zhao X, Xiao H, et al. Transcriptomic analysis of a tertiary relict plant, extreme xerophyte *Reaumuria soongorica* to identify genes related to drought adaptation. PLoS One. 2013;8(5):e63993.
16. Sloan DB, Keller SR, Berardi AE, Sanderson BJ, Karpovich JF, Taylor DR. De novo transcriptome assembly and polymorphism detection in the flowering plant *Silene vulgaris* (Caryophyllaceae). Mol Ecol Resour. 2012;12(2):333–43.
17. Zhang L, Yan H-F, Wu W, Yu H, Ge X-J. Comparative transcriptome analysis and marker development of two closely related Primrose species (*Primula poissonii* and *Primula wilsonii*). BMC Genomics. 2013;14(1):329.
18. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671–82.
19. Kumar S, Blaxter ML. Comparing de novo assemblers for 454 transcriptome data. BMC Genomics. 2010;11(1):571.
20. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. Genomics. 2006;88(1):127–31.
21. Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, et al. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics. 2010;11(1):663.
22. Wit P, Pespeni MH, Ladner JT, Barshis DJ, Seneca F, Jaris H, et al. The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. Mol Ecol Resour. 2012;12(6):1058–67.
23. Pallavicini A, Canapa A, Barucca M, Alföldi J, Biscotti MA, Buonocore F, et al. Analysis of the transcriptome of the Indonesian coelacanth *Latimeria menadoensis*. BMC Genomics. 2013;14(1):538.
24. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008;24(3):142–9.
25. De Wit P, Palumbi SR. Transcriptome-wide polymorphisms of red abalone (*Haliotis rufescens*) reveal patterns of gene flow and local adaptation. Mol Ecol. 2013;22(11):2884–97.
26. Pinosio S, González-Martínez S, Bagnoli F, Cattonaro F, Grivet D, Marroni F, et al. First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill. Mol Ecol Resour. 2014;14(4):846–56.
27. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
28. Xu Q, Chen L-L, Ruan X, Chen D, Zhu A, Chen C, et al. The draft genome of sweet orange (*Citrus sinensis*). Nat Genet. 2013;45(1):59–66.
29. Judd WS, Olmstead RG. A survey of tricolpate (eudicot) phylogenetic relationships. Am J Bot. 2004;91(10):1627–44.
30. Li X, Acharya A, Farmer AD, Crow JA, Bharti AK, Kramer RS, et al. Prevalence of single nucleotide polymorphism among 27 diverse alfalfa genotypes as assessed by transcriptome sequencing. BMC Genomics. 2012;13(1):568.
31. Audigeos D, Buonamici A, Belkadi L, Rymer P, Boshier D, Scotti-Saintagne C, et al. Aquaporins in the wild: natural genetic diversity and selective pressure in the PIP gene family in five Neotropical tree species. BMC Evol Biol. 2010;10(1):202.
32. Kaldenhoff R, Ribas-Carbo M, Sans JF, Lovisolo C, Heckwolf M, Uehlein N. Aquaporins and plant water balance. Plant Cell Environ. 2008;31(5):658–66.
33. Steuer B, Stuhlfauth T, Fock HP. The efficiency of water use in water stressed plants is increased due to ABA induced stomatal closure. Photosynth Res. 1988;18(3):327–36.
34. Tyerman SD, Niemietz CM, Bramley H. Plant aquaporins: multifunctional water and solute channels with expanding roles. Plant Cell Environ. 2002;25(2):173–94.
35. Ferrer JL, Austin MB, Stewart Jr C, Noel JP. Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. Plant Physiol Biochem. 2008;46(3):356–70.
36. Small ID, Peeters N. The PPR motif - a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci. 2000;25(2):45–7.
37. Kotera E, Tasaka M, Shikanai T. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature. 2005;433(7023):326–30.
38. Schmitz-Linneweber C, Small I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci. 2008;13(12):663–70.
39. Kobe B, Deisenhofer J. The leucine-rich repeat: a versatile binding motif. Trends Biochem Sci. 1994;19(10):415–21.
40. Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol. 2001;11(6):725–32.
41. Wang T, Zhang N, Du L. Isolation of RNA of high quality and yield from *Ginkgo biloba* leaves. Biotechnol Lett. 2005;27(9):629–33.
42. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. Integrative analysis of environmental sequences using MEGAN4. Genome research. 2011;21:1552-1560.
43. Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. ISMB. 1999;99:138–47.

Chapter 5 - Targeted capture to assess neutral genomic variation in the narrow-leaf hopbush across a continental biodiversity refugium

"Man selects only for his own good: Nature only for that of the being which she tends."

Darwin, 1859

Statement of Authorship

Title of Paper	Determining the level and structure of population genomic variation for the narrow-leaf hopbush across a continental biodiversity refugium - the Adelaide Geosyncline
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	This chapter has been prepared in manuscript style for submission for publication

Principal Author

Name of Principal Author (Candidate)	Matthew Christmas
Contribution to the Paper	Designed the study; carried out field collections of samples; performed all lab work; analysed sequencing data; wrote manuscript as principal author
Overall percentage (%)	85%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 9.12.15

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Ed Biffin
Contribution to the Paper	Assisted with study design; assisted with field collections of samples; assisted with lab work; advised on and edited manuscript. 5%
Signature	Date 9.12.15

Name of Co-Author	Martin Breed
Contribution to the Paper	Advised on study design and analysis and interpretation of the data; advised on and edited manuscript. 5%
Signature	Date 8/12/15

Name of Co-Author	Andrew Lowe
Contribution to the Paper	As principal supervisor, obtained funding for the research; advised on study design; advised on development of the manuscript. 5%
Signature	Date 8.17.15

Targeted capture to assess neutral genomic variation in the narrow-leaf hopbush across a continental biodiversity refugium

Matthew J. Christmas[1], Ed Biffin[2], Martin F. Breed[1] and Andrew J. Lowe[1*]

[1]Environment Institute and School of Biological Sciences, The University of Adelaide, North Terrace, SA 5005, Australia

[2]State Herbarium of South Australia, Hackney Road, Adelaide, SA 5000, Australia

***Author for correspondence**:

Andrew J. Lowe, tel +61 8 8313 1140, fax: +61 8 8303 4364, Email: andrew.lowe@adelaide.edu.au

Key words: applied genomics, *Dodonaea viscosa*, genetic structure, hybrid-capture, restoration

Manuscript prepared for submission to Nature Scientific Reports

Abstract

The Adelaide geosyncline is purported to be an important continental refugium for Mediterranean and semi-arid Australian biota, yet few population genetic studies have been conducted to test this theory. Here, we focus on a plant species distributed widely throughout the region, the narrow-leaf hopbush, *Dodonaea viscosa* ssp. *angustissima*, and examine its genetic diversity and population structure. We used a hybrid-capture target enrichment technique to selectively sequence over 700 genes from 89 individuals across 17 sampling locations. We compared 815 single nucleotide polymorphisms (SNPs) among individuals and populations to investigate population genetic structure. Three distinct genetic clusters were identified; a Flinders/Gammon ranges cluster, an Eastern cluster, and a Kangaroo Island cluster. Higher genetic diversity was identified in the Flinders/Gammon Ranges cluster, confirming that this area is likely to have acted as a refugium during past climate oscillations. We discuss these findings and consider the historical range dynamics of these populations. We discuss the application of our findings to restoration in this species across the region. Finally we provide methodological considerations for population genomics studies that aim to use novel genomic approaches (such as target capture methods) on non-model systems.

Introduction

Climate change impacts are already being realised across the globe, as evidenced by species' responses[1,2]. These impacts are predicted to continue to have widespread effects as conditions become more extreme[3,4]. Persistence of plant populations under climate change will be in large part driven by their ability to overcome constraints to adaptation[5]. For example, large populations with high genetic diversity and wide spread gene flow should be able to maximise adaptive and migrational responses to such pressures. Understanding the distribution of population genetic diversity within species helps to forecast their potential to successfully adapt *in situ* or migrate in response to environmental pressures[6-8].

During past climate oscillations, particularly those experienced during the Pleistocene, refugia have played a major role in the persistence of a vast number of species[9-11]. Refugia are areas that provide species with spatial and/or temporal protection from disturbances[12] and, under climate change, can act as safe havens and shelter species from the harshest conditions. For example, the Adelaide geosyncline region in South Australia (figure 1), the focal region of this study, has been identified as an important historical refugium, where Kangaroo Island and the Flinders Ranges acted as refugia for species to retreat to during colder drier periods[9,13]. Under contemporary climate change, the Adelaide geosyncline also has the potential to serve a similar purpose, offering altitudinal and latitudinal gradients for species to migrate across and avoid climatic extremes. However, the capacity of this area

to be an effective future refugium may be compromised by its highly fragmented state, where habitat loss over the last 200 years has led to little of the historical woodlands and forest remaining. Despite its potential importance, the area remains largely understudied in terms of the population genetic structure and diversity of component species. To our knowledge, only one other published plant population genetic study has focussed on this region[14].

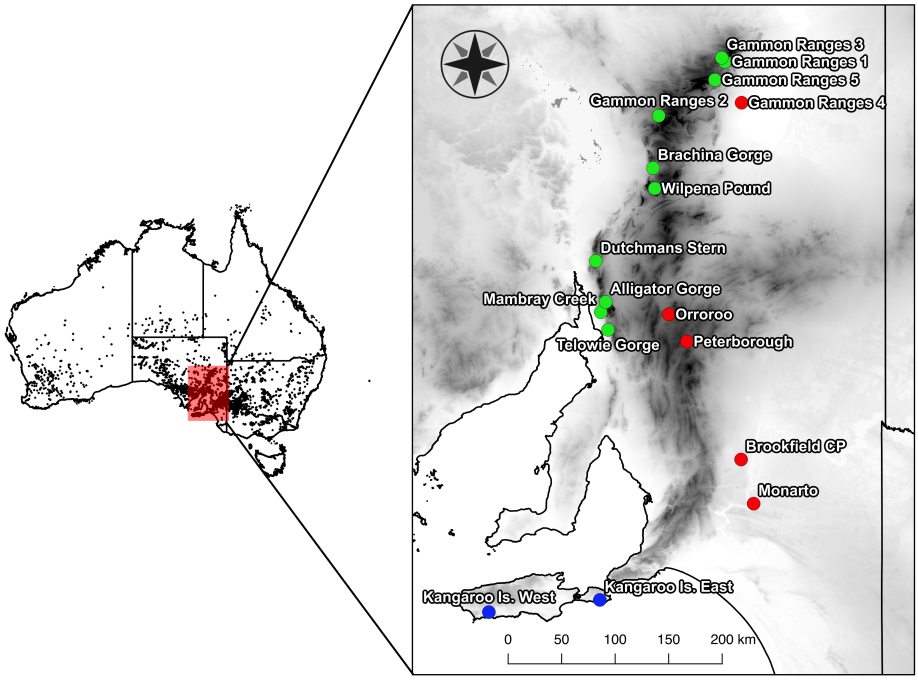

Figure 1. Map of sampling region in South Australia. Population sampling locations of *Dodonaea viscosa* ssp. *angustissima* are indicated by coloured circles, where colours represent genetic cluster assignment from population genetic structure analysis: blue= Kangaroo Island cluster; green = Flinders/Gammon ranges cluster; red = Eastern cluster. Map shading represents elevation with darker shading indicating higher elevation. Black dots on Australian continent map represent all post-1980 *D. viscosa* ssp. *angustissima* sampling locations, downloaded from the Atlas of Living Australia (June, 2016).

Contractions to and expansions from refugia leave genetic signatures across the genome, which contribute to the structuring of genetic diversity in contemporary populations. Populations persisting in past refugia generally maintain higher genetic diversity than the populations that have expanded from them[15,16]. Measures of genetic diversity and structure in contemporary populations therefore allow us to make inferences about past responses to climate change.

In this study, we focused on the narrow-leaf hopbush, *Dodonaea viscosa* ssp. *angustissima* (*D. v. angustissima* hereafter), a widely distributed endemic woody shrub of Australia with a range extending throughout the southern and central regions of the continent. Its hardy nature is reflected in its wide distribution across diverse habitats such as open woodlands, sand plains, and on margins of sand dunes[17]. It is commonly used for restoration, however very little is known about the level and structure of genetic diversity and there is an increasing call for this type of information to be incorporated into restoration planning[18-20]. In particular, measures of population genetic diversity and structure can help ensure the sourcing of high quality and genetically diverse seed in order to maximise adaptive potential of restored populations under climate change[20].

Here, we focus on *D. v. angustissima*'s distribution across the Adelaide geosyncline region (figure 1). This region spans a wide temperature and

rainfall gradient, with cooler, wetter conditions in the south and warmer, drier conditions in the north and east. The Mount Lofty, Flinders, and Gammon Ranges are significant mountain ranges traversing the region. The region has been extensively cleared since European settlement with, for example, less than 10% of the original vegetation cover remaining in the Mount Lofty Ranges[21,22].

With the onset of the 'genomics era', genome-wide data are now easily accessible for non-model species[23]. Genome-wide datasets are superior to more traditional genetic markers (e.g. microsatellites) in estimating the levels and structuring of population genetic diversity[24,25]. For example, the use of hundreds to tens of thousands of single nucleotide polymorphism (SNP) markers distributed throughout the genome means that population genetic studies no longer need as many individual samples per population for accurate allele frequency estimates as was needed when measuring relatively few microsatellite markers[26,27]. As a result, more populations can be included in a study without added expense.

We utilised a novel target capture method and genotyped 89 *D. v. angustissima* samples from 17 populations for single nucleotide polymorphisms (SNPs) to examine population genetic structure and diversity of an ecologically important species across the understudied Adelaide geosyncline. We hypothesised that populations present along the Flinders ranges may show signs of being remnants of a past refugium through elevated levels of genetic diversity, as has been suggested in a previous plant

population genetics study across the region[14], with the topographic nature of the ranges providing ideal refugial habitat.

Results

Sequence data, SNP filtering and outlier analysis

Sequencing of hybrid-capture libraries from all 89 individuals resulted in a total of ~332 million reads, with the number of reads sequenced per individual ranging from 2.3 million to 5 million (mean 3.6 million reads per individual). The percentage of reads that mapped back to the transcriptome reference was 15.74%, which is low but not to be unexpected with the approach taken. By designing capture baits based on a transcriptome reference, alternate splicing and introns, for example, cannot be accounted for, which results in the sequencing of genomic regions that will not map to the transcriptome. Of the reads that mapped, 67.7% mapped in pairs. Following the calling of variants by identifying SNP differences between the reference and mapped sequences, rigorous and stringent filtering steps were taken to provide a reliable set of neutral SNP calls with high coverage across all individuals. Filtering of raw SNPs on depth of coverage, minimum minor allele frequency, and percentage of missing data per SNP resulted in a set of 25,329 SNPs. These SNPs were then pruned of SNPs in LD, reducing the SNP set to 8,462. The requirement of at least 100 bp between each SNP reduced the SNP set further to 2,800 SNPs. We excluded an additional 342 F_{ST} outlier SNPs as they were deemed to be non-neutral. Of the remaining 2,458 SNPs, a further

1,643 SNPs were removed for having negative F_{IS} values as a conservative method of excluding paralogs. This resulted in a final SNP set of 815 SNPs for population genetic diversity and structure analysis.

Population genetic structure

In a discriminant analysis of principal components (DAPC), K = 3 had the lowest Bayesian information criterion (BIC) value, with a clear 'elbow' in the graph at this K value (figure 2). Two discriminant functions were retained, explaining 84.1 % of the variance. Three distinct clusters were identified, one containing all Kangaroo Island samples (KI cluster), one containing samples from within the Flinders and Gammon Ranges (FGR cluster), and one containing all samples to the east of the ranges (Eastern cluster) (figures 1 and 2). The STRUCTURE analysis revealed two to be the most likely value of K (ΔK = 9,835.32) with K=3 the second most likely (ΔK = 1,716.64). When K=2, the same FGR cluster and Eastern cluster as in the DAPC analysis were identified, with the Kangaroo Island samples appearing to be admixed from these clusters (figure 3a). When K = 3, the Kangaroo Island samples formed a third distinct genetic cluster (figure 3b), matching the DAPC result

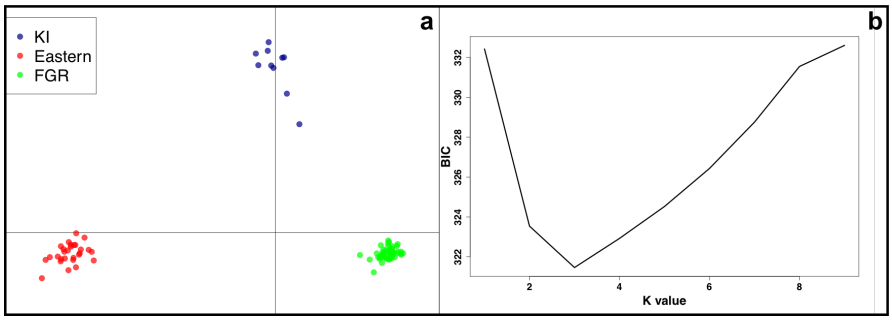

Figure 2. Discriminant analysis of principal components (DAPC) results. (a) Principal component scatter plot of all individuals, based on the DAPC output, and (b) the optimal number of clusters (*K*) as determined by 'k-means', a clustering algorithm which looks for the value of *K* that maximises the variation between groups. The Bayesian Information Criterion (BIC) is plotted for *K* = 1-9 and the 'elbow' in the graph at *K* = 3 indicates this to be the most likely value of *K*. KI = Kangaroo Island cluster, Eastern = Eastern cluster, FGR = Finders/Gammon Ranges cluster.

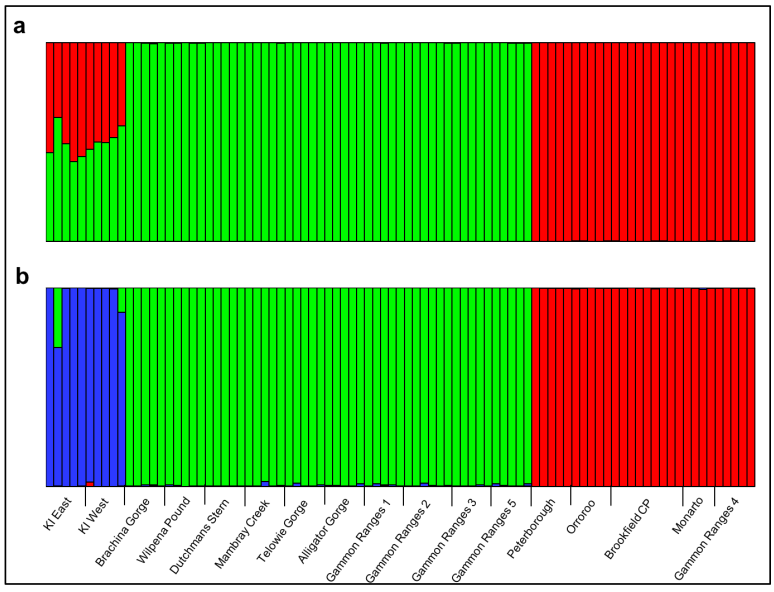

Figure 3. Individual genetic cluster assignments from STRUCTURE results. Results shown are the combined results from ten replicate runs per K value using the admixture model with 200,000 burn in followed by 1,000,000 iterations. (a) *K*=2 (most likely, ΔK = 9,835.32) and (b) *K*=3 (second most likely, ΔK = 1,716.64). Coloured bars represent percentage assignment of individuals to each of the two (a) or three (b) identified clusters. Sampling site locations are listed across the bottom.

Nested AMOVA analysis revealed that the majority of the genetic variance was within individuals (69%, table 1). There was very little variation among individuals within sample populations or among populations within the identified genetic clusters (table 1). Among genetic cluster variance was significant and equal to 16.9% of the total variance (table 1), supporting the clustering identified by the DAPC and STRUCTURE analyses. Average pairwise F_{ST} estimates indicated the greatest differentiation was between the Eastern and KI clusters (F_{ST} = 0.280; table 2), with the least differentiation between the FGR and KI clusters (F_{ST} = 0.138; table 2).

Table 1. Nested analysis of molecular variance (AMOVA) with individuals (n = 89) nested within populations (n = 17), and populations nested within genetic clusters identified from genetic structure analyses (n = 3). The significance of the F statistics was tested using 10,000 permutations in a series of permutation tests.

Source of Variation	Nested in	%var	F-stat	F-value	P
Within Individual	--	69.3	F_{IT}	0.31	--
Among Individual	Population	9.7	F_{IS}	0.12	<0.001
Among Population	Genetic clusters	4.1	F_{SC}	0.05	<0.001
Among genetic clusters	--	16.9	F_{CT}	0.17	<0.001

Table 2. Average pairwise F_{ST} among the three population clusters identified from a discriminant analysis of principal components (DAPC) and STRUCTURE analysis. FGR = Flinders/Gammon Ranges cluster, comprised of Brachina Gorge, Wilpena Pound, Dutchmans Stern, Mambray Creek, Telowie Gorge, Alligator Gorge, and Gammon Ranges 1, 2, 3 and 5 populations (n = 51); Eastern = Eastern cluster comprised of Peterborough, Orroroo, Brookfield Conservation Park, Monarto, and Gammon Ranges 4 populations (n = 28); KI = Kangaroo Island cluster, comprised of Kangaroo Island East and West populations (n = 10).

	KI	Eastern
FGR	0.138	0.189
Eastern	0.280	-

Genetic diversity

Overall observed (H_O) and expected (H_E) heterozygosity were 0.123 (95% CI = ±0.007) and 0.141 (95% CI = ±0.007) respectively, with lowest H_O and H_E in the Peterborough subpopulation (0.066 and 0.076 respectively), greatest H_O in the Telowie Gorge population (0.168) and greatest H_E in the Brachina Gorge population (0.202) (table 3). The FGR cluster had the highest genetic diversity (H_O = 0.157), with the Eastern and KI clusters harbouring similarly lower levels (H_O = 0.077 and 0.071 respectively; table 3).

Table 3. Population genetic summary statistics for the 17 sampling sites showing the number of individuals sampled per population (*n*), and observed (H_O) and expected (H_E) heterozygosity. KI = Kangaroo Island.

Sampling site	***n***	**H_O**	**H_E**
KI East	5	0.073	0.070
KI West	5	0.068	0.078
Peterborough	5	0.066	0.076
Orroroo	5	0.074	0.083
Brachina Gorge	5	0.167	0.202
Wilpena Pound	5	0.129	0.167
Dutchmans Stern	5	0.141	0.163
Mambray Creek	5	0.157	0.181
Telowie Gorge	5	0.168	0.181
Alligator Gorge	5	0.150	0.168
Brookfield CP	8	0.086	0.097
Monarto	5	0.075	0.092
Gammon Ranges 1	5	0.163	0.182
Gammon Ranges 2	6	0.163	0.184
Gammon Ranges 3	5	0.165	0.185
Gammon Ranges 4	5	0.080	0.092
Gammon Ranges 5	5	0.165	0.194
Genetic clusters			
Kangaroo Is.	10	0.071	0.079
Flinders/Gammon	51	0.157	0.189
Eastern	28	0.077	0.092
Overall	89	0.123	0.141

Isolation by distance

Redundancy analysis (RDA) performed on all samples demonstrated that 58% of the total genetic variation was constrained by spatial variables (ANOVA, F = 3.078, P < 0.001; figure 4). By multiplying the percentage of constrained variation (58%) by the overall F_{ST} (0.153) we ascertained that the proportion of the total genetic variation that is explained by the spatial variables is equivalent to an F_{ST} of 0.089. For the Flinders/Gammon Ranges cluster, 15.8% of the total genetic variation was constrained by latitude (ANOVA, F = 1.50, P < 0.01). Overall F_{ST} in this cluster was 0.044, and so the proportion of the total genetic variation explained by the spatial variables is equivalent to an F_{ST} of 0.007. Spatial variables did not explain significant levels of total genetic variation in the Eastern cluster.

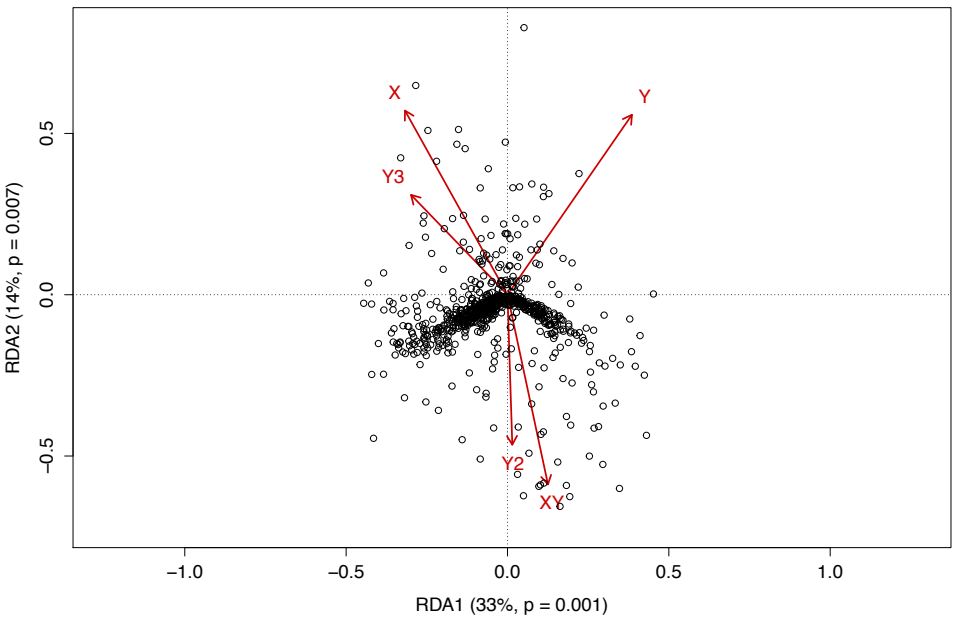

Figure 4. Redundancy analysis (RDA) biplot representing the output of an RDA performed on allele frequency data from 89 *Dodonaea viscosa* ssp. *angustissima* samples from 17 populations. Open circles represent the ordinated allele frequencies (response variable); Red arrows represent spatial polynomials (explanatory variables) plotted as vectors. 58% of the total variation in the genetic data was constrained by the spatial explanatory variables. Of this constrained variation, 33% (p = 0.001) was constrained by axis one (RDA1), and 14% (p = 0.007) by axis two (RDA2). Significance of RDA was assessed using an analysis of variance (ANOVA).

Discussion

Our analysis of neutral SNP variation, distributed across 411 genes, in *D. v. angustissima* detected strong signals of population genetic structure throughout the Adelaide geosyncline region, identifying three distinct clusters. Populations sampled along the Flinders and Gammon Ranges, a significant mountain range in the region, showed distinct genetic signals from populations sampled to the east of the ranges, as well as those from Kangaroo

Island (KI). The Flinders and Gammon Ranges (FGR) cluster also demonstrated higher genetic diversity across the sequenced genes compared to the other two clusters.

Genetic structure and diversity

We used model-based (STRUCTURE) and non-model-based (DAPC) methods to investigate population genetic structure. Both methods assigned the same individuals to three separate clusters: a Kangaroo Island cluster, a FGR cluster, and an Eastern cluster. The DAPC analysis provided strong evidence for KI populations forming a distinct genetic cluster from the FGR and Eastern clusters, and this was also the second most likely scenario in the STRUCTURE analysis. The three clusters make biological sense as the contemporary ranges of mainland populations of *D. v. angustissima* do not extend to coastal regions of the Fleurieu peninsula, the closest part of the mainland to KI. Also, KI has been separated from the mainland since the retreating ice sheets led to sea level rise at the end of the Pleistocene, around 10,000 years ago[28]. It is possible that the KI populations are more closely related to unsampled populations from the Yorke and/or Eyre peninsulas, west of Adelaide. Further sampling would need to be undertaken to test this.

The STRUCTURE analysis provided more support for two distinct clusters, with the KI samples being an admixture of the FGR and Eastern clusters. Admixing of the FGR and Eastern clusters suggests that the KI populations may have resulted from gene flow from the mainland. KI is only 13.5 km offshore, *D. v. angustissima* seed can remain viable in sea water for extended

periods of time[29] and has dispersed out of Australia to as far as South America and Madagascar[30]. Consequently, seed dispersal from the mainland to KI is a real possiblility.

Pairwise F_{ST} estimates between the genetic clusters showed the FGR and KI clusters to be less differentiated from each other than from the Eastern cluster. This suggests greater connectivity between FGR and KI, potentially via seed transported down Spencer Gulf, the body of water to the west of the Southern Flinders Ranges which leads out to the Southern Ocean and Kangaroo Island. The observed differentiation between the FGR and Eastern clusters could be a result of the changes in ecology and environment on and off of the ranges, such as a steep rainfall gradient with rainfall rapidly decreasing to the east of the ranges, resulting in isolation by environment. For example, Gammon Ranges population 4 was less than 35 km from population 5, yet the two populations fall into distinct genetic clusters. Considering the high genetic similarity between all FGR populations, which extend over a much larger distance, gene flow would be expected between these two populations. The two sampling sites differ greatly in their elevation (27 m at GR4 versus 700 m at GR5), their annual mean precipitation (13 mm at GR4 and 24 mm at GR5), and their annual mean aridity index values (0.07 at GR4 and 0.14 at GR5), so gene flow may be unsuccessful despite the short distance, resulting in isolation by environment (precipitation and aridity data obtained from the Atlas of Living Australia, June 2016).

Greater genetic diversity within the FGR cluster compared to the KI and Eastern clusters provides some evidence towards the history of these populations. The presence of high genetic diversity in comparison to surrounding populations is an indicator of past refugia, as refugial populations tend to maintain higher genetic diversity than the populations that have expanded from them[15,16]. Regions of varied topography and high elevation provide ideal refugial conditions as they enable species to remain within their preferred climatic envelopes with only short migration distances[12]. The Southern Flinders Ranges have been suggested to have acted as a genetic refugium for the needle bottlebrush (*Callistemon teretifolius*) during Mid-Pleistocene climate oscillations[14]. Similarly, the results presented here suggest that the Flinders Ranges may have provided a refugium for *D. v. angustissima*.

The low levels of genetic diversity in the KI cluster may be a result of being an island population separated from the mainland and so subject to smaller population sizes and greater genetic drift. However, genotypes for only ten samples within the KI cluster were considered here, in comparison to 51 in the FGR cluster and 28 in the Eastern cluster, and so the lower genetic diversity may also be attributed to sampling error.

Redundancy analysis showed that 58% of the genetic variation across all samples was constrained by space, suggesting isolation by distance. However, as most of the genetic variation was distributed among genetic clusters as well as the fact that the three identified clusters are (mostly) spatially separated, the constrained variation cannot be attributed to isolation by distance. Testing

for the influence of space on within cluster variation found that space constrained only a small percentage of genetic variation in the FGR cluster and none in the Eastern Cluster. This adds to the evidence that most of the genetic variation is distributed among the identified clusters, rather than within.

In terms of *D. v. angustissima*'s continental distribution, the sampling in this study is restricted and so conclusions as to the origins of the populations we have focussed on are difficult to make. Whether the populations in the Eastern cluster are post-Pleistocene expansions from a Flinders/Gammon Ranges refugium or perhaps the margins of a range expansion from a more distant refugium in the east is unresolved without a greater, range-wide sampling effort. Considering the considerable levels of genetic differentiation and the evident lack of admixture between the FGR and Eastern clusters, the second hypothesis appears more likely.

Conservation and restoration implications

The Adelaide Geosyncline has a number of National and Conservation Parks where natural stands of native vegetation are protected. Between these protected areas much land has been cleared, leaving protected areas fragmented across the landscape. Large-scale restoration is carried out across the region to increase the cover of native vegetation, re-connect these fragments, and return functional, native ecosystems. Recent work has focussed on improving success rates of plantings under climate change, due to the questionable success rates of locally sourced material[18-20]. Supplementing local gene pools to increase their adaptive potential should provide restored

populations with the best chances of thriving into the future, whilst avoiding outbreeding depression and maladaptation to local conditions[18,19]. For *D. v. angustissima*, a species commonly used in revegetation projects, the distinct genetic clustering and clear assignment of individuals to these clusters demonstrates that the three populations are genetically isolated from one another, and adaptive differences are likely to be present. As such, movement of seed between these regions may result in maladapted plants and outbreeding depression. Further investigation into the phenotypic differences among plants across these genetic clusters through reciprocal transplant experiments are required to fully assess the risks of mixing seed from across the identified genetic clusters.

Developing genomic resources for non-model species

The target capture method used in the current study[31,32] is yet to be widely utilised in the fields of population and conservation genetics, in comparison to other genome partitioning methods such as Genotyping by Sequencing (GBS) and RADSeq[33]. Here, we chose a more targeted approach as it allowed us to sequence specific genes of interest identified and designed from the assembled transcriptome for the species[34]. This resulted in reliably sequencing over 700 gene regions with putative functions assigned for each individual. The main advantage of this approach was that, for a non-model organism without a reference genome, identified variants could be assigned to the specific gene they occurred in and their functional significance could be ascertained. Although this type of information is not necessarily informative

for population genetic analyses, where the aim is to estimate neutral processes, the development of such a genetic marker dataset provided the neutral markers required for the types of analyses presented here (as most of the variation, even in functional genes, is expected to be neutral) as well as providing a set of markers located within transcribed genes that can be explored for evidence of non-neutral processes such as selection.

In our study, 5-8 individuals were sampled per population. These relatively small numbers were constrained by the fact that, as in most population genetic studies, compromises must be made between the number of populations sampled and the number of samples per population due to budget restrictions. This is a potential issue as estimates of F_{ST} can be biased if sample sizes are too small[35-37]. It has been suggested that power in F_{ST} estimates can more readily be increased by sampling more individuals per population rather than sampling more markers per individual, particularly when F_{ST} is low[37]. However, with the advent of next-generation sequencing, it is now cheaper to increase the number of markers compared to increasing the number of individuals genotyped. In their simulations of the effect of number of individuals on inferential power for different number of SNPs, Morin *et al*[37] demonstrated that a sample size of 10 individuals per population and only 20 SNPs provided complete power to detect differentiation at the level of F_{ST} = 0.2. As few as four samples per population have been shown to be sufficient for F_{ST} estimates when using a large number of markers (>1,000)[26,27].

In our study, average pairwise F_{ST} among populations was 0.16, which is quite low, and so our use of only 5-8 individuals per population may have resulted in low power for our F_{ST} estimates. However, the assignment of individuals to genetic clusters through the genetic structure analyses meant that we were actually working with sample sizes of 51, 28, and 10 for the FGR, Eastern and KI clusters respectively. This, along with our use of a large number of SNPs (815) should have provided sufficient power to reliably detect differentiation among the clusters without having to compromise on the number of sampling sites.

Methods

Study system and sampling

We sampled *D. v. angustissima* throughout the Adelaide geosyncline, with sampling effort stretching from Kangaroo Island in the south, through the Mount Lofty and Flinders Ranges to the Gammon Ranges in the north (figure 1). This sampling design enabled us to collect samples covering multiple environmental gradients, with a strong north-south temperature and rainfall gradient as well as an independent east-west rainfall gradient. Avoiding a single, latitudinal transect for sampling and sampling populations that are geographically close but environmentally dissimilar makes the detection of population genetic structure driven by adaptation (isolation by ecology) as well as by distance possible, as large environmental distances between populations are likely to have resulted in local adaptation[38,39]. *D. v.*

angustissima leaf samples were collected from 89 plants, which included 5-8 plants per site at 17 sites across the region. Leaf samples were stored in teabags on silica gel prior to DNA extraction.

Genome-wide data generation

Capture probe design

The previously published transcriptome for this species[34] was used to design hybrid-capture probes for selectively sequencing hundreds of gene regions reliably across all samples. Previous annotation of the transcriptome via BLAST searches to the NCBI non-redundant database meant that genes and their putative functions had already been identified (details in [34]). This information was used to design a probe set that could generate data on functional regions of the genome to inform on both neutral (the present study) and adaptive (a separate study) genetic variation. Functional information was used to select a set of 353 genes that were assigned gene ontology classifications relating to a response to water stress as well as, more specifically, all genes identified as relating to aquaporin and abscisic acid (ABA) functions. A second set of 617 genes was also selected on the basis of the presence of non-synonymous SNPs in a subspecies comparison in [34]. This resulted in a set of 970 target genes. Hybrid capture probes for the capture of these 970 genes were designed and synthesised by MYcroarray (MI, USA) using their 80-mer MyBaits custom bait library system with 2x tiling.

Although the targeted gene sequences were mainly selected based on *a priori* expectations they may be under selection and so informative for a separate study focussing on signatures of selection, it is expected that a significant proportion of the variation in these targeted genes will be neutral. By identifying the neutral variation in this dataset we were able to use it to address questions of neutral population genetic diversity and structure in the current study.

DNA extraction, hybrid-capture enrichment and sequencing

DNA was extracted using the Machery-Nagel Nucleospin Plant II Kit at the Australian Genome Research Facility (AGRF, Adelaide, Australia). The extracted DNA was then sonicated for random sheering and 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 (www.mycroarray.com/pdf/MYbaits-manual-v2.pdf) using the high stringency wash buffer and 12 cycles of post-capture PCR. Following capture 100 bp paired-end sequencing with dual indexing of 89 samples was performed on one lane of an Illumina HiSeq 2000 at AGRF (Melbourne, Australia). Sequence data was subsequently processed using the Illumina CASAVA pipeline (version 1.8.2).

Sequence quality, SNP discovery and filtering

Sequence quality was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Raw sequence quality was very high, negating the need for any trimming. Mapping

of raw sequence reads to the reference transcriptome from [34] was performed using BWA[40]. The indexed reference was created using default settings. Picard tools (http://broadinstitute.github.io/picard/) were used to compress the resulting SAM files, sort the sequences by reference contig and mark duplicated sequence reads. Mapping characteristics were assessed using SAMtools[41] . Variant calling was performed per individual on the mapped reads using the SAMtools utility "mpileup". Settings used are listed in the supplementary methods. Variants were output as genotype probabilities in one VCF file per individual. Output VCF files were then merged and a custom script was used to convert variant calls from genotype probabilities to genotype calls.

SNPs were subsequently filtered using VCFTools[42] as follows: minimum depth of 10 reads per individual, minor allele frequency >10%, missing data per SNP <25% across all individuals. The mean number of base pairs between SNPs for each contig was also calculated and contigs containing fewer than 10 base pairs per SNP were removed in order to control for mapping errors. We then filtered out SNPs that were likely to be in linkage disequilibrium (LD) using the LD pruning tool in PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/). This ran independent pairwise regressions between all SNPs. A cut-off $r^2 > 0.5$ was used, whereby one of a pair of SNPs was removed from the dataset if the coefficient of determination between the pair was greater than 0.5, thus removing SNPs

showing strong signals of LD. A further requirement of at least 100bp between each SNP was also implemented.

We then removed outlier SNPs using an F_{ST}-based outlier analysis implemented in BayeScan ver. 2.0[43]. BayeScan implements a reversible-jump MCMC algorithm to estimate the posterior probability of models of neutrality and selection. The use of posterior probabilities adjusts for inflated false discovery rates (FDR; the expected proportion of false positives among outlier markers). Q-values (the minimum FDR at which a locus may become significant) are calculated for each locus and used to set an FDR threshold of 0.05 (a 5% false positive rate). Default settings were used, including prior odds = 10. Such low prior odds increase the risk of false positives[44] and therefore will result in a very conservative set of neutral SNPs, which in our case is ideal.

The hybrid capture baits were designed based on transcriptome sequences and, as a reference genome was lacking, the presence of duplicated or paralogous sequences of the bait targets within the *D. v. angustissima* genome was unknown. If present, paralogous sequences may map together during the mapping stage. This could skew allele frequency estimates and bias results. Paired-end sequencing was employed in this study, and the requirement of both members of a pair to be present when mapping to a reference can help to reduce the chance of mapping paralogous regions together. As an extra control, F_{IS} values of the generated SNP set were calculated in GENODIVE and SNPs displaying significantly negative F_{IS} values

(indicating greater than expected heterozygosity under Hardy-Weinburg equilibrium, which may be indicative of paralogous regions mapping together; significance assessed using permutation tests with 10,000 permutations) were removed using VCFTools.

Population genetic analysis

Population genetic structure was explored using a number of methods, as follows. Population genetic clustering analyses were performed using the non-model based method discriminant analysis of principle components (DAPC[45]), and the model-based method STRUCTURE[46]. Firstly, DAPC[45], implemented in `adegenet` in R[47], was used in order to ascertain the number and assignment of individuals to genetic clusters. DAPC is a non-model-based multivariate approach, which seeks discriminating functions between groups of individuals while minimising variation within clusters. Genetic data were first transformed into uncorrelated components using principal component analysis (PCA). The number of genetic clusters was then defined using k-means, a clustering algorithm that looks for the value of k that maximises 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`, in order to efficiently describe the genetic clusters and assign samples to each cluster.

Secondly, the most likely number of clusters and individual assignment to those clusters was assessed using STRUCTURE ver. 2.3.4. An admixture model was used to determine the number of population clusters (K) with a burn-in of 200,000 followed by 1,000,000 iterations. *K* values 1-10 were assessed, with 10 replicates per *K* value. ΔK[48] was calculated for each *K* value in Structure Harvester ver.0.6.94[49] in order to assess the most likely *K*. Results from replicate runs of the most likely *K* were combined using CLUMPP[50] with default settings.

A nested analysis of molecular variance (AMOVA)[51] was performed using GENODIVE, with individuals nested within populations and populations nested within the genetic clusters identified by genetic structure analysis. Fixation indices and the proportion of genetic variation found within individuals (F_{IT}), among individuals nested within populations (F_{IS}), among populations nested within genetic clusters (F_{SC}), and among genetic clusters (F_{CT}) were calculated. Significance of each fixation index was evaluated using permutation tests with 10,000 permutations in order to assess the partitioning of genetic variation among subpopulations as well as among the genetic clusters. Pairwise F_{ST} [52] was calculated between each of the genetic clusters identified by the structure analyses. Genetic diversity was assessed through measures of expected and observed heterozygosity for each sampling site, as well as for the genetic clusters determined by population structure analyses, in GENODIVE ver. 2.0b27[53].

In order to measure the spatial component of the among-population variation a redundancy analysis (RDA) was performed on the population allele frequencies using a modified R script from [39]. Briefly, allele frequencies for one allele per locus were calculated for each population. A matrix of spatial variables was made by calculating orthogonal third-degree polynomials based on population coordinates using the command "poly" in R[39,54]. The command "OrdiStep" in the R package VEGAN was used for forward selection of spatial variables in order to prevent overfitting. RDA was then performed, using the command "rda" (VEGAN), with the allele frequency matrix as dependent and spatial polynomials matrix as independent variables. The output from the RDA was then used to calculate the percentage of the total genetic variation that is explained by the spatial variables by multiplying the proportion of constrained variation with the overall value of F_{ST}[39]. ANOVA was used to assess the significance of the RDA.

In order to account for the co-correlation of geographic distance and identified genetic clusters, the RDA analysis was performed separately on only populations from the FGR cluster, only populations from the Eastern cluster, as well as all samples together. The Kangaroo Island populations were not analysed separately due to the low number of samples and limited geographic variation.

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, the Field Naturalist Society of South Australia and the Australian Wildlife Society Student Grant awarded to MJC. Thanks also to QFAB (Queensland, Australia) for assistance with bioinformatic processing.

Author contributions

MJC, EB, MFB, and AJL designed the research. MJC and EB performed field collections and laboratory work. MJC analysed the data. MJC wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

Additional information

Supplementary information accompanies this paper online.

Competing financial interests: The authors declare there are no conflicts of interest.

Data archiving: Sequence reads are archived at the NCBI SRA with accession number SRP077342. The variants file is available as a supplementary file in variant call format (Neutral_SNPs_File.vcf).

References

1 Franks, S. J., Weber, J. J. & Aitken, S. N. Evolutionary and plastic responses to climate change in terrestrial plant populations. *Evol. Appl.* **7**, 123-139 (2014).
2 Parmesan, C. Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol., Evol. Syst.* **37**, 637-669, doi:Doi 10.1146/Annurev.Ecolsys.37.091305.110100 (2006).
3 Corlett, R. T. & Westcott, D. A. Will plant movements keep up with climate change? *Trends Ecol. Evol.* **28**, 482-488, doi:10.1016/j.tree.2013.04.003 (2013).
4 Jump, A. S. & Peñuelas, J. Running to stand still: adaptation and the response of plants to rapid climate change. *Ecol. Lett.* **8**, 1010-1020, doi:10.1111/j.1461-0248.2005.00796.x (2005).
5 Christmas, M. J., Breed, M. F. & Lowe, A. J. Constraints to and conservation implications for climate change adaptation in plants. *Conserv. Genet.* **17**, 305-320, doi:10.1007/s10592-015-0782-5 (2016).
6 Larcombe, M. J., McKinnon, G. E. & Vaillancourt, R. E. Genetic evidence for the origins of range disjunctions in the Australian dry sclerophyll plant Hardenbergia violacea. *J. Biogeogr.* **38**, 125-136 (2011).
7 Petit, R. J., Hu, F. S. & Dick, C. W. Forests of the past: a window to future changes. *Science* **320**, 1450-1452 (2008).
8 Temunović, M., Frascaria‐Lacoste, N., Franjić, J., Satovic, Z. & Fernández‐Manjarrés, J. F. Identifying refugia from climate change using coupled ecological and genetic data in a transitional Mediterranean‐temperate tree species. *Mol. Ecol.* **22**, 2128-2142 (2013).
9 Byrne, M. Evidence for multiple refugia at different time scales during Pleistocene climatic oscillations in southern Australia inferred from phylogeography. *Quat. Sci. Rev.* **27**, 2576-2585 (2008).
10 Petit, R. J. *et al.* Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *For. Ecol. Manage.* **156**, 49-74 (2002).
11 Stewart, J. R. & Lister, A. M. Cryptic northern refugia and the origins of the modern biota. *Trends Ecol. Evol.* **16**, 608-613 (2001).
12 Keppel, G. *et al.* Refugia: identifying and understanding safe havens for biodiversity under climate change. *Global Ecol. Biogeogr.* **21**, 393-404, doi:10.1111/j.1466-8238.2011.00686.x (2012).
13 Crisp, M. D., Laffan, S., Linder, H. P. & Monro, A. Endemism in the Australian flora. *J. Biogeogr.* **28**, 183-198 (2001).
14 McCallum, K. P., Guerin, G. R., Breed, M. F. & Lowe, A. J. Combining population genetics, species distribution modelling and field assessments to understand a species vulnerability to climate change. *Austral Ecol.* **39**, 17-28, doi:10.1111/aec.12041 (2013).
15 Hewitt, G. Genetic consequences of climatic oscillations in the Quaternary. *Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci.* **359**, 183-195 (2004).

16 Lewis, P. O. & Crawford, D. J. Pleistocene refugium endemics exhibit greater allozymic diversity than widespread congeners in the genus *Polygonella* (Polygonaceae). *Am. J. Bot.*, 141-149 (1995).
17 West, J. G. A revision of *Dodonaea* Miller (Sapindaceae) in Australia. *Brunonia* **7**, 1-194 (1984).
18 Breed, M. F., Stead, M. G., Ottewell, K. M., Gardner, M. G. & Lowe, A. J. Which provenance and where? Seed sourcing strategies for revegetation in a changing environment. *Conserv. Genet.* **14**, 1-10, doi:10.1007/s10592-012-0425-z (2012).
19 Broadhurst, L. M. *et al.* Seed supply for broadscale restoration: maximizing evolutionary potential. *Evol. Appl.* **1**, 587-597 (2008).
20 Prober, S. M. *et al.* Climate-adjusted provenancing: a strategy for climate-resilient ecological restoration. *Front. Ecol. Evol.* **3**, 65 (2015).
21 Bradshaw, C. J. Little left to lose: deforestation and forest degradation in Australia since European colonization. *J. Plant Ecol.* **5**, 109-120 (2012).
22 Paton, D. & O'Connor, J. The state of Australia's birds 2009: restoring woodland habitats for birds. (2010).
23 Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133-141, doi:10.1016/j.tig.2007.12.007 (2008).
24 Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K. & Hedrick, P. W. Conservation genetics in transition to conservation genomics. *Trends Genet.* **26**, 177-187 (2010).
25 Wheeler, N. & Sederoff, R. Role of genomics in the potential restoration of the American chestnut. *Tree Genet. Genom.* **5**, 181-187 (2009).
26 Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489-494 (2009).
27 Willing, E.-M., Dreyer, C. & Van Oosterhout, C. Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLoS One* **7**, e42649 (2012).
28 Hope, J., Lampert, R., Edmondson, E., Smith, M. & Van Tets, G. Late Pleistocene faunal remains from Seton rock shelter, Kangaroo Island, South Australia. *J. Biogeogr.*, 363-385 (1977).
29 Baskin, J. M., Davis, B. H., Baskin, C. C., Gleason, S. M. & Cordell, S. Physical dormancy in seeds of *Dodonaea viscosa* (Sapindales, Sapindaceae) from Hawaii. *Seed Sci. Res.* **14**, 81-90 (2004).
30 Harrington, M. G. & Gadek, P. A. A species well travelled–the *Dodonaea viscosa* (Sapindaceae) complex based on phylogenetic analyses of nuclear ribosomal ITS and ETSf sequences. *J. Biogeogr.* **36**, 2313-2323 (2009).
31 Jones, M. R. & Good, J. M. Targeted capture in evolutionary and ecological genomics. *Mol. Ecol.* **25**, 185-202 (2016).
32 Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111-118, doi:10.1038/nmeth.1419 (2010).
33 Davey, J. W. & Blaxter, M. L. RADSeq: next-generation population genetics. *Brief. Funct. Genomics* **9**, 416-423 (2010).
34 Christmas, M. J., Biffin, E. & Lowe, A. J. Transcriptome sequencing, annotation and polymorphism detection in the hop bush, *Dodonaea viscosa*. *BMC Genomics* **16**, 803 (2015).

35 Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat. Rev. Genet.* **10**, 639-650 (2009).
36 Kalinowski, S. Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity* **94**, 33-36 (2005).
37 Morin, P. A., Martien, K. K. & Taylor, B. L. Assessing statistical power of SNPs for population structure and conservation studies. *Mol. Ecol. Res.* **9**, 66-73 (2009).
38 Hereford, J. A quantitative survey of local adaptation and fitness trade‐offs. *Am. Nat.* **173**, 579-588 (2009).
39 Meirmans, P. G. Seven common mistakes in population genetics and how to avoid them. *Mol. Ecol.* **24**, 3223-3231 (2015).
40 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
41 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
42 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
43 Foll, M. & Gaggiotti, O. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* **180**, 977-993, doi:10.1534/genetics.108.092221 (2008).
44 Lotterhos, K. E. & Whitlock, M. C. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol. Ecol.* **23**, 2178-2192 (2014).
45 Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
46 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).
47 Team, R. C. (ISBN 3-900051-07-0, 2014).
48 Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611-2620 (2005).
49 Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Res.* **4**, 359-361, doi:10.1007/s12686-011-9548-7 (2012).
50 Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806 (2007).
51 Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491 (1992).
52 Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution*, 1358-1370 (1984).

53 Meirmans, P. G. & Van Tienderen, P. H. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol. Ecol. Notes* **4**, 792-794 (2004).
54 Borcard, D., Legendre, P. & Drapeau, P. Partialling out the spatial component of ecological variation. *Ecology* **73**, 1045-1055 (1992).

Supplementary Methods

Flags used in SNP calling using mpileup:

-l selected.snps.list – use a pre-defined list of alleles for genotyping rather than selecting SNPs on basis of deviation from the reference base
-f use a fasta format sequence reference
-b bam.file.list – perform the genotyping on the list of filenames contained within this file
-I = skip INDELs
-C 50 = adjust the mapping quality to 50 – as recommended by samtools authors
-t DP,SP = include (high quality) depth-of-coverage and strand bias information in the VCF output file
-v = output results in VCF format
-A = do not discard anomalous read pairs (the reference collection is fragmented and it is seen that a fraction of ~10% of mapping reads are anomalous in the flagstat section)
-E = recalculate BAQ on the fly – the base mapping quality information is used in calculation of likelihood that a variant is real / noise

Chapter 6 - Finding needles in a genomic haystack: targeted sequencing to identify signatures of selection in a non-model species

"But Natural Selection, as we shall hereafter see, is a power incessantly ready for action, and is immeasurably superior to man's feeble efforts, as the works of Nature are to those of Art."

Darwin, 1859

Statement of Authorship

Title of Paper	Finding needles in a genomic haystack: targeted sequencing to identify signatures of selection in a non-model species
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	This chapter has been prepared in manuscript style for submission for publication

Principal Author

Name of Principal Author (Candidate)	Matthew J Christmas
Contribution to the Paper	Designed the study; carried out field collections of samples; performed all lab work; analysed sequencing data; wrote manuscript as principal author
Overall percentage (%)	85%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 9.12.15

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Ed Biffin
Contribution to the Paper	Assisted with study design; assisted with field collections of samples; assisted with lab work; advised on and edited manuscript. 5%
Signature	Date 9.12.15.

Name of Co-Author	Martin F Breed
Contribution to the Paper	Advised on study design and analysis and interpretation of the data; advised on and edited manuscript. 5%
Signature	Date 8/12/15

Name of Co-Author	Andrew J Lowe
Contribution to the Paper	As principal supervisor, obtained funding for the research; advised on study design; advised on development of the manuscript. 5%
Signature	Date 8.17.15

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

Running title: Targeted capture for identifying selection

Matthew J. Christmas[1], Ed Biffin[2], Martin F. Breed[1] and Andrew J. Lowe[1]

[1]Environment Institute and School of Biological Sciences, The University of Adelaide, North Terrace, SA 5005, Australia; [2]State Herbarium of South Australia, Hackney Road, Adelaide, SA 5000, Australia

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

Author for correspondence: Andrew J. Lowe, tel +61 8 8313 1140, fax: +61 8 8303 4364, Email: andrew.lowe@adelaide.edu.au

Accepted for publication in Molecular Ecology, July 2016

Abstract

Teasing apart neutral and adaptive genomic processes and identifying loci that are targets of selection can be difficult, particularly for non-model 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 non-model 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 8,462 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 non-model organisms.

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 non-model organisms (Davey *et al.* 2011). New tools are regularly being developed in order 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 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 non-functional (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 in order 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 non-model 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 non-functional 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 of 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 non-model plant *Dodonaea viscosa* ssp. *angustissima* (hereafter *D. v. angustissima*) in the absence of a reference genome. *D. v. 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 know 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). In order to increase the chances of uncovering signatures of selection, we chose candidate genes with a putative function 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 to provide evidence towards answering two main questions: (1) to what extent has environmental variation among our target populations driven selection, leading to gene-environment correlations in functional genes? and (2), 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 5-8 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 in order to minimise the chances of sampling close relatives. Leaf samples were stored in teabags on silica gel prior to DNA extraction.

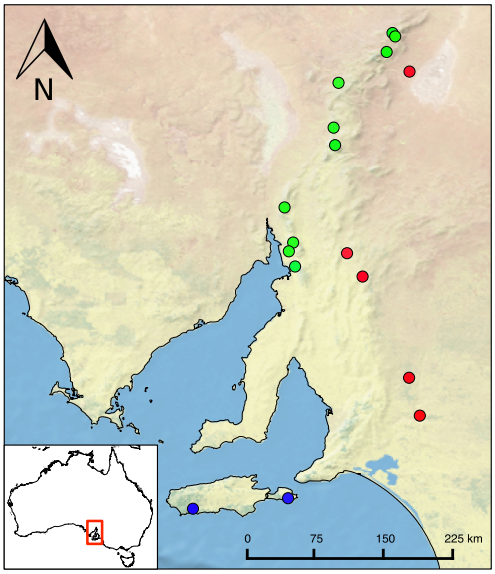

Figure 1. Sampling locations along the Adelaide Geosyncline Region where leaf samples from 5-8 individuals were collected for DNA extraction from each of 17 sites. Colours of dots represent population assignment to one of three genetic clusters identified in both STRUCTURE and DAPC analysis (blue= Kangaroo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

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	$_{ENV}$PC1	$_{ENV}$PC2
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
Aligator 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

Environmental data

We sampled across a steep temperature and rainfall gradient, where mean annual maximum temperature ranged from 19.5°C to 28.5°C and mean annual precipitation from 467 mm to 134 mm in the most southern and most northern sites respectively. To summarise and reduce redundancy among the many environmental variables that vary across the region, we ran a principle components analysis (PCA) on fifteen environmental parameters downloaded from the Atlas of Living Australia (http://www.ala.org.au, accessed 20 September 2015; table S1) using the `PCA` function in the `R` package `FactoMineR`, after having first standardised 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 principle components 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).

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/5x5km/year)	0.02	-0.73

Candidate genes and genotyping

We used the published *D. viscosa* transcriptome (Christmas *et al.* 2015b) to design MYbaits hybrid capture baits (MYcroarray, USA) for the capture of a set of 970 genes. 80mer baits with a 2x 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 non-synonymous mutations among two *D. viscosa* 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 Machery-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, Michigan, USA) with 12 cycles of post-capture 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 and Durbin, 2009). The indexed reference was created using default settings. Resulting SAM files were compressed, 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 in order 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 dataset if the coefficient of determination between the pair was greater than 0.5, thus removing SNPs showing strong signals of LD. Following these filtering steps 8,462 SNPs remained for downstream analysis.

Neutral genetic structure

Neutral population genetic structure amongst our samples was assessed using a putatively neutral subset of the main SNP dataset (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 of 0.05, resulting in a conservative output of neutral SNPs. The resultant SNP set was filtered to reduce linkage disequilibrium 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 dataset, 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 non-model-based multivariate approach, which seeks discriminating functions between groups of individuals while minimising 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 and vonHoldt 2012) was used to calculate ΔK in order to assess the most likely value of *K* and then results from replicate runs for this *K* value were combined using CLUMPP (Jakobsson and Rosenberg 2007) with default settings.

For DAPC, genetic data were first transformed into uncorrelated components using principal component analysis (PCA). The number of genetic clusters was then defined using k-means, a clustering algorithm that looks for the value of *K* that maximises 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`, in order 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 false discovery rates (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 in order to conservatively identify those SNPs that consistently showed signatures of selection.

Outlier detection. 8,462 SNPs were analysed in the F_{ST}-based outlier analysis software BayeScan v. 2.0 (Foll & Gaggiotti 2008) in order to identify outlier SNPs, i.e. SNPs that demonstrated stronger and weaker differentiation among populations than would be expected under neutral expectations. Such SNPs may be demonstrating signatures of diversifying or stabilising 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 in order to assess which is 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 false discovery rate 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 generalised 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 dataset 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 to 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 non-parametric Spearman's rank correlation coefficient, ρ. SNPs with a $\log_{10}$BF >0.5 as well as an absolute value of ρ>0.3 (where ρ ranges from -1 to 1) were therefore considered as robust candidates demonstrating signatures of selection.

Latent factor mixed models (LFMM) 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 false discovery rate (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 & S2).

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 8,462 input SNPs (2.36%), with F_{ST} values ranging from 0.265 to 0.518 (fig. S3).

Neutral genetic structure

Both STRUCTURE and DAPC outputs agreed that the most likely value of *K* was three (fig. S4), 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 in order 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 ρ >0.3) to PC1 and PC2 respectively (fig. 2). LFMM identified 1,011 (11.95 %) and 919 (10.86 %) SNPs with significant correlations (corrected P < 0.001) to PC1 and PC2 respectively. For PC1, 7 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 6 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).

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

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 elements-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

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 elements-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	2.788	1.82E-07	0.531	-0.308	-2.191	0.123	BE,LFMM

	protein							
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

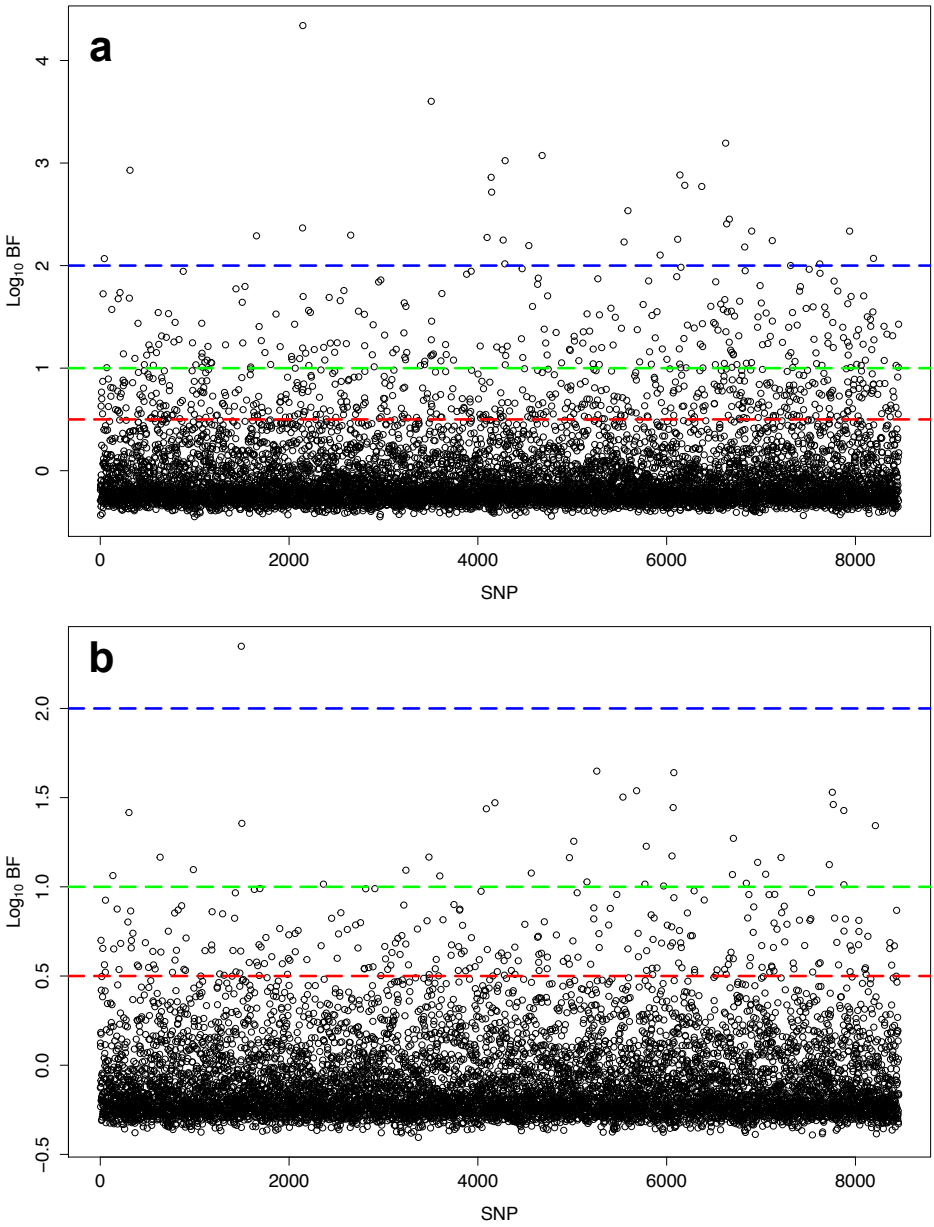

Figure 2. Manhattan plots of genetic differentiation associated with environmental parameters. Log_{10} Bayes Factors (BF) from BAYENV2 output represent associations between SNP allele frequency variation with variation in (a) PC1 and (b) PC2, the first two principal components from a principal component 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.

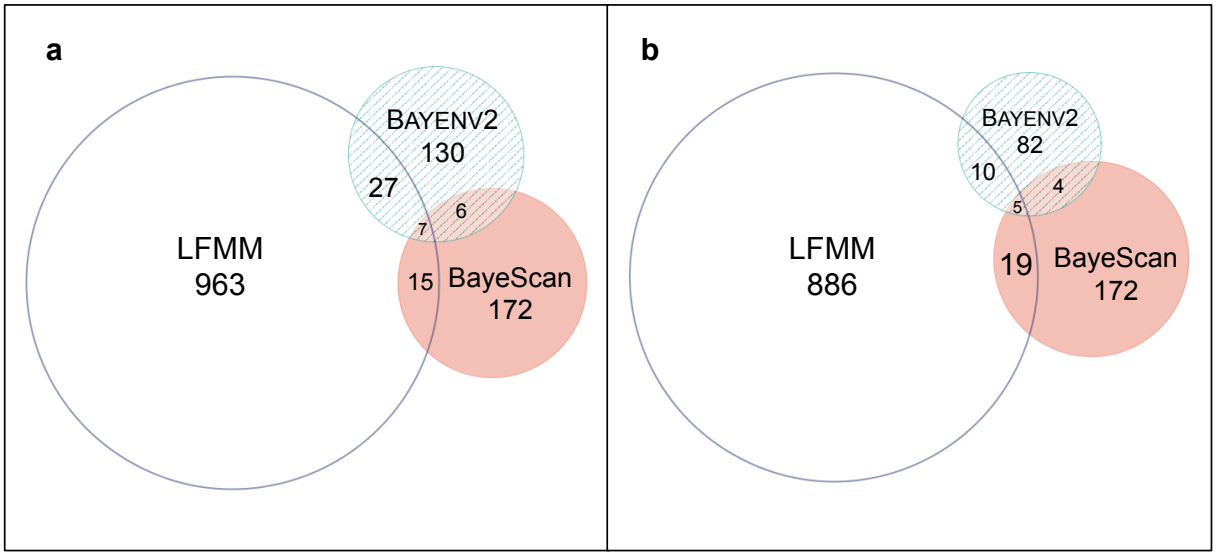

Figure 3. Venn diagrams comparing the significant outputs from an outlier detection method (BayeScan) and two genotype-environment analysis methods (LFMM and BAYENV2), which look for significant correlations between SNPs and (a) PC1 and (b) PC2 from a principal components analysis of environmental variables.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

Significant SNP gene functions

BLAST searches of assembled transcriptome contigs to the NCBI non-redundant protein database followed by gene ontology 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 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.

Of the 45 targeted aquaporin and ABA genes, two aquaporin and two ABA-related genes were found to 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 abscisic acid 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).

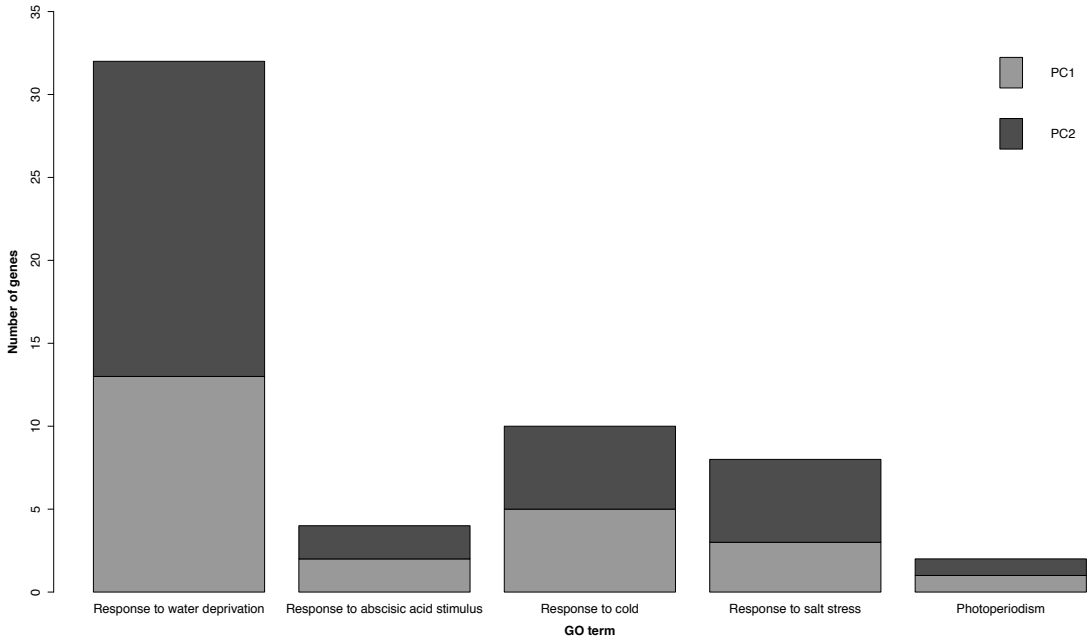

Figure 4. Number of genes (with GO term relating to an environmental response) containing SNPs that correlate significantly with either PC1 or PC2.

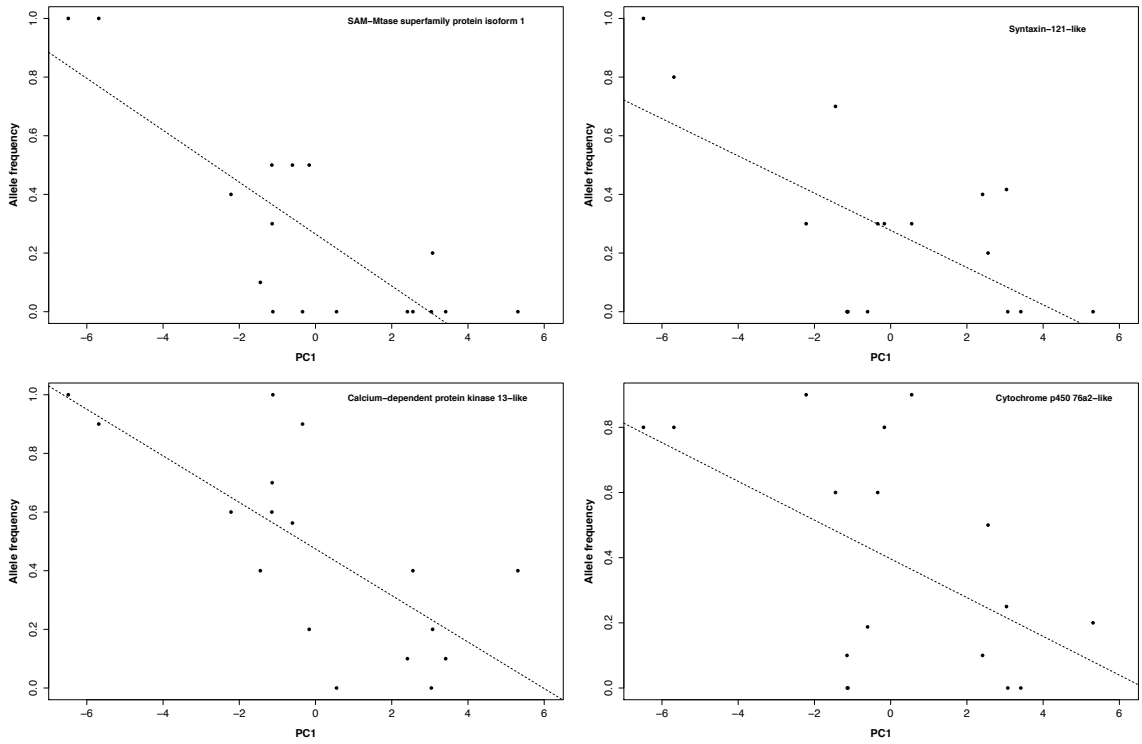

Figure 5. Example correlations between allele frequency and environmental factors summarised by PC1 (temperature- and water-related) for the four SNP markers with the greatest Bayes Factor score that were identified as being significant outliers by all methods (BayeScan, LFMM, BAYENV2). Linear regression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

Redundancy analysis

Redundancy analysis showed that 12.3% of the variation across all 8,462 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 to PC1 and PC2 indicates IBD is present within the data, which may confound the identification of signatures of selection.

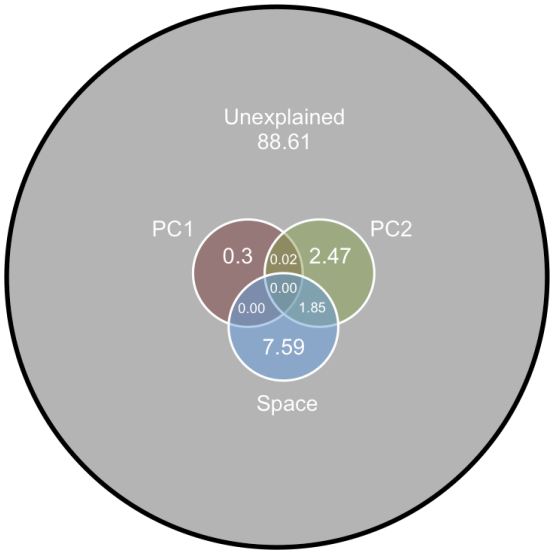

Figure 6. Decomposition of among-population variation of 17 populations of *Dodonaea viscosa* ssp. *angustissima* distributed along an environmental gradient using redundancy analysis (RDA). Variation is decomposed into environmental components (PC1; red, and PC2; green), a spatial component (blue), overlap between these components, and an unexplained component (grey) for all 8,462 SNPs.

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. By 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 redundancy analysis showed that only a small fraction of the variation (<2.82%) across the 8,462 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 8,462 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 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, since 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 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 summarised 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 *et al.* 2012; Guerin & Lowe 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 gene ontology 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 gene ontology 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.

Non-model 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 gene ontology annotations can then provide information on the putative functions of transcribed sequences (De Wit *et al.* 2012) and target 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 in order 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).

Acknowledgments

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.
Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**, 697-709.
Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics* **9**, 323.
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.
Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980.
Bonhomme M, Chevalet C, Servin B*, et al.* (2010) Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* **186**, 241-262.
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*, 1-16.
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, Albers CA, Banks E, DePristo MA *et al* (2011). The variant call format and VCFtools. *Bioinformatics* **27**(15)**:** 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.
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*, pp. 43-58. Springer.
Guerin GR, Lowe AJ (2012) Leaf morpholgy shift: new data and analysis support climate link. *Biology Letters*, rsbl20120860.
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.
Hartmann M-A, Benveniste P (1987) Plant membrane sterols: Isolation, identification, and biosynthesis. *Methods in Enzymology* **148**, 632-650.
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**(14)**:** 1754-1760.
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 JONE (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. 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**(7)**:** 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.
Roje S (2006) S-Adenosyl-L-methionine: beyond the universal methyl group donor. *Phytochemistry* **67**, 1686-1698.
Savolainen O, Lascoux M, Merila J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics* **14**, 807-820.

Schaller H (2003) The role of sterols in plant growth and development. *Progress in Lipid Research* **42**, 163-175.
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.
Wentzinger LF, Bach TJ, Hartmann M-A (2002) Inhibition of squalene synthase and squalene epoxidase in tobacco cells triggers an up-regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Plant Physiology* **130**, 334-346.
Wright S (1943) Isolation by distance. *Genetics* **28**, 114.
Zhu C, Schraut D, Hartung W & Schaffner AR (2005) Differential responses of maize MIP genes to salt stress and ABA. *Journal of Experimental Botany* **56,** 2971–2981.
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.
Zhu J-K (2002) Salt and drought stress signal transduction in plants. *Annual Review of Plant Biology* **53**, 247.

Data accessibility

Sequencing reads are archived at the NCBI SRA under the accession SRP077342, associated with BioProject PRJNA326697. The variants file is available as a supplementary file (File_S1.vcf). Details of transcriptome data archiving can be found in Christmas *et al*. (2015b).

Author contributions

MJC, EB, MFB, and AJL designed the research. MJC and EB performed field collections and laboratory work. MJC analysed the data. MJC wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

Supplementary information

Table S1: Environmental data downloaded from the Atlas of Living Australia for each of the sampling locations. These data were reduced to two principal components in a principal component analysis.

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

Figures S1 & 2: Histograms of corrected P-values from the Latent Factor Mixed Model (LFMM) analysis for PC1 and PC2 respectively.

Figure S3: Global outlier detection among 8,462 SNPs in 17 *Dodonaea viscosa* ssp. *angustissima* populations from South Australia in BayeScan with prior odds = 100. The vertical line represents a false discovery threshold of 0.05.

Figure S4: Graphs showing outputs from neutral population genetic analyses in STRUCTURE and DAPC. A) Bayesian Information Criterion (BIC) value for each value of K tested in DAPC where the elbow in the graph indicates the most likely value of K to be three; B) Delta K value for each value of K tested in STRUCTURE, where K=3 has the greatest delta K value and therefore the greatest support; C) Percentage assignment of each individual to the clusters identified in STRUCTURE when K=3, with colours representing each of the three clusters (blue = Kangaroo Island cluster; green = Flinders Ranges cluster; red= Eastern cluster).

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

Conclusions and future directions

In this thesis, I have developed the first genome-wide resources for *D. viscosa*, a widely distributed, ecologically important species previously lacking such resources. I have demonstrated the utility of these resources in addressing questions regarding phylogeography, population genetic structure and diversity as well as investigation of adaptive processes along a strong environmental gradient. Analyses similar to those performed here applied to populations across the wider distribution of the species could be used to further confirm the findings of this thesis, as well as address questions of population connectivity and adaptive potential over a larger scale.

The transcriptome work presented in this thesis represents an excellent alternative to whole-genome sequencing for non-model species lacking in genomic resources. Sequencing the transcriptome provided a route to understanding potential functional significance of the observed genetic variation in the absence of an annotated genome. One potential pitfall of the use of the transcriptome in this way is that its annotation relies upon matching the sequenced gene regions to putatively orthologous genes of known function in closely related species. If the required similarity between sequences in the BLAST analysis is too strict, orthologous genes that have diverged sufficiently may be missed. If too relaxed, sequences maybe incorrectly matched with non-orthologous genes. This can result in the incorrect annotation of gene sequences. Also, any genes without orthologs within the available databases will be left unannotated and so ignored, despite their potential functional importance. This is something to keep in mind as research moves from annotation, variant screening and identification of signatures of selection presented here to more functional analysis such as gene expression analysis and drawing links between genotype and phenotype.

The sampling efforts for the studies presented here were sufficient for the questions I set out to address, but not extensive. In order to achieve a more complete picture of the phylogenetic relationships presented in chapter three, both within the *D.viscosa* species complex and with its sister taxa, a wider sampling effort would be needed including samples from its overseas distribution. As well as this, genetic information could be paired with phenotypic measurements of, for example, leaf traits in order to better understand the links between phylogeny and morphology: an assessment of the variation in leaf morphology in this highly divergent species could help to identify the ancestral state of the species, from which different forms have evolved. A wider population sampling effort outside of the Adelaide geosyncline, particularly into more eastern and western areas may help to better ascertain the origins of the genetically distinct population clusters identified in chapter five.

I believe the main strength of this thesis lies in the development of a genomic resource that has then been used to demonstrate clear signatures of selection driven by environmental factors. This shows that the environment, particularly climate, has acted as a strong selective force on these populations in the past. Under contemporary climate change, climate will continue to act as strong selection pressure on these populations and further adaptation via migration and gene flow, as well as plasticity, will be required. As a widely distributed and diverged species, *D. viscosa* appears to be well placed to adapt to these on-going environmental changes. Our findings, although strong in their own right, can be further developed in a number of ways, as discussed below.

Isolation by distance along environmental gradients can result in confounding signals in the genetic data, where differences between populations due to neutral processes can show similar patterns of genetic variation similar to those expected from selection. Although a number of methods, including those used in chapter six, attempt to account for this, a more robust experimental design whereby population pairs are

sampled along a gradient could provide even stronger evidence for selection. This design would aim to sample geographically close (maximising gene flow) but environmentally distinct (maximising selection) pairs of populations, thus helping to prise apart signals of 'isolation by distance' from 'isolation by environment'.

A recent review into the use of landscape genetics to detect genetic adaptive variation has highlighted that current landscape genomic studies are mostly still at the exploratory stage and the next advancement should be to a confirmatory stage whereby the adaptive significance of loci identified as being under selection is tested (Manel et al. 2010). The work presented in this thesis could be described as exploratory and further confirmation of the findings is required. Evidence for adaptive variation in *D. viscosa* is growing, with the genomic data presented here supported by evidence of changes in leaf width through space and time (Guerin & Lowe 2012; Guerin et al. 2012) and changes in stomatal density in response to temperature (Hill et al. 2015).

The next steps required to further test the importance of genetic differentiation among populations and to relate these differences to phenotypic traits is to carry out common garden and reciprocal transplant experiments. A common garden experiment, whereby seed from multiple populations across the gradient is grown under common conditions, can be informative on whether the phenotypic differences observed in the field are due to genetic differences or are more a plastic response to environment. If phenotypic differences are observed then looking for correlations between phenotypes and genotypes would advance our findings to the more confirmatory stage. Reciprocal transplant experiments would allow for tests of local adaptation and plasticity and we could see whether the putatively adaptive differences identified through the genome scans actually are correlated with greater fitness in home environment, again providing further evidence that the genomic differences identified are adaptively important. To truly confirm the adaptive significance of our findings, rather than just strengthening

the identified correlations, robust breeding experiments where the effects of genotype and environment on phenotype can be fully partitioned would be required. Genome-wide association studies (GWAS), where associations are made between phenotype and genome-wide variation (as opposed to the relatively narrow distribution of variation investigated here) may also provide stronger links between phenotype and the causative genetic variation.

Growth of populations under common and controlled conditions would also allow for studies of gene expression. Here, we could test whether outlier genes identified in this thesis display different expression levels amongst the populations and, through a series of environmental manipulations, whether the expression levels change in response to environment, i.e. are some populations better at up-regulating or down-regulating certain genes in response to higher temperatures or drought? These types of studies could provide insight into how long-lived, sessile plant species are able to deal with environmental variation throughout their lifetime via adaptive plastic responses.

As part of this thesis, I did hope to include details of a common garden experiment. Unfortunately, and as is the way with science, this aspect of my research did not go to plan. I collected seed from several populations but germination rates were prohibitively low. Details of this study are included in a published report in the appendix. I feel that it is important to move from the exploratory types of studies presented within this thesis to the more confirmatory studies, as suggested by Manel et al. (2010), if we are to truly understand the functional importance of what we have found and therefore strengthen our predictions as to what it may mean for population persistence under contemporary climate change.

References

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. Nature Reviews Genetics, 11, 697-709.

Black I, William C, Baer C, Antolin M, DuTeau N (2001) Population genomics: genome-wide sampling of insect populations. Annual Review of Entomology, 46, 441-469.

Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. Molecular Ecology, 17, 3583-3584.

Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature, 431, 931-945.

Darwin, C (1859) The origin of species by means of natural selection, John Murray, London.

Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. Briefings in Functional Genomics, 9, 416-423.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, 269, 496-512.

Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, Feldmann H, Galibert F, Hoheisel J, Jacq C, Johnston M (1996) Life with 6000 genes. Science, 274, 546-567.

Guerin GR, Lowe AJ (2012) Leaf morpholgy shift: new data and analysis support climate link. Biology Letters, rsbl20120860.

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

Harrington MG, Gadek PA (2009) A species well travelled–the *Dodonaea viscosa* (Sapindaceae) complex based on phylogenetic analyses of nuclear ribosomal ITS and ETSf sequences. Journal of Biogeography, 36, 2313-2323.

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.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nature Reviews Genetics, 4, 981-994.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nature Methods, 7, 111-118.

Manel S, Joost S, Epperson BK, Holderegger R, Storfer A, Rosenberg MS, Scribner KT, Bonin A, Fortin MJ (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. Molecular ecology, 19, 3760-3772.

Mendel G (1866) Versuche über Pflanzenhybriden. Verhandlungen des Naturforschenden Vereines in Brunn 4: 3, 44.

Ouborg N, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. Trends in Genetics, 26, 177-187.

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. Trends in Ecology & Evolution, 24, 192-200.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, Smith M (1977a) Nucleotide sequence of bacteriophage [phi]X174 DNA. Nature, 265, 687-695.

Sanger F, Nicklen S, Coulson AR (1977b) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 74, 5463-5467.

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

Stapley J, Reger J, Feulner PG, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J (2010) Adaptation genomics: the next generation. Trends in Ecology & Evolution, 25, 705-712.

Van Verk MC, Hickman R, Pieterse CM, Van Wees S (2013) RNA-Seq: revelation of the messengers. Trends in Plant Science, 18, 175-179.

Watson JD, Crick FH (1953) Molecular structure of nucleic acids. Nature, 171, 737-738.

West JG (1984) A revision of *Dodonaea* Miller (Sapindaceae) in Australia. Brunonia, 7, 1-194.

Appendix

A1. List of publications. List of all publications, published, submitted or to be submitted, completed during my postgraduate studies as a named author.

A2. Report published in The South Australian Naturalist. This report was for a common garden experiment funded by The Field Naturalist Society of South Australia Lirabenda Endowment Fund. The report was a requirement of the funding and reported on the outcomes of the study.

A3. PLOS One paper. I was involved in the editing of the manuscript and creating figures for the publication.

A4. Transect paper submitted to *Trends in Ecology and Evolution.* I was involved in the initial workshop, which this paper developed from, provided text for the manuscript, assisted in reviewing and editing the manuscript.

A5. Global Ecology and Biogeography paper. I was involved in the initial discussions of the paper, planning content and structure, as well as reviewing and editing the manuscript.

A6. Down but not out. Overcoming the PhD battle.

A1. Publications

The following list contains publications that have either been published, submitted, or are soon to be submitted for publication that I have been involved in during the course of my postgraduate studies.

Breed MF, **Christmas MJ**, Lowe AJ (2014) Higher Levels of Multiple Paternities Increase Seedling Survival in the Long-Lived Tree *Eucalyptus gracilis*. PLOS One, 9, e90478. **Appendix A3**

Caddy-Retalic S, Andersen A, Aspinwall MJ, Breed MF, Byrne M, **Christmas MJ**, Dong N, Evans BJ, Fordham DA, Guerin GR, Hoffmann AA, Hughes AC, van Leeuwen SJ, McInerney FA, Prober SM, Rossetto M, Rymer PD, Steane DA, Lowe AJ (2015) Networked bioclimatic transects are powerful observatories of global change. **Submitted to Trends in Ecology and Evolution. Appendix A4**

Christmas MJ, Biffin E, Lowe AJ (2015) Transcriptome sequencing, annotation and polymorphism detection in the hop bush, *Dodonaea viscosa*. BMC Genomics, 16, 803. **(Chapter 4)**

Christmas MJ, Breed MF, Lowe AJ (2016) Constraints to and conservation implications for climate change adaptation in plants. Conservation Genetics, 17, 305-320. **(Chapter 2)**

Christmas MJ, Biffin E, Breed MF, Lowe AJ (2016) Determining the level and structure of population genomic variation for the narrow-leaf hopbush across a continental biodiversity refugium - the Adelaide Geosyncline. **To be submitted to Nature Scientific Reports. (Chapter 5)**

Christmas MJ, Biffin E, Breed MF, Lowe AJ (2016) Finding needles in a genomic haystack: targeted sequencing to identify signatures of selection in a non-model species. Molecular Ecology**,** accepted**. (Chapter 6)**

Christmas MJ, Biffin E, Lowe AJ (2016) Geography, not morphological variation, defines genetic partitioning in a genome-wide SNP screen of *Dodonaea viscosa* (Sapindaceae). **To be submitted to The Journal of Biogeography. (Chapter 3)**

Guerin GR, Martín-Forés I, Biffin E, Baruch Z, Breed MF, **Christmas MJ**, Cross HB, Lowe AJ (2014) Global change community ecology beyond species-sorting: a quantitative framework based on Mediterranean-biome examples. Global Ecology and Biogeography, 23, 1062-1072. **Appendix A5.**

Christmas, M.J. (2015). Adaptation along a climatic gradient: is trait plasticity or genetic adaptation responsible in the narrow-leaf Hop-bush, Dodonaea viscosa ssp. Angustissima? A lesson in collecting seed for common garden experiments. *The South Australian Naturalist,* 89(1), 27-33.

A3: PLOS One Paper

OPEN ACCESS Freely available online

PLOS ONE

Higher Levels of Multiple Paternities Increase Seedling Survival in the Long-Lived Tree *Eucalyptus gracilis*

Martin F. Breed[1], Matthew J. Christmas[1], Andrew J. Lowe[1,2]*

1 Australian Centre for Evolutionary Biology and Biodiversity (ACEBB) and School of Earth and Environmental Sciences, University of Adelaide, Adelaide, South Australia, Australia, **2** State Herbarium of South Australia, Science Resource Centre, Department of Environment, Water and Natural Resources, Adelaide, South Australia, Australia

Abstract

Studying associations between mating system parameters and fitness in natural populations of trees advances our understanding of how local environments affect seed quality, and thereby helps to predict when inbreeding or multiple paternities should impact on fitness. Indeed, for species that demonstrate inbreeding avoidance, multiple paternities (*i.e.* the number of male parents per half-sib family) should still vary and regulate fitness more than inbreeding – named here as the 'constrained inbreeding hypothesis'. We test this hypothesis in *Eucalyptus gracilis*, a predominantly insect-pollinated tree. Fifty-eight open-pollinated progeny arrays were collected from trees in three populations. Progeny were planted in a reciprocal transplant trial. Fitness was measured by family establishment rates. We genotyped all trees and their progeny at eight microsatellite loci. Planting site had a strong effect on fitness, but seed provenance and seed provenance × planting site did not. Populations had comparable mating system parameters and were generally outcrossed, experienced low biparental inbreeding and high levels of multiple paternity. As predicted, seed families that had more multiple paternities also had higher fitness, and no fitness-inbreeding correlations were detected. Demonstrating that fitness was most affected by multiple paternities rather than inbreeding, we provide evidence supporting the constrained inbreeding hypothesis; *i.e.* that multiple paternity may impact on fitness over and above that of inbreeding, particularly for preferentially outcrossing trees at life stages beyond seed development.

Citation: Breed MF, Christmas MJ, Lowe AJ (2014) Higher Levels of Multiple Paternities Increase Seedling Survival in the Long-Lived Tree *Eucalyptus gracilis*. PLoS ONE 9(2): e90478. doi:10.1371/journal.pone.0090478

Editor: Giovanni G. Vendramin, CNR, Italy

Received August 1, 2013; **Accepted** February 3, 2014; **Published** February 28, 2014

Funding: This work was supported by Australian Research Council Linkage project (www.arc.gov.au/, LP110200805) and South Australian Premier's Science and Research Fund awarded to AJL (http://www.dfeest.sa.gov.au/); funding from the Native Vegetation Council of South Australia (www.nvc.sa.gov.au/, grant 09/10/27), Nature Foundation SA Inc. (http://www.naturefoundation.org.au/), Australian Geographic Society (http://www.australiangeographic.com.au/society/), Biological Society of South Australia (http://www.biologysocietysa.com/), Field Naturalist Society of South Australia (http://www.fnssa.org.au/), Wildlife Preservation Society of Australia (http://www.australianwildlife.net.au/); and NCCARF Travel Grants (http://www.nccarf.edu.au/) awarded to MFB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: andrew.lowe@adelaide.edu.au

Introduction

The realised inbreeding rate of monoecious trees, when estimated from mature seeds or seedlings, is usually constrained below their actual inbreeding rate [1]. This occurs because inbreeding usually imposes fitness costs at early stages of reproduction [1–4], via expression of deleterious recessive alleles [5–7], leading to the abortion of these inbred offspring. Selfing and biparental inbreeding (*i.e.* related breeding events) should both be constrained, but selfing more so than biparental inbreeding due to the higher inbreeding coefficients that are generated during selfing.

Monoecious trees also routinely exhibit multiple paternities because they receive great amounts of pollen from a large diversity of donors [8–10]. Regulation of the supply and diversity of pollen is largely controlled by a tree's local environment, often regulated by local variation in pollination services and the effective density of pollen donors [11,12]. Fertilisation success is then filtered by the availability of receptive flowers (*i.e.* tree phenology) and the genetic compatibility of pollen-ovule combinations.

Tree fitness should increase with more multiple paternities because, firstly, higher levels of multiple paternities should facilitate more complimentary pollen-ovule combinations by generating greater opportunities for female choice for superior pollen and/or pollen competition [8,13]. Thus, females that have lower levels of multiple paternities have, by definition, placed greater weight on their compatibility with fewer pollen donors and undergone suboptimal levels of mate discrimination – the bet hedging hypothesis [13]. Secondly, higher levels of multiple paternities should also give rise to greater genetic diversity within progeny arrays (*i.e.* greater genotype × environment interactions within progeny arrays) [13]. Theory predicts that greater offspring genetic diversity should facilitate higher mean offspring fitness as a result of, for example, more effective resource exploitation from offspring [13]. Consequently, females that receive lower levels of multiple paternities should have less genetically diverse offspring, increasing the risk that a high proportion of her offspring will be poorly adapted to local environments, particularly in unpredictable or changing environments.

Given that multiple paternities are likely to impose fitness benefits, and that trees usually maintain low realised inbreeding levels, divergence from this high multiple paternity state is expected to produce lower fitness offspring - named here as the 'constrained inbreeding hypothesis'. Furthermore, trees should be good candidates to detect the effect of multiple paternities because

Appendix

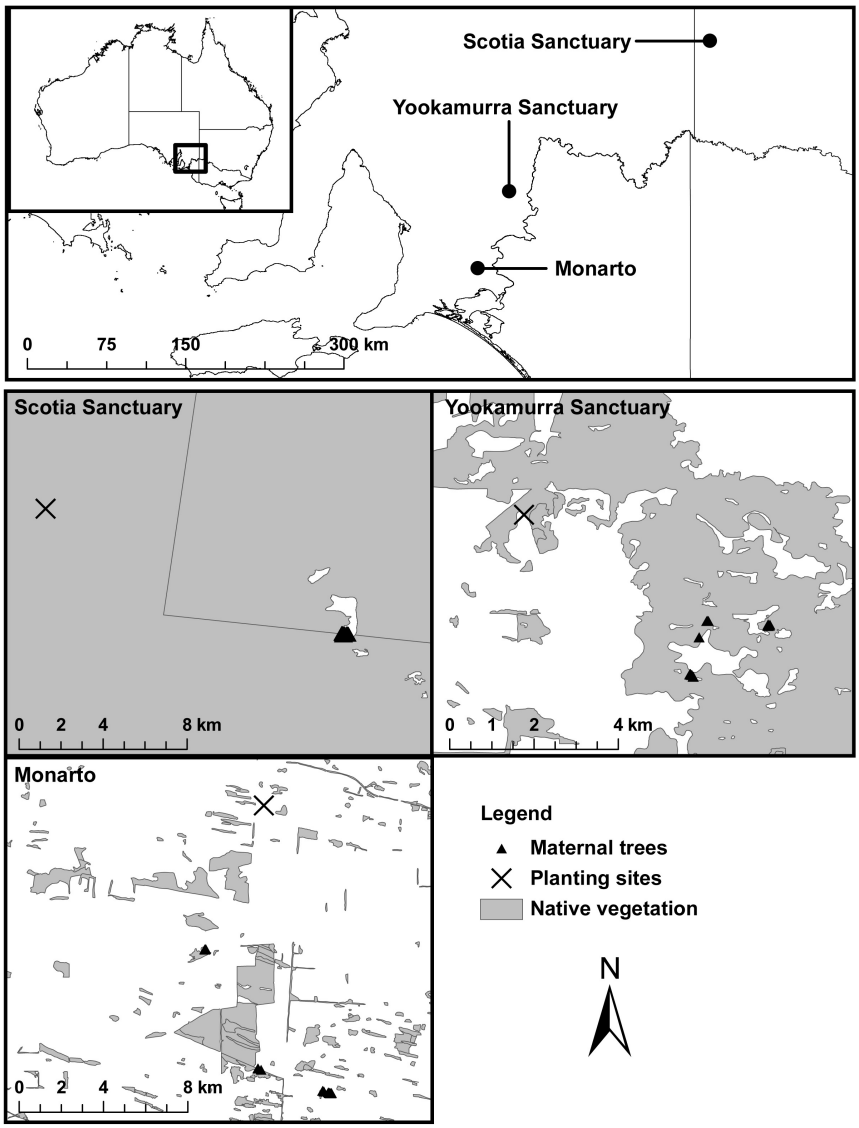

Figure 1. Map showing location of *Eucalyptus gracilis* maternal trees and planting sites. Maps show samples from the three populations in the Murray Darling Basin, Australia. Insert maps show greater spatial information on sampled populations. Reciprocal transplant planting locations shown at each planting site by a cross (x).
doi:10.1371/journal.pone.0090478.g001

Appendix

trees tend to have large lifetime fecundities, which result in strong selection acting at early life stage (*e.g.* seedlings) [9,14].

Inbreeding and inbreeding depression are possibly present at late stages of offspring development in natural tree populations (*e.g.* [15]), but there has been little emphasis on the potential magnitude of fitness effects of multiple paternities in natural tree populations. Indeed, there are numerous examples of fragmented tree populations where mature seed/seedling fitness was studied (*i.e.* after the effects of early inbreeding depression), but the studied trees maintained strong inbreeding avoidance [11,16,17]. Consequently, there was insufficient variation in inbreeding to observe inbreeding-fitness correlations in these studies [18], but there were detectable effects of multiple paternities on fitness [11,16,17]. These are significant observations since seeds and seedlings are both important life stages for the demographic trajectories of tree populations [19], and both are critical life stages for land managers who require good quality trees (*e.g.* for ecological restoration; [20,21]).

Variation in local pollination services can greatly affect the extent of multiple paternities, related breeding events, and selfing observed in trees [22–25]. Furthermore, over 20 years of effort has been invested into reporting changes in local pollination by estimating tree mating system parameters in fragmented populations [11,26–28].

Despite the established theory of optimal pollinator foraging behaviour in natural populations [29,30], as well as the observed fitness impacts due to shifts in pollinator foraging in fragmented populations [11,16,17,22,31–34], it is surprising that not more studies have explored how natural variation in these mating system parameters may impact on fitness of intact tree populations. Indeed, it may be that natural populations of trees also tend to experience minimal inbreeding depression at later life stages (*i.e.* after seed development) because of strong early inbreeding depression [1–4], but this is currently unknown. As a consequence of this expectation, the degree of multiple paternities may have stronger effects on fitness at late life stages than inbreeding (*e.g.* mature seed, seedling), even in natural populations of trees.

To test whether multiple paternities have stronger effects on fitness at late life stages than inbreeding, we combined assessments of mating system parameters and seedling establishment rates of open-pollinated progeny arrays (measured 16 months after germination) of *Eucalyptus gracilis* F. Muell. (white mallee or yorrell). We estimated four different mating system parameters; two commonly used measures of multiple paternities (correlated paternity and the number of full-sibships) and two measures of inbreeding (outcrossing rate and biparental inbreeding). We sampled progeny arrays from three natural and mostly intact populations across the Murray-Darling Basin in southern Australia (Figure 1). Like other eucalypts [3,25], we expect that *E. gracilis* will express strong inbreeding avoidance, constraining variation in the measured inbreeding parameters, and therefore we expect to detect high levels of outcrossing and low levels of biparental inbreeding. This inbreeding constraint should reduce the probability of detecting inbreeding-seedling fitness correlations [18]. However, the levels of multiple paternities within progeny arrays of *E. gracilis* should not be constrained, but rather should vary across the families we sampled according to their local pollination services. Additionally, *E. gracilis* should be a good candidate to explore the effects of inbreeding and multiple paternities because, like other eucalypts, *E. gracilis* should be a strong outcrosser and have high lifetime fecundity [3,35], and as a consequence of these life history traits, *E. gracilis* offspring should have significant genetic load [5] and strong selection should act at the seedling life stage [9]. We therefore expect that this study system is suitable for isolating and detecting multiple paternity-seedling fitness correlations.

Materials and Methods

Study Species

Eucalyptus gracilis is a multi-stemmed, sclerophyllous tree common throughout the sand and sand-over-limestone soils of the Murray-Darling Basin, southern Australia [36,37]. *Eucalyptus gracilis* generally grows from 4 to 8 m high, it has small white hermaphroditic flowers (diameter of mature flowers with reflexed stamens: <15 mm) and is pollinated primarily by small insects and, to a lesser degree, by birds and small marsupials [38,39].

Eucalypt flowers are protandrous (*i.e.* male reproductive phase precedes female phase within flowers) and flower development within and between inflorescences is sequential and gradual. Therefore, flowers in male or female phase may be in close proximity, allowing geitonogamous selfing to occur (*i.e.* pollination from another flower on the same plant; [40]). Data from closely related eucalypts suggest that the species investigated here probably has a late-acting self-incompatibility mechanism, resulting in mixed mating to preferential outcrossing (t_m generally $>$ 0.70; [3]). Serotinous fruit (*i.e.* seeds released in response to an environmental trigger) are held over numerous years, with drying triggering seed-release. Seeds are small (<2 mm diameter) and gravity dispersed. Based on data from *E. incrassata* and our field observations, ants generally exhaust soil seed banks, except during particularly heavy seed release such as post fire [41,42].

Seed Collection

Open-pollinated seeds were collected from across the canopies of trees located in three sites in the Murray-Darling Basin (Figure 1). Scotia Sanctuary and Yookamurra Sanctuary trees ($n_{Scotia} = 18$; $n_{Yookamurra} = 20$; Figure 1) were from large intact woodlands, with no history of known anthropogenic disturbance. Monarto Woodland trees ($n = 20$; Figure 1) were from small remnant woodlands. Small remnant woodlands were natural habitats surrounded by agricultural land, but again with no history of known anthropogenic disturbance. *E. gracilis* is a common overstory tree at each site ($N > 1000$), and one of many *Eucalyptus* species common throughout the semi-arid Murray Darling Basin [36]. We avoided sampling nearest neighbours and we sampled numerous stands per site where possible, although road access limited our sampling to one stand at Scotia Sanctuary. *E. gracilis* stands at Yookamurra Sanctuary had significant higher density than stands at both Scotia Sanctuary and Monarto Woodland (trees ha^{-1}: Monarto Woodland = 23.67, SD = 2.29; Yookamurra Sanctuary = 49.33, SD = 4.63; Scotia Sanctuary = 20.42, SD = 3.24).

Seedling Establishment Trials

Fifteen replicates of approximately 20 seeds from each tree were sown on February 1st 2010. Germination was conducted under semi-controlled glasshouse conditions in Adelaide, South Australia (S34°55'05", E138°36'18"). All seedlings were moved to full-sun at the Mt Lofty Botanic Gardens, South Australia (S34°59'03", E138°43'08") after four weeks in glasshouse conditions. Crates of seedlings were shifted and rotated approximately weekly to avoid confounding effects of location in glasshouse/nursery. The most central seedling within each pot was chosen, and non-central additional seedlings were removed over the subsequent weeks prior to planting. We hoped to minimise selection on seedling fitness with this process, but cannot rule out that selection for fitter individuals may have taken place. Glasshouse and nursery

Appendix

environments may allow inferior seedlings to survive when compared to seedling survival under natural woodland conditions. This bias should be consistent across progeny arrays and, under glasshouse/nursery environments, additional biases should be controlled for (*e.g.* competition, demographic or environmental stochastic effects).

Plantings took place at Scotia Sanctuary, Yookamurra Sanctuary and at Monarto Woodland between May and June 2010 (seed source sample sizes: $n_{\text{Monarto Woodland}} = 294$; $n_{\text{Yookamurra}} = 295$; $n_{\text{Scotia}} = 264$; reciprocal transplant experiment locations shown in Figure 1). We implemented a randomised complete block design [43]. Planting sites were located in close proximity to 'local' maternal trees (<13 km in all cases; Figure 1). Planting sites were prepared by rotary hoeing to remove residual surface vegetation, parallel rip-lines were drawn through at 3 m intervals, and seedlings were spaced at 2 m intervals. A 200×200×500 mm tree guard (Global Land Repairs, Fyshwick) surrounded each seedling to protect against herbivores (*e.g.* rabbits). This reciprocal transplant experiment was originally planned to explore adaptive divergence in *E. gracilis* as well as this mating system analysis. However, since we found such weak neutral genetic differentiation and no divergence in establishment rate (see results presented below), this investigation was set aside and the mating system analysis was conducted in more depth.

In May 2011, we counted the number of seedlings that had died 12 months after planting (*i.e.* 16 months after germination). This fitness proxy included deaths that had occurred at each planting site, whether local or non-local. We used the ratio of mortality counts and progeny array size as the variable in subsequent analyses. Using seedling mortality as our only fitness proxy means that we can only speculate about earlier or later stages of the life cycle of *E. gracilis* (*e.g.* germination, fecundity), and therefore our results need to be interpreted in this context.

Microsatellite Genotyping

Leaf tissue was collected from each seedling prior to planting and DNA was extracted using the Machery-Nagel Nucleospin Plant II Kit at the Australian Genome Research Facility (AGRF, Adelaide, Australia). Eight direct-labelled microsatellite markers were selected from the set of EST-derived markers by Faria *et al.* ([44]; EMBRA1382; EMBRA2002; EMBRA1445; EMBRA1284; EMBRA1928; EMBRA1468; EMBRA1363). A BLAST search was performed for each microsatellite sequence using accession numbers in Faria *et al.* [44] and resulted in no significant hits with genes with a known function. EMBRA1363 produced two unlinked and scoreable PCR products (EMBRA1363a and b). PCR was performed in a single 10 μL multiplex PCR containing 1 μL template DNA (*ca.* 20 ng μL^{-1}), 5 μL 2× Qiagen Multiplex PCR Master Mix (Qiagen, Hilden, Germany), 3 μL of nuclease-free water, 1 μL of primer mix with each primer at 2 μM concentration. Standard Qiagen Multiplex PCR conditions were used with an initial activation step at 95°C for 15 minutes, 40 cycles of denaturation at 94°C for 30 seconds, annealing at 60°C for 90 seconds and extension at 60°C for 60 seconds, with final extension at 60°C for 30 minutes. LIZ500 size standard was added to samples and fragments were separated on an AB3730 genetic analyser with a 36 cm capillary array (Applied Biosystems, Foster City, MA, USA) at AGRF. Alleles were automatically called using GeneMapper software (Applied Biosystems) and double-checked manually.

Data Analysis

Each maternal tree was presumed to reflect patterns of population genetics pre-clearance since all sampled trees were estimated to be >80 years old [45,46] and most land clearance occurred <80 years ago [47]. Maternal genotypes were used to screen for null alleles in MICRO-CHECKER [48] and INEst [49], where INEst employs a method that produces un-biased estimates of null allele frequencies for populations that experience inbreeding. GENEPOP on the web (http://genepop.curtin.edu.au) was used for tests for heterozygote deficit/excess and linkage disequilibrium, applying sequential Bonferroni correction for multiple testing where appropriate. Additionally, the per-locus probability of paternity exclusion (Q) and combined probability of paternity exclusion (QC) were estimated in GENALEX [50]. Pairwise population genetic differentiation parameters G_{ST_est} [51] and D_{est} [52] were estimated in GENODIVE [53].

We estimated the following genetic diversity parameters for maternal tree and progeny groups using GENALEX: number of alleles (A), Nei's unbiased expected and observed heterozygosity (H_E and H_O, respectively; [54]). In addition, the fixation index (F) was estimated for each population. To account for differences in sample size, we estimated the rarefied mean number of alleles per locus (AR) using HP-RARE [55]. All samples that failed amplification at more than three loci were excluded ($n = 5$).

We estimated the following mating system parameters in MLTR [56]: multilocus outcrossing rate (t_m), biparental inbreeding (t_m–t_s) and multilocus correlated paternity (r_p). Families were bootstrapped 1000 times to calculate variance estimates for each parameter. Family-level mating system parameters were estimated in the same way except that individuals within families were bootstrapped 1000 times to calculate variance estimates. To further investigate the role of the multiple paternities, we estimated the number of full-sib groups within progeny arrays using KINALYZER [57,58], implementing the 2-allele algorithm, and scaled this value to the size of progeny arrays (k_n). Selfed offspring were excluded from this analysis.

We used general linear models in a maximum likelihood, multi-model inference framework in R v. 2.12.1 (R Project for Statistical Computing, http://www.r-project.org; [59] to test for hypothesised relationships between *E. gracilis* establishment success (counts of seedling mortality per family, scaled to size of family) and four genetic predictors: multilocus outcrossing rate (t_m), biparental inbreeding (t_m–t_s), correlated paternity (r_p) and the number of full-sibships within progeny arrays scaled to size of progeny array (k_n). We relied on Akaike's Information Criterion corrected for small sample sizes (AIC_c) for model selection [59].

Ethics Statement

All relevant permits and approvals were obtained for the work presented in this study. Work conducted in Scotia and Yookamurra Sanctuary was done with written approval from the landowner, Australian Wildlife Sanctuary. Work conducted in Monarto Woodland was approved by Primary Industries and Regions SA (PIRSA-ForestrySA and Rural Solutions). Work conducted on Ferries-McDonald Conservation Park and Monarto Conservation Park was approved by the South Australian Department of Environment and Heritage (now Department of Environment, Water and Natural Resources). No protected species were sampled.

Data Access

Data accession numbers have not yet been obtained, but they will be provided in the event that our manuscript is accepted for publication.

Results

Genetic Marker Quality

We genotyped open-pollinated progeny from 20 trees from Monarto Woodland ($n = 287$), 20 trees Yookamurra Sanctuary ($n = 291$) and 18 trees from Scotia Sanctuary ($n = 260$) (progeny array size data reported in Table 1). A total of 115 different alleles were identified across progeny (Table S1). The combined probability of paternity exclusion if neither parent is known indicates good resolution for the genetic markers used (QC = 1.00). No significant excesses or deficits of heterozygotes were observed in the groups of maternal trees and we found no significant null alleles at any loci within any population. No significant linkage disequilibrium was observed between pairs of loci scored in maternal trees after adjustment for multiple testing.

Genetic Diversity and Population Differentiation

There were no significant differences in allelic richness, expected and observed heterozygosity between progeny and maternal trees (all *t*-test $P > 0.05$; Table 1). Genetic differentiation between populations was weak but significant (all genetic differentiation values < 0.15; all $P < 0.05$; Table S2). Yookamurra Sanctuary was more genetically similar to Monarto Woodland than Scotia Sanctuary, reflecting the spatial proximity of populations (Figure 1). Accordingly, Monarto Woodland and Scotia Sanctuary were the most genetically differentiated population pair.

Mating System Parameters, Stand Density and Seedling Establishment

Each population was strongly outcrossed ($t_m > 0.95$; Table 2). Biparental inbreeding and correlated paternity were generally low across populations ($t_m - t_s < 0.20$; $r_p < 0.15$), and significantly lower in Yookamurra Sanctuary than Monarto Woodland and Scotia Sanctuary. No significant differences were present in the number of full-sib groups scaled to progeny array size across populations ($k_n = 0.40$ to 0.44). There were only weak correlations among mating system parameters when estimated at the family level ($r^2 < 0.10$), except for between correlated paternity and number of full-sibships scaled to progeny array size ($r^2 < 0.32$), the two measures of multiple paternities.

Seedling establishment was significantly higher at Monarto Woodland and Yookamurra Sanctuary sites than Scotia Sanctuary. There were no significant differences in seedling establishment according to seed provenance and there was no significant interaction between seed provenance and planting site (Generalized linear model: link function = binomial; seed provenance $\chi^2 = 2.24$, $d.f. = 2$, $P = 0.33$; planting site $\chi^2 = 48.98$, $d.f. = 2$, $P < 0.001$; seed provenance × planting site $\chi^2 = 1.95$, $d.f. = 4$, $P = 0.75$; Table 1; Table S3; Table S4).

Across all families ($n = 58$), the number of full-sibships within progeny arrays (k_n) and correlated paternity (r_p) had strong effects on seedling establishment rate (k_n had a positive effect on establishment rate: per cent deviance explained = 16.4%; r_p had a negative effect on establishment rate: per cent deviance explained = 10.3%; ΔAIC_c between these top two models = 3.81; ΔAIC_c to next best model = 5.49; ΔAIC_c to null model = 7.50; Table 3).

Biparental inbreeding had a negative effect on establishment rate, but its effect was much weaker than the number of full-sibships within progeny arrays and correlated paternity ($t_m - t_s$: per cent deviance explained = 9.07%; ΔAIC_c to best fitting model = 5.17). Outcrossing rate did not associate with growth (per cent deviance explained < 1%; ΔAIC_c to best model = 9.65; ranked worse than null model).

Table 1. Genetic variability of *Eucalyptus gracilis* populations at eight microsatellite markers, progeny array size and seedling establishment data.

Group and parameter	Monarto Woodland	Yookamurra Sanctuary	Scotia Sanctuary
Adults			
n	20	20	18
AR	5.31 (0.29)	4.96 (0.31)	4.95 (0.23)
H_E	0.85 (0.04)	0.81 (0.05)	0.83 (0.04)
H_O	0.85 (0.04)	0.86 (0.03)	0.83 (0.03)
F	−0.03 (0.05)	−0.11 (0.05)	−0.04 (0.04)
Progeny			
Progeny array size	14.70 (0.13)	13.44 (0.20)	14.75 (0.14)
$n_{planted\ seedlings}$	294	295	264
$n_{alive\ seedlings}$	244	255	210
$n_{dead\ seedlings}$	50	40	54
Establishment rate (%)	85.02	87.63	80.77
AR	5.17 (0.08)	4.89 (0.08)	4.92 (0.07)
H_E	0.83 (0.05)	0.80 (0.05)	0.82 (0.04)
H_O	0.71 (0.08)	0.72 (0.09)	0.70 (0.07)
F	0.17 (0.08)	0.14 (0.09)	0.16 (0.06)

n, number of samples.
AR, rarefied allelic richness.
H_E and H_O, unbiased expected and observed heterozygosity, respectively.
F, fixation index.
standard deviations in parentheses.
doi:10.1371/journal.pone.0090478.t001

Appendix

Table 2. Mating system parameter estimates for *Eucalyptus gracilis* from each population.

Source population	Density (trees ha^{-1})	t_m	t_m–t_s	r_p	k_n
Monarto Woodland	23.67 (2.29)[a]	0.97 (0.02)[a]	0.15 (0.01)[a]	0.12 (0.02)[a]	0.40 (0.03)[a]
Yookamurra Sanctuary	49.33 (4.63)[b]	0.98 (0.01)[a]	0.11 (0.02)[b]	0.06 (0.01)[b]	0.44 (0.02)[a]
Scotia Sanctuary	20.42 (3.24)[a]	0.95 (0.02)[b]	0.16 (0.03)[a]	0.11 (0.04)[a]	0.43 (0.03)[a]

t_m, outcrossing rate.
t_m–t_s, biparental inbreeding.
r_p, correlated paternity.
k_n, the number of full-sibships within progeny arrays scaled to progeny array size.
standard deviations in parentheses.
95% confidence interval homogeneous subgroups indicated by 'a' and 'b'.
doi:10.1371/journal.pone.0090478.t002

We explored the variance of estimated family-level mating system parameters further because of possible problems of estimating these parameters from a limited progeny array sample size. We observed that most of the upper 50% of estimated mating system parameters had 95% confidence intervals that did not overlap zero, indicating significant levels of these parameters in these families, which also suggests that estimation of these parameters in our study was largely robust to our sample sizes (Figure S1). However, we do recommend that attention should be paid to the potential for high variance of family-level estimates in future studies.

We also explored the leverage and influence of the outlier on the significant regressions (see Figure 2). When the outlier was removed and these regressions were re-run, correlated paternity and the number of full-sibships within progeny arrays were still the best fitting predictors of establishment rate (r_p and k_n per cent deviance explained = 16.6 and 9.3%; Table S5). Additionally, when the original regressions that included the outlier were bootstrapped, the 5 and 95% bootstrapped percentiles of the multiple paternity-establishment rate regression slopes only marginally overlapped zero (Table 2). Thus, we conclude that this outlier had high leverage but had marginal influence on the regressions and was thus retained in our analyses.

Table 3. General linear model comparisons of relationships between genetic predictors and establishment rate (%) of *Eucalyptus gracilis* progeny arrays.

Model	% DE	*w*AIC	ΔAIC_c	*k*	*ß*
Establishment rate ~ k_n	16.37	0.80	0.00	2	−4.09 (−9.74 to 1.52)
Establishment rate ~ r_p	10.26	0.12	3.81	2	30.24 (−41.85 to 100.00)
Establishment rate ~ t_m–t_s	7.42	0.05	5.49	2	
Establishment rate ~ 1	0.00	0.02	7.50	1	
Establishment rate ~ t_m	0.01	0.01	9.65	2	

% DE, per cent deviance explained by model.
*w*AIC, Akaike weight that shows the relative likelihood of model *i*.
ΔAIC_c, indicator of differences between model AIC_c (a measure of model goodness-of-fit scaled to the number of parameters in the model) and minimum AIC_c in the model set.
k, number of parameters in each model.
ß, unstandardized regression slope with 5 and 95% bootstrapped percentiles in parentheses in models that were either the best fitting model or had $\Delta AIC_c < 4$.
t_m, outcrossing rate.
t_m–t_s, biparental inbreeding.
r_p, correlated paternity.
k_n, the number of full-sibships within progeny arrays scaled to progeny array size.
1, null model.
doi:10.1371/journal.pone.0090478.t003

The stand density of Yookamurra Sanctuary was significantly higher than both Monarto Woodland and Scotia Sanctuary (Table 1, 2). Progeny arrays collected from Yookamurra Sanctuary exhibited significantly lower biparental inbreeding and more multiple paternities than both other populations (Table 2), and fits with expectations based on density differences between these populations. Establishment rates also tended to be higher in families from the higher density Yookamurra Sanctuary, but this effect was not significant (see text above; Table S4).

During the sampling period, rainfall was substantially higher than the long-term average at all sites (1.7, 2,1 and 2.7 times the recent past for Monarto Woodland, Yookamurra Sanctuary and Scotia Sanctuary, respectively; Table S6). This suggests that the degree of water stress acting on seedlings was somewhat lower than expected. Since selection against low fitness phenotypes should be weaker during these periods of reduced stress [60], and because it is likely that *E. gracilis* is sensitive to water availability [61], we expect to observe lower seedling mortality than during an average year. Thus, the correlations we derive between mating system parameters and fitness are probably underestimates.

Discussion

We explored whether inbreeding avoidance in monoecious trees constrains inbreeding-fitness correlations at life stages beyond seed development [18], and whether within such systems and life stages, levels of multiple paternities had a greater influence on offspring fitness than inbreeding - the constrained inbreeding hypothesis. Further to maximising fitness in inbreeding-constrained systems, multiple paternities are likely to be positively correlated with the degree of mate discrimination/bet hedging and genetic diversity of offspring, and therefore competition among male gametes and/or female choice for superior male gametes [8,13]. Indeed, in this study we provide evidence to support the constrained inbreeding hypothesis by showing that open-pollinated families of *Eucalyptus gracilis* had little variation in inbreeding (measuring both selfing and biparental inbreeding at the seedling stage), but demonstrated a correlation between the number of multiple paternities and fitness at the seedling life stage.

Levels of multiple paternities in *E. gracilis* seedling families should not be constrained by strong inbreeding avoidance. Therefore, as we predicted, we found that levels of multiple paternities were the strongest predictors of seedling fitness as

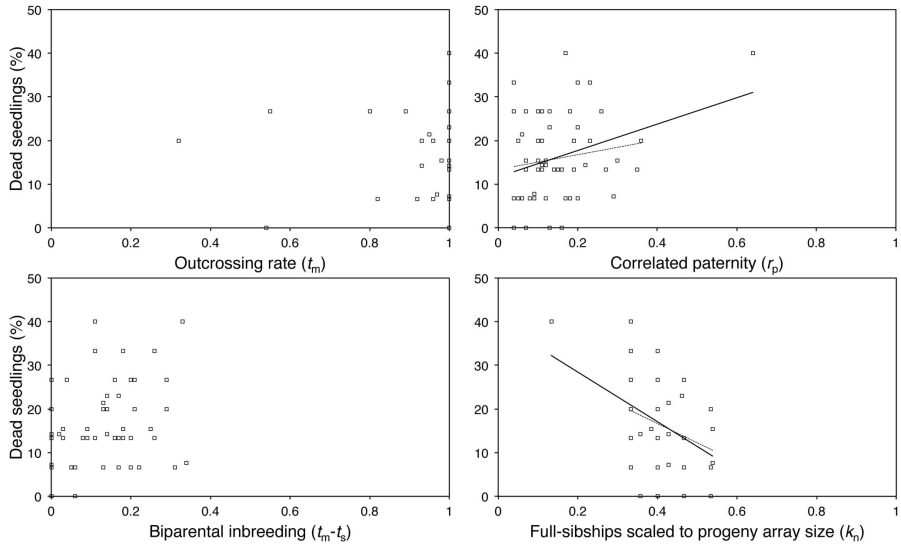

Figure 2. Scatterplots showing relationships between *Eucalyptus gracilis* family-level establishment rates and mating system parameters. Establishment rate percentages per progeny array are shown on the y-axis and mating system parameter values shown on the x-axis. Linear trend lines between genetic parameters and growth shown for relationships where $\Delta AIC_c < 4$ (ΔAIC_c values presented in Table 3). Trend lines are for all data are solid and trend lines are for data without the outlier are dashed.
doi:10.1371/journal.pone.0090478.g002

measured by establishment rates; indeed these effects were over and above the realised effects of inbreeding-related parameters on seedlings (Table 3). Our results are consistent with previous studies that have documented fitness impacts of reduced multiple paternities, of which most of these studies were done in fragmented tree populations [11,16,17,31,32,34].

To make our conclusion, we explored the correlations of four different mating system parameters, estimated from early seedlings for three natural populations of *E. gracilis*, with the fitness of these seedlings measured by monitoring seedling establishment rates. The four different mating system parameters we observed were two commonly used measures of multiple paternities (correlated paternity and the number of full-sibships) and two measures of inbreeding (outcrossing rate and biparental inbreeding). The two multiple paternity measures are estimated independently of inbreeding [11,58]. Including these multiple paternity measures in our study was particularly important since species that undergo strong inbreeding avoidance when measured at the seedling life stage, like many eucalypts [3], are unlikely to express significant levels of inbreeding depression when observed at this life stage ([16,18], but see [15]). Consequently, we suggest that in species with strong inbreeding avoidance, the degree of multiple paternities could be as important to observe as inbreeding levels when investigating intermediate stage fitness consequences of variation in mating system parameters [16]. However, with our design, we can only conclude that this fitness effect is acting on seedling establishment and not earlier (*e.g.* germination) or later life stages (*e.g.* fecundity). Thus, we encourage future studies to explore these multiple paternity effects in natural populations outside of this life stage, with special attention made to maximising the numbers of progeny used per family (to improve mating system parameter estimates), and the number of families used, since our sample size ($n = 58$ families) is low for studies of plant fitness.

It should be noted that stand density might be an important factor underlying much of the variation in observed multiple paternities here. Yookamurra Sanctuary had a stand density significantly higher than both Monarto Woodland and Scotia Sanctuary, which were similar, and accordingly Yookamurra had significantly lower biparental inbreeding and more multiple paternities. However, despite the fact that establishment rates of seedlings from Yookamurra Sanctuary tended to be higher, this effect was not significant. Thus, further investigations are required to identify the effect of population-level ecological characteristics, such as stand density, that may explain our observed patterns. However, consistent with previous studies, our data do generally support positive density-dependent establishment as a function of multiple paternity in mostly outcrossing animal-pollinated trees [16,25].

Supporting Information

Figure S1 Frequency histograms of family-level estimated mating system parameters and 95% confidence intervals.
(DOCX)

Table S1 Genetic variability at each microsatellite locus for *Eucalyptus gracilis* maternal trees.
(DOCX)

Appendix

Table S2 Genetic differentiation of *Eucalyptus gracilis* populations.
(DOCX)

Table S3 Planting sample sizes and seedling establishment information after 16 months of growth at each of the three sites and for each seed provenance.
(DOCX)

Table S4 Generalized linear models of effects of seed provenance and planting site on establishment rate of *Eucalyptus gracilis*.
(DOCX)

Table S5 General linear model comparisons of relationships between genetic predictors and establishment rate of *Eucalyptus gracilis* without the outlier.
(DOCX)

Table S6 Rainfall observations collected from the closest weather stations to the planting sites with data extending >100 years.
(DOCX)

Acknowledgments

The authors would like to thank the Mt Lofty Botanic Gardens staff for assistance rearing seedlings, Matt Hayward and Phil Scully from Australian Wildlife Conservancy, Rob Murphy from Rural Solutions South Australia and the many volunteers for assistance with fieldwork.

Author Contributions

Conceived and designed the experiments: MFB AJL. Performed the experiments: MFB. Analyzed the data: MFB MJC. Contributed reagents/materials/analysis tools: MFB MJC. Wrote the paper: MFB MJC AJL.

References

1. Salzer K, Gugerli F (2012) Reduced fitness at early life stages in peripheral versus core populations of Swiss stone pine (*Pinus cembra*) is not reflected by levels of inbreeding in seed families. Alpine Botany 122: 75–85.
2. Pound LM, Wallwork MAB, Potts BM, Sedgley M (2002) Early ovule development following self- and cross-pollinations in *Eucalyptus globulus* Labill. ssp. *globulus*. Annals of Botany 89: 613–620.
3. Horsley TN, Johnson SD (2007) Is *Eucalyptus* cryptically self-incompatible? Annals of Botany (London) 100: 1373–1378.
4. Hirao AS (2010) Kinship between parents reduces offspring fitness in a natural population of *Rhododendron brachycarpum*. Annals of Botany (London) 105: 637–646.
5. Klekowski EJ (1988) Genetic load and its causes in long-lived plants. Trees-Structure and Function 2: 195–203.
6. Charlesworth D, Morgan MT, Charlesworth B (1990) Inbreeding depression, genetic load, and the evolution of outcrossing rates in a multilocus system with no linkage. Evolution 44: 1469–1489.
7. Crnokrak P, Barrett SCH (2002) Purging the genetic load: a review of the experimental evidence. Evolution 56: 2347–2358.
8. Skogsmyr IO, Lankinen Å (2002) Sexual selection: an evolutionary force in plants? Biological Reviews (Cambridge) 77: 537–562.
9. Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. Annual Review of Ecology, Evolution, and Systematics 37: 187–214.
10. Nason J, Herre E, Hamrick J (1998) The breeding structure of a tropical keystone plant resource. Nature 391: 685–687.
11. Breed MF, Gardner MG, Ottewell K, Navarro C, Lowe A (2012) Shifts in reproductive assurance strategies and inbreeding costs associated with habitat fragmentation in Central American mahogany Ecology Letters 15: 444–452.
12. Llorens TM, Byrne M, Yates CJ, Nistelberger HM, Coates DJ (2011) Evaluating the influence of different aspects of habitat fragmentation on mating patterns and pollen dispersal in the bird-pollinated *Banksia sphaerocarpa* var. *caesia*. Molecular Ecology 21: 314–328.
13. Yasui Y (1998) The 'genetic benefits' of female multiple mating reconsidered. Trends in Ecology & Evolution 13: 246–250.
14. Hufford KM, Hamrick JL (2003) Viability selection at three early life stages of the tropical tree, *Platypodium elegans* (Fabaceae, Papilionoideae). Evolution 57: 518–526.
15. Silva JCE, Hardner C, Tilyard P, Pires AM, Potts BM (2010) Effects of inbreeding on population mean performance and observational variances in Eucalyptus globulus. Annals of Forest Science 67.
16. Breed MF, Marklund MHK, Ottewell KM, Gardner MG, Harris JCB, et al. (2012) Pollen diversity matters: revealing the neglected effect of pollen diversity on fitness in fragmented landscapes. Molecular Ecology 21: 5955–5968.
17. Cascante A, Quesada M, Lobo JJ, Fuchs EA (2002) Effects of dry tropical forest fragmentation on the reproductive success and genetic structure of the tree *Samanea saman*. Conservation Biology 16: 137–147.
18. Szulkin M, Bierne N, David P (2010) Heterozygosity-fitness correlations: a time for reappraisal. Evolution 64: 1202–1217.
19. Petit RJ, Brewer S, Bordacs S, Burg K, Cheddadi R, et al. (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. Forest Ecology and Management 156: 49–74.
20. Breed MF, Stead MG, Ottewell KM, Gardner MG, Lowe AJ (2013) Which provenance and where? Seed sourcing strategies for revegetation in a changing environment. Conservation Genetics 14: 1–10.
21. Broadhurst LM, Lowe A, Coates DJ, Cunningham SA, McDonald M, et al. (2008) Seed supply for broadscale restoration: maximizing evolutionary potential. Evolutionary Applications 1: 587–597.
22. Quesada M, Stoner KE, Lobo JA, Herrerias-Diego Y, Palacios-Guevara C, et al. (2004) Effects of forest fragmentation on pollinator activity and consequences for plant reproductive success and mating patterns in bat-pollinated Bombacaceous trees. Biotropica 36: 131–138.
23. Ghazoul J (2005) Pollen and seed dispersal among dispersed plants. Biological Reviews (Cambridge) 80: 413–443.
24. Bianchi EJJA, Cunningham SA (2012) Unravelling the role of mate density and sex ratio in competition for pollen. Oikos 121: 219–227.
25. Breed MF, Ottewell KM, Gardner MG, Marklund MHK, Dormontt ED, et al. (in press) Mating patterns and pollinator mobility are critical traits in forest fragmentation genetics. Heredity.
26. Lowe AJ, Boshier D, Ward M, Bacles CFE, Navarro C (2005) Genetic resource impacts of habitat loss and degradation; reconciling empirical evidence and predicted theory for neotropical trees. Heredity 95: 255–273.
27. Eckert CG, Kalisz S, Geber MA, Sargent R, Elle E, et al. (2010) Plant mating systems in a changing world. Trends in Ecology & Evolution 25: 35–43.
28. Aldrich PR, Hamrick JL (1998) Reproductive dominance of pasture trees in a fragmented tropical forest mosaic. Science 281: 103–105.
29. Charnov EL (1976) Optimal foraging, the marginal value theorem. Theoretical Population Biology 9: 129–136.
30. Ottewell KM, Donnellan SC, Lowe AJ, Paton DC (2009) Predicting reproductive success of insect- versus bird-pollinated scattered trees in agricultural landscapes. Biological Conservation 142: 888–898.
31. Fuchs E, Lobo J, Quesada M (2003) Effects of forest fragmentation and flowering phenology on the reproductive success and mating patterns of the tropical dry forest tree *Pachira quinata*. Conservation Biology 17: 149–157.
32. González-Varo JP, Albaladejo RG, Aparicio A, Arroyo J (2010) Linking genetic diversity, mating patterns and progeny performance in fragmented populations of a Mediterranean shrub. Journal of Applied Ecology 47: 1242–1252.
33. Hoebee SE, Young AG (2001) Low neighbourhood size and high interpopulation differentiation in the endangered shrub *Grevillea iaspicula* McGill (Proteaceae). Heredity 86: 489–496.
34. Hirayama K, Ishida K, Setsuko S, Tomaru N (2007) Reduced seed production, inbreeding, and pollen shortage in a small population of a threatened tree, *Magnolia stellata*. Biological Conservation 136: 315–323.
35. House SM (1997) Reproductive biology of eucalypts. In: Williams JE, Woinarski JCZ, editors. Eucalypt Ecology: Individuals to Ecosystems. Cambridge: Cambridge University Press. 30–55.
36. Parsons RF (1969) Physiological and ecological tolerances of *Eucalyptus incrassata* and *E. socialis* to edaphic factors. Ecology 50: 386–390.
37. Nicolle D (1997) Eucalypts of South Australia. Adelaide, South Australia: Lane Print Group.
38. Slee A, Brooker M, Duffy S, West J (2006) EUCLID: Eucalyptus of Australia. Canberra: Centre for Plant Biodiversity Research.
39. Morrant D, Petit S, Schumann R (2010) Floral nectar sugar composition and flowering phenology of the food plants used by the western pygmy possum, *Cercartetus concinnus*, at Innes National Park, South Australia. Ecological Research 25: 579–589.
40. House SM (1997) Reproductive Biology of Eucalypts. In: Williams J, Woniarski J, editors. Eucalypt ecology: individuals to ecosystems: Cambridge University Press.
41. Wellington AB, Noble IR (1985) Seed dynamics and factors limiting recruitment of the mallee *Eucalyptus incrassata* in semi-arid, south-eastern Australia. Journal of Ecology 73: 657–666.
42. Wellington AB, Noble IR (1985) Post-fire recruitment and mortality in a population of the mallee *Eucalyptus incrassata* in semi-arid, south-eastern Australia. Journal of Ecology 73: 645–656.
43. Addelman S (1969) The generalized randomized block design. The American Statistician 23: 35–36.

Appendix

44. Faria DA, Mamani EMC, Pappas MR, Pappas GJ, Jr, Grattapaglia D (2010) A selected set of EST-derived microsatellites, polymorphic and transferable across 6 species of *Eucalyptus*. Journal of Heredity 101: 512–520.
45. Clarke MF, Avitabile SC, Brown L, Callister KE, Haslem A, et al. (2010) Ageing mallee eucalypt vegetation after fire: insights for successional trajectories in semi-arid mallee ecosystems. Australian Journal of Botany 58: 363–372.
46. Vranckx GUY, Jacquemyn H, Muys B, Honnay O (2011) Meta-analysis of susceptibility of woody plants to loss of genetic diversity through habitat fragmentation. Conservation Biology 26: 228–237.
47. Bradshaw CJA (2012) Little left to lose: deforestation and forest degradation in Australia since European colonisation. Journal of Plant Ecology 5: 109–120.
48. Oosterhout CV, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4: 535–538.
49. Chybicki IJ, Burczyk J (2009) Simultaneous estimation of null alleles and inbreeding coefficients. Journal of Heredity 100: 106–113.
50. Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6: 288–295.
51. Nei M, Chesser R (1983) Estimation of fixation indexes and gene diversities. Annals of Human Genetics 47: 253–259.
52. Jost LOU (2008) GST and its relatives do not measure differentiation. Molecular Ecology 17: 4015–4026.
53. Meirmans PG, Van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. Molecular Ecology Notes 4: 792–794.
54. Nei M (1973) Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences of the United States of America 70: 3321–3323.
55. Kalinowski ST (2005) HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. Molecular Ecology Notes 5: 187–189.
56. Ritland K (2002) Extensions of models for the estimation of mating systems using n independent loci. Heredity 88: 221–228.
57. Berger-Wolf TY, Sheikh SI, DasGupta B, Ashley MV, Caballero IC, et al. (2007) Reconstructing sibling relationships in wild populations. Bioinformatics 23: 49–56.
58. Ashley MV, Caballero IC, Chaovalitwongse W, Dasgupta B, Govindan P, et al. (2009) KINALYZER, a computer program for reconstructing sibling groups. Molecular Ecology Resources 9: 1127–1131.
59. Burnham KP, Andersen DR (2002) Model selection and multimodel inference. New York: Springer.
60. Cheptou P-O, Donohue K (2010) Environment-dependent inbreeding depression: its ecological and evolutionary significance. New Phytologist 189: 395–407.
61. Parsons R, Rowan J (1968) Edaphic range and cohabitaion of some Mallee Eucalypts in South-Eastern Australia. Australian Journal of Botany 16: 109–116.

A4: Paper submitted to Trends in Ecology and Evolution

Networked bioclimatic transects are powerful observatories of global change

Stefan Caddy-Retalic[1,2,3], Alan N. Andersen[1,4], Michael J. Aspinwall[1,5], Martin F. Breed[1,2], Margaret Byrne[1,6], Matthew J. Christmas[1,2], Ning Dong[7,8], Brad J. Evans[8,9], Damien A. Fordham[1,2], Greg R. Guerin[1,2], Ary A. Hoffmann[1,10], Alice C. Hughes[11], Stephen J. van Leeuwen[1,6], Francesca A. McInerney[3], Suzanne M. Prober[1,12], Maurizio Rossetto[1,13], Paul D. Rymer[1,5], Dorothy A. Steane[1,14,15], Andrew J. Lowe[1,2,16]

[1]Australian Transect Network, Terrestrial Ecosystem Research Network (TERN)
[2]School of Biological Sciences and Environment Institute, University of Adelaide, North Terrace, Adelaide, SA 5005 Australia
[3]Sprigg Geobiology Centre and School of Physical Sciences, University of Adelaide, North Terrace, Adelaide, SA 5005 Australia
[4]CSIRO Land & Water Flagship, Tropical Ecosystems Research Centre, Winnellie, NT, Australia
[5]Hawkesbury Institute for the Environment, University of Western Sydney, NSW, Australia
[6]Science and Conservation Division, Western Australian Department of Parks and Wildlife, Locked Bag 104, Bentley Delivery Centre, WA 6983, Australia
[7]Department of Biological Sciences, Macquarie University, North Ryde NSW 2109, Australia
[8]Ecosystem Modelling and Scaling Infrastructure (eMAST), Terrestrial Ecosystem Research Network (TERN)
[9]Department of Environmental Sciences, Faculty of Agriculture and Environment, University of Sydney, NSW 2006, Australia
[10]School of BioSciences, Bio21 Institute, The University of Melbourne, 30 Flemington Park Road, Parkville, Victoria 3052, Australia
[11]Centre for Integrative Conservation, Xishuangbanna Tropical Botanic Garden, Chinese Academy of Sciences, Menglun, Mengla County, Yunnan 666303, P.R. China
[12]CSIRO Land and Water Flagship, Private Bag 5, Wembley, Western Australia 6913, Australia
[13]National Herbarium of NSW, Royal Botanic Gardens and Domain Trust, Mrs Macquaries Road, Sydney 2000, Australia
[14]School of Biological Sciences, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia
[15]Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Locked Bag 4, Maroochydore, Queensland 4558, Australia
[16]Science, Monitoring and Knowledge Branch, Department of Environment, Water and Natural Resources, Hackney Road, Kent Town, SA 5005, Australia.

Corresponding author: Lowe, A.J. (andrew.lowe@adelaide.edu.au).

Keywords: change detection, community turnover, ecological forecasting, environmental gradients, functional traits, phenotypic plasticity, spatial analogues

Abstract

Using transects to measure how functional traits, genotypes and species vary along climate gradients is a powerful approach to investigate genotypic and phenotypic turnover within species, and the resilience of ecological communities to climate change. This paper highlights examples of such species- and community-level changes. We describe how these levels can be integrated with observations from multiple transects, manipulative experiments, genomics and ecological modelling. This integration will help improve understanding of past and future climate-induced changes, derive generalised insights into biodiversity change and guide conservation practice. A continental or global network of transects will help determine the adaptive drivers and limits of species and ecosystems, and allow more accurate forecasting of ecosystem change.

Large-scale bioclimatic transects

Understanding how ecosystems respond to global change is vital for developing improved management strategies that harness the adaptive potential (see Glossary) of species and resilience of communities. A particular challenge is to scale up knowledge from detailed local studies to understand ecological dynamics at regional scales. Large-scale transects that traverse major climate gradients, and associated turnover in functional traits and communities, are ideal platforms for environmental change research [1, 2]. Building networks of such transects will help underpin the development of generalised models of how climate affects biodiversity at gene, species and community levels.

The systematic use of bioclimatic transects as platforms for studying ecosystem change has its origins in the International Geosphere-Biosphere Program (IGBP), which aimed to investigate the roles of climate and land use as ecosystem drivers in major biomes globally [3, 4]. In the two decades since the IGBP was established, interest in exploring the impacts of global change on species and ecosystems has led to a proliferation of independent studies designed around transects using spatial bioclimatic change as a proxy for temporal climate change [2, 5]. These transects are attractive research platforms because they are powerful and cost-effective (Box 1). Transects explicitly maximise variation in environmental variables for study site stratification, reducing the number of sites (and therefore resources) required to describe variability compared to alternative sampling designs [1].

Bioclimatic transects have been used to examine variation at multiple biological scales, from functional traits and genes within species, to turnover of entire

ecosystems, and provide valuable insights into the relationships between abiotic variables and the adaptive limits of species and communities. Such studies can improve our understanding of the patterns and processes of micro- and macro-evolution, as well as enhance our understanding of the processes that facilitate species persistence and ecosystem resilience, particularly in relation to climate change. As such, bioclimatic transect research has addressed the following fundamental questions:

1. To what degree is trait turnover in a changing climate dictated by underlying genetic potential?
2. What climatic thresholds limit the distribution of species and discrete ecological assemblages?
3. How does climate change lead to change in major biomes?

In this paper, we draw on the experience of members of the Australian Transect Network (ATN; Supplementary Material Box S1), a facility of Australia's Terrestrial Ecosystem Research Network, to describe how networked bioclimatic transect research provides cost-effective insight into ecology and evolutionary biology in the context of global change. Similar to other global networks (e.g. the IGBP Transect Network [6] and the Pacific-Asia Biodiversity Transect Network [7]), the ATN has developed several large-scale plot networks that traverse bioclimatic gradients. The ATN collects species and community-level data to investigate the impacts of, and adaptation to, climate variation. The Australian continent is an ideal location for establishing a successful and globally informative network of bioclimatic transects: it incorporates all of the world's major climate zones and climatic variation occurs in largely continuous gradations. The continent enjoys political and funding stability,

allowing >1000 km transects to be established across jurisdictions (and environments) with comparative ease. Australia's suitability as a proxy for the world's climate zones aside, a globally-representative transect network could be achieved through reinvestment in an international network of transects as envisaged by the IGBP.

Here we use important insights from the ATN to provide an overview of how bioclimatic transect research can help frame key ecological insights into responses to global change. We break down the study of biotic responses into intra-specific (i.e. phenological, functional trait and genetic variation) and inter-specific (i.e. community turnover) variation. We summarise key aspects of transect establishment and design to mitigate possible shortcomings of transect methods, and highlight the opportunities provided by this type of work through genomic and modelling approaches.

Exploring intra-species change

Discounting extinction and migration, populations have three main modes of adapting to climate change: (1) plasticity, which involves non-heritable phenotype alteration, potentially through changes in gene expression, to increase fitness [8]; (2) heritable epigenetics, which improves fitness through the activation and/or deactivation of genes across multiple generations [9]; and (3) evolution, whereby phenotypes can adapt over generations through shifts in genotype to improve fitness under new conditions [10]. Despite biological responses to climate change being well documented in the literature (e.g. phenological change [11, 12]), distinguishing their mechanism is often difficult (e.g. plastic vs. heritable epigenetic changes). A key strength of networked transects is that replicated observations of phenotypic change can be linked to both spatial and temporal climate change, helping to identify the

potential drivers that can then be examined in more detail through manipulative experiments, such as reciprocal transplant trials.

Studies of phenotypic responses of plant species to spatial and temporal climate change have been undertaken along the TRansect for ENvironmental monitoring and Decision making (TREND) in South Australia. The sticky hop bush, *Dodonaea viscosa*, exhibited clinal leaf area variation, narrowing with increasing temperature and decreasing rainfall along the TREND [13]. Leaf narrowing in plants lowers surface area (reducing transpiration and limiting radiation loads), potentially making these plants more resilient to aridification. A subsequent analysis of historical herbarium specimens revealed a 40% decrease in leaf width over the last 127 years, with most change occurring since 1950 [14]. An analysis of the flowering times of wallflower orchid, *Diuris orientis*, from herbarium records over the last 100 years identified a shift towards earlier flowering. This was thought to be a response to avoid flowering during increasingly arid summers, as seen with recent climate shifts across its natural range [12]. A similar phenological change was observed in the field along an altitudinal transect of natural populations, indicating that ongoing phenological shifts are expected for this species [12]. These results are consistent with an adaptive response to climate change.

Transects provide a scientifically robust and comparatively cheap platform for conducting experiments such as reciprocal transplant trials, which can be used as a less expensive alternative to experiments in controlled climate facilities. For example, experimental transplants can be incorporated into studies of phenotypic change along bioclimatic gradients, allowing differentiation of plastic and genetic adaptive changes

[e.g. 15, 16]. Indeed, combining growth experiments with genetic data collected along gradients to reveal associations between phenotypic and genetic variation with climate is a major focus of many transect research programs. This approach has been used to study the New South Wales waratah, *Telopea speciosissima,* and red ironbark, *Eucalyptus tricarpa,* along the Biodiversity and Adaptation Transect Sydney (BATS) and Victorian *Eucalyptus* Adaptation Transect (VEAT) respectively [16-18]. Local adaptation in functional traits was demonstrated for *E. tricarpa* using common gardens at each end of the VEAT aridity gradient [16, 18]. Some traits displayed complex combinations of plasticity and genetic divergence along the VEAT, and for several traits there was complex clinal genetic variation in plasticity itself [16]. A combination of adaptive genetic and plastic responses was also suggested in studies of york gum, *E. loxophleba*, and gimlet, *E. salubris*, on the South-West Australian Transitional Transect (SWATT) [19, 20]. Similarly, studies of *T. speciosissima* along the BATS revealed genetic differentiation of coastal and upland genotypes, with substantial mixing at mid-elevations [17]. Germination trials showed significant interactions between genotype and germination temperature in growth cabinets and field conditions, where coastal and upland genotypes showed highest germination rates at 30°C and 10°C respectively, suggesting differential selection by optimal germination temperatures in these ecotypes [17].

Advances in observing micro-evolutionary processes of climate adaptation have been made through study of fruit flies (*Drosophila*) along the East Australian Drosophila Transect (EADrosT) [21, 22]. Genetic differentiation among populations has been demonstrated in numerous traits by culturing flies under uniform conditions for multiple generations. Clear differentiation has also been demonstrated in

chromosome inversions, specific genes, transposable elements and maternally inherited bacteria [21-23]. Many of these genetic changes have been shown to be adaptive. For example, cold temperatures were shown to be an agent of selection on body size and winter egg retention. Geographic patterns in genetic changes have been associated with climate adaptation. Indeed, shifts in gene and inversion clines through time have provided some of the first evidence of adaptive evolution under contemporary climate change [24].

Understanding change in ecological communities

When confronted with climate change, species are forced to adapt, migrate or die [25].

Studies of adaptation to climate change by species (as described above) indicate that some species are being pushed towards and potentially beyond their capacity to adapt. When species are tipped beyond their capacity to adapt, localised changes in species assemblages will occur (Figure 1). A combination of transects and experimental studies can also disentangle the relative effects of physiological thresholds and competition in determining the limits of a species range.

Analysis of woody plants along the Northern Australian Tropical Transect (NATT) revealed a systematic decline in woody vegetation species richness with declining rainfall [26]. In contrast, ant species richness is highly resilient to changes in rainfall, remaining uniformly high across the NATT [27]. However, ant species composition shows marked disjunctions between arid and monsoonal zones in the south and between the semi-arid and mesic zones in the north. Similarly, plant species and family turnover has been observed along the TREND aridity gradient (Figure 1).

Families characteristic of mesic environments (e.g. Cyperaceae and Xanthorrhoeaceae) dominate communities at the temperate end, giving way to a greater prevalence of arid-adapted families (e.g. Amaranthaceae and Solanaceae) at the drier end. Particularly rapid species turnover occurs in the range of 15-16°C in mean annual temperature and mean annual rainfall of 400-600 mm [28].

The ability of transect studies to identify community tipping points, where rapid, non-linear ecological change occurs (Figure 1) is especially powerful in the context of future climate change. Examples of such tipping points are the arid-monsoonal ecotone on the NATT [27], the temperature/rainfall threshold delineating mesic and arid vegetation on the TREND [28], and the abrupt transition from mesic eucalypt woodlands to arid acacia woodlands on the SWATT [29]. Identifying such tipping points is important for predicting ecosystem change and planning management responses, allowing conservation action to be focused on climate-sensitive biological communities and regions (Figure 2).

Transect studies are also particularly useful for investigating the interactions between environmental drivers and land use. For example, plant composition at intermittently grazed sites on the Box-gum East-West Transect (BoxEW) showed greater affinity with the dry end of the gradient than with ungrazed sites. Characteristic taxa from drier woodlands (e.g. grasses, annual forbs, succulents) become more prominent in mesic woodlands with grazing. Conversely, mesic grasses and some perennial forbs that occurred along the whole gradient in ungrazed sites were rare in drier woodlands with livestock grazing [30]. The interaction of community composition and land use history demonstrates the potential for rapid and

extensive shifts in plant composition associated with the aridification effects of grazing [31].

Maximising value from a transect network

The benefit of individual transect studies has been widely extolled in the scientific literature.

However, deriving causation from analyses of single transects is fraught. Covariation of many variables (e.g. temperature, rainfall, soil and land use) [32] can make it difficult to interpret patterns across single transects, even when manipulative studies can be undertaken. Additionally, confounding impacts (such as fire or grazing) may occur on a single transect that could be mistaken for a climate-only signal. However, a network of transects (such as the ATN; Box 1; Figure 3) helps ameliorate these limitations, as it enables comparison of occurrence, and variation in genes and traits between species and communities on independent transects. For example, there are consistent patterns across numerous taxa of genotypic and phenotypic variation with climate along the EADrosT, TREND, BATS and SWATT. These consistent patterns suggest generalised responses.

Interpretations of patterns of adaptive change would be further facilitated by replicating transects along analogous environmental gradients [33]. The use of multiple taxa and/or replicated transects can help identify whether many genes with small effect, or a few genes with large effect, provide the basis of adaptive evolution. If, for example, the same genes are associated with adaptation in multiple functional traits across species (and transects), this may indicate that there are only a few genetic solutions available to cope with climate change [34]. Conversely, if many genes or

combinations of genes are shown to be adaptive across replicated gradients, there may be substantial flexibility in genetic (or epigenetic) responses. Similar investigations of community attributes (such as species diversity) are likely to improve prognostic understanding of community-level change. While there has been extensive discussion of the theoretical expectations of the predictability of evolution [e.g. 35], well-designed transect-based studies will help resolve this question.

While multiple species and replicated transects can provide a picture of the extent to which traits, species and communities can vary along climate gradients, they are also potentially affected by evolutionary and ecological processes that are disconnected from adaptive processes. For example, habitat fragmentation may limit gene flow and therefore the spread of adaptive genes across a landscape [36]. Differences between populations might then be interpreted as representing adaptation, whereas in reality they may simply reflect neutral divergence that happens to match an abiotic gradient in a continuous manner [37]. To avoid this problem, particularly in fragmented landscapes, studies should ideally integrate multiple gradients, such as the elevational/latitudinal sampling undertaken on the TREND [13] and EADrosT [38] (Figure 3). This approach can disentangle the relative contribution of neutral (e.g. migration – isolation by distance) and adaptive (e.g. selection – isolation by environment) processes to avoid interpreting divergence due to isolation as adaptation [39].

***Transect*omics**

New genomics tools are rapidly being developing and applied to understand climate adaptation [40]. Recent applications in Australian transect research include

exploring variation in genome-wide Diversity Arrays Technology (DArT) markers to understand local adaptation in eucalypts [18] and the nature of genetic changes within chromosomal inversions in *Drosophila* [22]. Genomic and transcriptomic approaches can also be used to test the importance of epigenetic changes and other modes of gene regulation in natural systems under climate change, which is still not yet well understood [41], but is likely to be significant [42]. For example, epigenetic changes have been implicated in drought responses in plants [43]. Transcriptomic studies also indicate that gene regulation is expected to have a considerable role in phenotypic plasticity and acclimation, and therefore is a likely target of selection [44]. Despite this potential, techniques for investigating epigenetic changes are still expensive when investigating non-model systems [45]. Moreover, when establishing causal associations between molecular changes and adaptations along transects, manipulative experiments are required. Relevant experiments would entail rearing organisms across multiple generations in common conditions to identify the importance of epigenetic effects, and transplants to isolate immediate environmental effects. Nevertheless, deploying these technologies from a transect platform will help identify local adaptation in functional traits through multi-generation common garden experiments, providing the basis for unravelling the role of gene regulation in climate adaptation.

Next generation ecological models

Recent advances in forecasting the range dynamics and distributions of species have focused on integrating physiological tolerance, adaptive potential, dispersal, metapopulation dynamics and species interactions [46-48]. Transect sampling remains the most efficient way to capture environmentally-driven variation across the ranges of species and communities [49]. Transect networks with wide spatial coverage of

bioclimatic space and temporal replication can therefore provide the detailed life-history data required to parameterise, validate and refine increasingly ecologically realistic models. Ecophysiological and genetic data collected across transect networks can further strengthen model predictions and can provide a unique source of validation data [50, 51; Figure 2]. For example, information on adaptation to climate and environmental variability can be used to modify vital rates in climate-biodiversity models, improving the reliability of ecological predictions and giving a better understanding of eco-evolutionary dynamics [52]. Re-sampling transect networks provides opportunities to quantify how species occurrence, abundance and demographic traits vary temporally as well as spatially. Integrating this information into ecological models is important because the range dynamics of species are sensitive to assumptions regarding inter-annual climate variability [53]. Building ecological models using transect network data is, therefore, likely to result in models that more accurately and explicitly reflect species' ecology and likely responses to changing conditions in both space and time.

Concluding remarks

We argue that a network of bioclimatic transects representative of multiple biomes is a powerful and efficient framework for structuring observational and manipulative studies to understand climate change impacts and adaptation. Bioclimatic transects permit a rapid exploration of links between genetic and phenotypic traits, and provide robust forecasts of how shifting climates affect biodiversity and ecosystem function. This understanding can be used to screen for genotypes that may be more resilient to future climates, to determine the benefits of assisted gene migration for key species (e.g. seed sourcing for restoration programs [18, 20, 54]). In addition, an

understanding of the trajectory and direction of change and the incorporation of physiological studies will strengthen the next generation of predictive models of species distributions under future climates [50].

Ultimately, information gleaned from transect studies will improve our knowledge of drivers of community turnover and help forecast species and ecosystem responses to climate change [51]. Integration of intraspecific adaptation with community level change will provide a unified model for understanding contemporary biodiversity shifts. Applying future climate projections to this model will strengthen predictions of future biodiversity impacts of climate change. Robust predictions will improve ecological outcomes by enabling managers and scientists to focus their attention and resources on protecting the most at-risk (or realistically conservable) taxa, assemblages and regions.

Acknowledgements

The ATN and eMAST are facilities of the Terrestrial Ecosystem Research Network (TERN) and supported by the Australian Government through the National Collaborative Research Infrastructure Strategy. Much of the work described here would not have been possible without long-term Australian Government funding provided through the TERN and other agencies. We thank the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for the Coupled Model Intercomparison Project, and the climate modelling groups (listed in Supplementary Material, Table 2) for their model output used in Figure 3. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provided coordinating support and led development of software

infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

1 De Frenne, P., *et al.* (2013) Latitudinal gradients as natural laboratories to infer species' responses to temperature. *Journal of Ecology* 101, 784-795
2 Parker, V.T., *et al.* (2011) Efficiency in assessment and monitoring methods: scaling down gradient-directed transects. *Ecosphere* 2, art99
3 Austin, M. and Heyligers, P. (1991) New approach to vegetation survey design: gradsect sampling. *Nature conservation: cost effective biological surveys and data analysis*, 31-36
4 Koch, G.W., *et al.* (1995) Terrestrial transects for global change research. *Vegetatio* 121, 53-65
5 Blois, J.L., *et al.* (2013) Space can substitute for time in predicting climate-change effects on biodiversity. *Proceedings of the National Academy of Sciences* 110, 9374-9379
6 Canadell, J.G., *et al.* (2002) IGBP/GCTE terrestrial transects: Dynamics of terrestrial ecosystems under environmental change – Introduction. *Journal of Vegetation Science* 13, 298-300
7 Mueller-Dombois, D. and Daehler, C.C. (2005) The PABITRA project: island landscapes under global change 1. *Pacific Science* 59, 133-139
8 Anderson, J.T. and Gezon, Z.J. (2015) Plasticity in functional traits in the context of climate change: a case study of the subalpine forb *Boechera stricta* (Brassicaceae). *Global Change Biology* 21, 1689-1703
9 Heard, E. and Martienssen, Robert A. (2014) Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* 157, 95-109
10 Pauls, S.U., *et al.* (2013) The impact of global climate change on genetic diversity within populations and species. *Molecular Ecology* 22, 925-946
11 Calinger, K.M., *et al.* (2013) Herbarium specimens reveal the footprint of climate change on flowering trends across north-central North America. *Ecology Letters* 16, 1037-1044
12 MacGillivray, F., *et al.* (2010) *Herbarium collections and photographic images: alternative data sources for phenological research. In Phenological Research: 425-461*. Springer
13 Guerin, G.R., *et al.* (2012) Leaf morphology shift linked to climate change. *Biology Letters* 8, 882-886
14 Guerin, G.R. and Lowe, A.J. (2013) Leaf morphology shift: new data and analysis support climate link. *Biology Letters* 9, 1-3
15 Grady, K.C., *et al.* (2013) Conservative leaf economic traits correlate with fast growth of genotypes of a foundation riparian species near the thermal maximum extent of its geographic range. *Functional Ecology* 27, 428-438
16 McLean, E.H., *et al.* (2014) Plasticity of functional traits varies clinally along a rainfall gradient in *Eucalyptus tricarpa*. *Plant, Cell & Environment* 37, 1440-1451
17 Rossetto, M., *et al.* (2011) The impact of distance and a shifting temperature gradient on genetic connectivity across a heterogeneous landscape. *BMC Evolutionary Biology* 11, 126
18 Steane, D.A., *et al.* (2014) Genome-wide scans detect adaptation to aridity in a widespread forest tree species. *Molecular Ecology* 23, 2500-2513
19 Steane, D.A., *et al.* (2015) Genome-wide scans reveal cryptic population structure in a dry-adapted eucalypt. *Tree Genetics & Genomes* 11, 1-14
20 Prober, S.M., *et al.* (2015) Climate-adjusted provenancing: a strategy for climate-resilient ecological restoration. *Frontiers in Ecology and Evolution* 3, 65

21 Hoffmann, A.A. and Weeks, A.R. (2007) Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in *Drosophila melanogaster* from eastern Australia. *Genetica* 129, 133-147
22 Rane, R.V.*, et al.* (2015) Genomic evidence for role of inversion 3RP of *Drosophila melanogaster* in facilitating climate change adaptation. *Molecular Ecology*, DOI: 10.1111/mec.13161
23 Levine, M.T.*, et al.* (2011) Whole-genome expression plasticity across tropical and temperate *Drosophila melanogaster* populations from eastern Australia. *Molecular Biology and Evolution* 28, 249-256
24 Umina, P.*, et al.* (2005) A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. *Science* 308, 691-693
25 Christmas, M.J.*, et al.* (in press) Constraints to and conservation implications for climate change adaptation in plants. *Conservation Genetics*
26 Bowman, D. (1996) Diversity patterns of woody species on a latitudinal transect from the monsoon tropics to desert in the Northern Territory, Australia. *Australian Journal of Botany* 44, 571-580
27 Andersen, A.N.*, et al.* (in press) Savanna ant species richness is maintained along a bioclimatic gradient of decreasing rainfall and increasing latitude in northern Australia. *Journal of Biogeography*
28 Guerin, G.R.*, et al.* (2013) Spatial modelling of species turnover identifies climate ecotones, climate change tipping points and vulnerable taxonomic groups. *Ecography* 36, 1086-1096
29 Butt, C.*, et al.* (1977) Uranium occurences in calcrete and associated sediments in Western Australia. pp. 67, CSIRO Division of Mineralogy
30 Prober, S.M. and Thiele, K. (2004) Floristic patterns along an east-west gradient in grassy box woodlands of central New South Wales. *Cunninghamia* 8, 306-325
31 Prober, S.M.*, et al.* (2014) Towards climate-resilient restoration in mesic eucalypt woodlands: characterizing topsoil biophysical condition in different degradation states. *Plant and Soil* 383, 231-244
32 Meirmans, P.G. (2015) Seven common mistakes in population genetics and how to avoid them. *Molecular Ecology* 24, 3223-3231
33 Savolainen, O.*, et al.* (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics* 14, 807-820
34 Bell, M.A. and Aguirre, W.E. (2013) Contemporary evolution, allelic recycling, and adaptive radiation of the threespine stickleback. *Evolutionary Ecology Research* 15, 377-411
35 Rockman, M.V. (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* 66, 1-17
36 Breed, M.F.*, et al.* (2011) Clarifying climate change adaptation responses for scattered trees in modified landscapes. *Journal of Applied Ecology* 48, 637-641
37 Warren, D.L.*, et al.* (2014) Mistaking geography for biology: inferring processes from species distributions. *Trends in Ecology & Evolution* 29, 572-580
38 Klepsatel, P.*, et al.* (2014) Similarities and differences in altitudinal versus latitudinal variation for morphological traits in *Drosophila melanogaster*. *Evolution* 68, 1385-1398
39 Sexton, J.P.*, et al.* (2014) Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution* 68, 1-15
40 Fitzpatrick, M.C. and Keller, S.R. (2015) Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters* 18, 1-16
41 Franks, S.J. and Hoffmann, A.A. (2012) Genetics of climate change adaptation. *Annual Review of Genetics* 46, 185-208
42 Prentis, P.J.*, et al.* (2008) Adaptive evolution in invasive species. *Trends in Plant Science* 13, 288-294

43 Rico, L.*, et al.* (2014) Changes in DNA methylation fingerprint of *Quercus ilex* trees in response to experimental field drought simulating projected climate change. *Plant Biology* 16, 419-427
44 Chen, Y.*, et al.* (2012) Genome-wide transcription analysis of clinal genetic variation in *Drosophila*. *PLoS One* 7, e34620
45 Shafer, A.B.*, et al.* (2014) Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution* 30, 78-87
46 Kearney, M.*, et al.* (2009) Integrating biophysical models and evolutionary theory to predict climatic impacts on species' ranges: the dengue mosquito *Aedes aegypti* in Australia. *Functional Ecology* 23, 528-538
47 Fordham, D.A.*, et al.* (2013) Tools for integrating range change, extinction risk and climate change information into conservation management. *Ecography* 36, 956-964
48 Fordham, D.A.*, et al.* (2013) Adapted conservation measures are required to save the Iberian lynx in a changing climate. *Nature Climate Change* 3, 899-903
49 Gillison, A.N. and Brewer, K.R.W. (1985) The use of gradient directed transects or gradsects in natural resource surveys. *Journal of Environmental Management* 20, 103-127
50 Fordham, D.A.*, et al.* (2014) Better forecasts of range dynamics using genetic data. *Trends in Ecology & Evolution* 29, 436-443
51 Wisz, M.S.*, et al.* (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews* 88, 15-30
52 Thuiller, W.*, et al.* (2013) A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology Letters* 16, 94-105
53 Bateman, B.L.*, et al.* (2012) Nice weather for bettongs: using weather events, not climate means, in species distribution models. *Ecography* 35, 306-314
54 Breed, M.F.*, et al.* (2013) Which provenance and where? Seed sourcing strategies for revegetation in a changing environment. *Conservation Genetics* 14, 1-10
55 White, A.*, et al.* (2012) *AusPlots Rangelands Survey Protocols Manual, Version 1.2.9*. University of Adelaide Press
56 Smartt, P. and Grainger, J. (1974) Sampling for vegetation survey: some aspects of the behaviour of unrestricted, restricted, and stratified techniques. *Journal of Biogeography* 1, 193-206
57 Austin, M.P. (1985) Continuum concept, ordination methods, and niche theory. *Annual Review of Ecology and Systematics* 16, 39-61
58 Hutchinson, M.*, et al.* (2014) Monthly daily maximum temperature: ANUClimate 1.0, 0.01 degree, Australian Coverage, 1970-2012. (Australian National University, C.A., ed)
59 Thomson, A.*, et al.* (2011) RCP4.5: a pathway for stabilization of radiative forcing by 2100. *Climatic Change* 109, 77-94
60 Riahi, K.*, et al.* (2011) RCP 8.5 - A scenario of comparatively high greenhouse gas emissions. *Climatic Change* 109, 33-57
61 CSIRO and Bureau of Meteorology (2015) Climate change in Australia: information for Australia's natural resource management regions: Technical Report.

Box 1. Defining Transects

The term transect is used in a broad sense to mean a path (usually linear) through an area along which data are collected. Measurements may include species presence and abundance (e.g. for biodiversity surveys), phenotypic traits, genetic sampling (e.g. for assessing population structure), and environmental variables. Transects can be utilised at varying scales. Transects spanning metres are used as a survey method for measuring vegetation structure within a plot [e.g. 55]. Transects spanning hundreds of kilometres (and major environmental change) are more commonly used to assess community composition and adaptive changes along environmental gradients on a large-scale (Figure i, and the focus of this article).

Figure i: Environmental change across three subcontinental transects.

Positioning a transect to follow a significant environmental gradient was proposed by Gillison and Brewer [49] as the most efficient method to capture habitat heterogeneity and maximise species detection in biodiversity surveys. This approach differed from traditional survey methods based on random, systematic or simple

stratified sampling [56]. Systematic sampling is resource intensive, and Gillison and Brewer criticised randomised sampling as potentially counter-productive, as species' distributions are rarely random. Instead, they proposed that greatest biodiversity would be found in line with the most significant environmental gradient or gradients within a study area, in a non-random distribution. They termed these gradient-orientated transects 'gradsects', which have remained a popular survey methodology. [e.g. 2, 3].

Large-scale (subcontinental) transects follow some gradsect principles. They are placed along a major environmental (often climatic) gradient; site selection is based on logistical considerations (e.g. accessibility); and they follow sound experimental design with opportunity for replication and randomisation within a transect. However, where gradsects were designed as a biodiversity survey tool,the goals of bioclimatic gradient studies are typically to assess composition, turnover, connectivity, and adaptation of species and communities, and to interpret these results in the context of the gradient. Such assessments can also help predict future outcomes of climate change through 'space-for-time' substitutions [5].

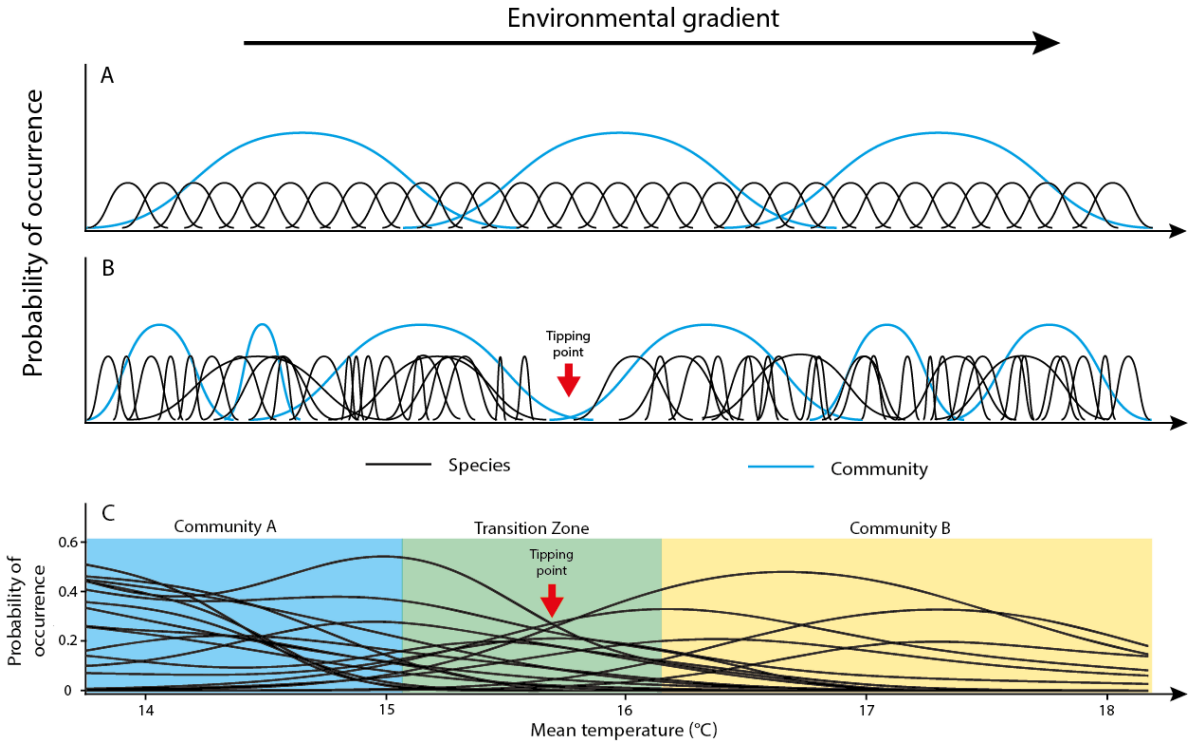

Figure 1. Turnover in species and communities on a hypothetical bioclimatic transect (A, B) and occurrence data from the TREND (C). If all species and communities have the same niche width and sensitivity, on an even gradient, regular species turnover can be expected (A). However, landscapes are likely to have a mix of generalist and specialist species with differing tolerances, adaptive potential or niche widths, potentially displaying an uneven response between taxonomic and/or functional groups (B). Red arrows indicate a non-linear ecological disjunction or "tipping point". (C) shows non-parametric distribution models for 19 common species on the TREND based on surveys of 3,567 field plots by the Biological Survey of South Australia. TREND data provided by the South Australian Department of Environment, Water and Natural Resources, accessed 20 August 2010 [28]. Conceptual diagrams after Austin [57].

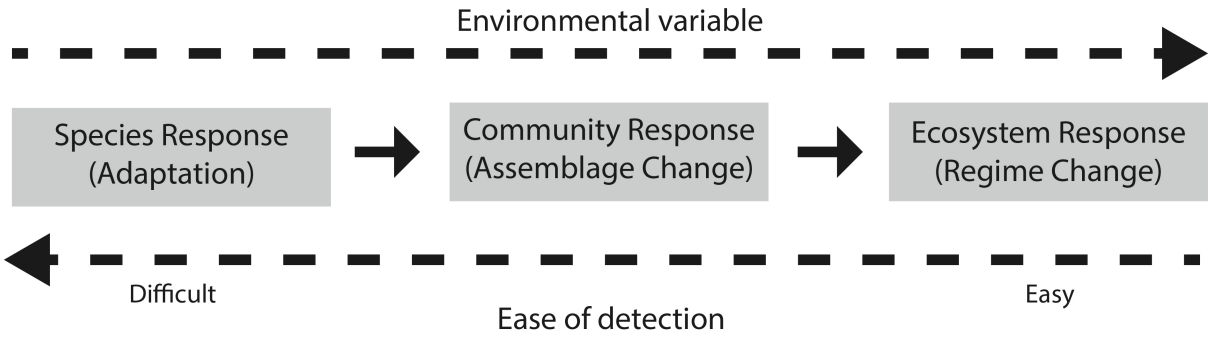

Figure 2. Schematic representation of the hierarchy of ecological change along an environmental gradient. Change progresses from sensitive (but difficult to detect) intraspecific changes in genes or traits (i.e. adaptation), through changes in species assemblage, generally requiring intensive field surveys; to profound (but more readily detected) biome level responses that can be detected using rapid surveys or remote sensing.

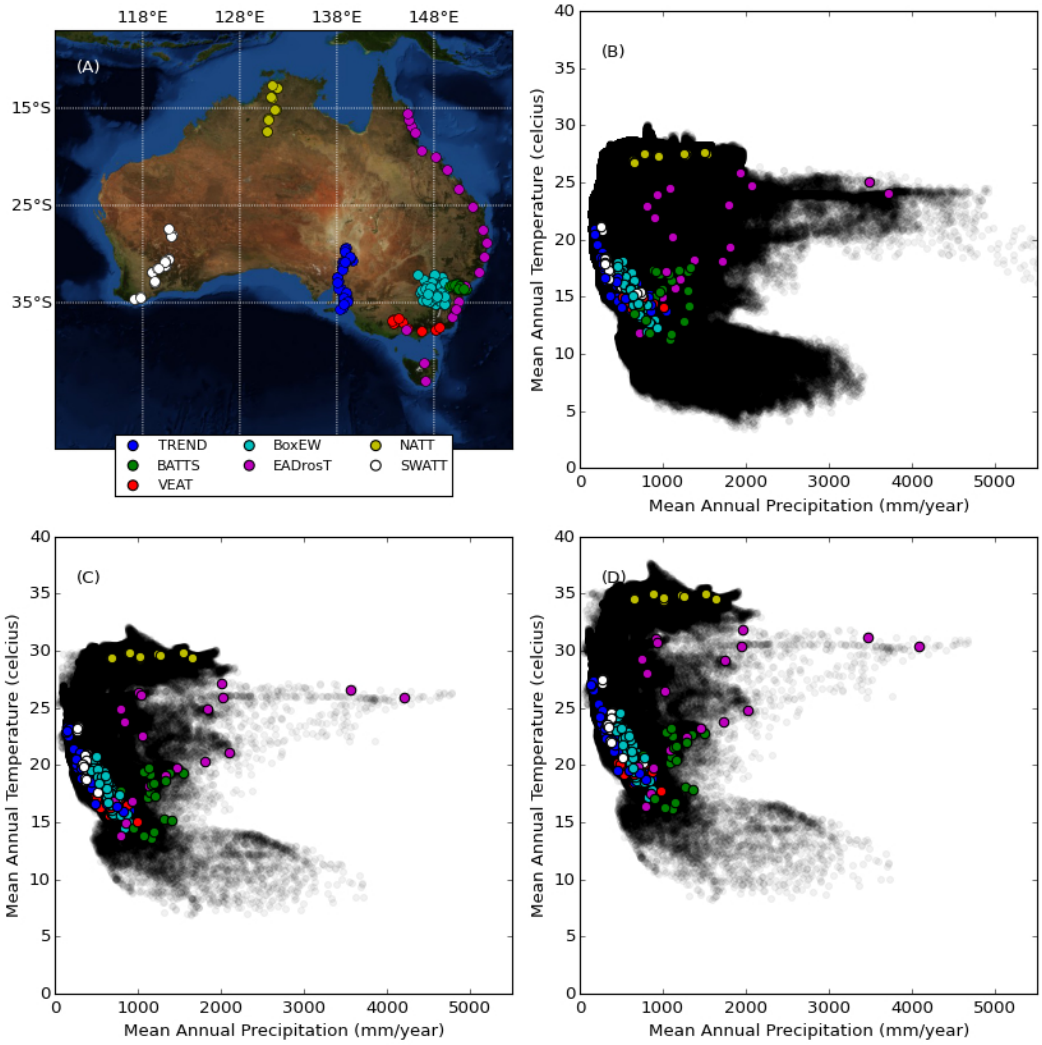

Figure 3. Spatial (A) and bioclimatic (B-D) context of ATN sites against recent (1970-2005) and projected (2006-2050) climate space. (B) Recent (1970-2005) ANUClimate v 1.0, 0.01 degree climate data [58] mean annual temperature and mean annual precipitation for each site, and all of Australia (grey circles). (C) 2006-2050 ensemble mean of seven global climate models for the RCP4.5 scenario (stabilisation of ~650ppm atmospheric CO_2 equivalent [59]). (D) 2006-2050 ensemble mean of seven global climate models for the RCP8.5 scenario (comparatively high greenhouse emissions [60]). Models selected to be consistent with current Australian Government climate modelling [61]. Refer to Supplementary Material Table S2 for details of climate models.

Glossary Box

Adaptation: A heritable change in genotype and/or gene expression in response to environmental change that improves a population's chances of persistence.

Adaptive potential: The capacity of a population, species, community or other biological system to undergo adaptation. Adaptive potential is both facilitated and limited by the levels of standing genetic variation, gene flow, *de novo* mutation and the inherent plasticity associated with a genotype.

Bioclimatic gradient: A continuous change in one or more climatic variable(s) with associated change in biodiversity. For example: a mesic woodland transitioning to an arid grassland.

Biome: A category of ecosystem determined by the structure of the dominant vegetation, such as savanna or tundra. Biomes may comprise a number of constituent ecological communities.

(Ecological) community: An assemblage of organisms that co-occur and interact in a steady state.

Ecological space: An *n*-dimensional hypervolume, where *n* represents every variable required for a species' persistence (e.g. sunlight, winter rainfall, food availability, etc.).

Epigenetic change: Gene expression moderated by one or more factors external to the gene – such as DNA methylation – that does not alter the gene sequence.

Functional group: A collection of organisms with shared traits, e.g. growth form or climatic requirements.

Functional trait: A trait that is indicative of an organism's interaction with its environment. Functional traits are often governed by balancing fitness trade-offs in biochemistry and/or physiology. For example, wood-density is a functional trait of trees that balances growth rate with durability.

Niche: The ecological space in which a species can persist. Generalist species occupy wide niches and are capable of persisting across most (or all) of a climate gradient and may, therefore, display greater adaptive potential. Specialist species occupy narrow niches and may be less likely to persist if environmental conditions change.

Non-linear change: Change occurring on a gradient associated with one or more tipping points. Non-linear change may be difficult to model or predict and may lead to transformative change within ecosystems.

Phenotypic plasticity: The potential of a genotype to produce variation in phenotype. Variation involves changes in one or more functional trait(s) without changes in gene frequency. Plastic responses may be temporary or permanent for an organism's lifespan. Genotypes vary in their plasticity; and evolution and plastic responses can occur in tandem. Examples include learning or non-heritable changes in gene expression. The mechanisms underlying phenotypic plasticity are not well understood but are likely to involve (epigenetic) changes in gene expression in many cases.

Tipping point: The point (in geographic or climate space) at which continuous change in a single environmental factor, or coalescence of multiple factors, reaches a threshold prompting a major ecological disjunction (e.g. a transition from one biome to another).

Tolerance: The ability of an individual, genotype, species, community or biome to persist in the face of extrinsic change.

Guerin, G.R., Martín-Forés, I., Biffin, E., Baruch, Z., Breed, M.F., Christmas, M.J., Cross, H.B. and Lowe, A.J. (2014). Global change community ecology beyond species-sorting: a quantitative framework based on mediterranean-biome examples. *Global Ecology and Biogeography*, 23(10), 1062-1072.

A6. Down but not out

This is an X-ray of what my left humerus looked like 6 weeks prior to submitting this thesis. I include it here as an eternal reminder to myself of the challenges I have faced in the past 3-½ years of my postgraduate studies. It is a reminder that, despite the setbacks (of which, the fractured humerus was probably the biggest and definitely the most painful), I can achieve what I set out for if I remain determined and committed. This thesis is a culmination of my achievements as a scientist thus far, and not even a broken arm could prevent me from achieving it. It also serves as a symbol of humility; I may be capable of completing a doctorate degree but I am also capable of some very stupid things.

"Nothing is easier than to admit in words the truth of the universal struggle for life, or more difficult--at least I have found it so--than constantly to bear this conclusion in mind."

Darwin, 1859