PROJECT TITLE

Genomic signatures of adaptive diversification in woodland Eucalyptus

AIMS AND BACKGROUND

An emerging pattern in speciation research is the importance of ancient standing variants and introgressed alleles, rather than new mutations, in recent divergence between sibling or incipient species [1, 2]. This is consistent with our understanding that, when available, both standing genetic variation and hybridisation between diverged lineages can promote rapid adaptation to new environments [3, 4].The source of the raw material for evolution (Figure 1) can have **significant impacts on the speed with which populations can adapt** and on the genomic signatures left by selection. These modes of adaptation can be identified by combining gene trees, species trees and geographic information [5]. There is growing evidence that interspecific gene flow has played an important role in at least one large vertebrate radiation [6], but its relative importance in plants is a major gap in our understanding of speciation.

***Eucalyptus* offers an unique opportunity to investigate the sources of allelic variation underlying adaptive radiation.** It is a species-rich (>700 species) genus, primarily consisting of essential foundation species on which entire ecosystems of interacting organisms depend. Many of its species co-occur over broad areas, creating replicated hybrid zones and the opportunity to investigate the relationship between divergence between species and local adaptation. Eucalypts are the dominant tree species in much of Australia and many have extremely large population sizes, which favours the persistence of standing genetic variation. One of the major outstanding questions in eucalypt biology is the process/es responsible for the large number of species in the genus, as well as the role of hybridisation in adaptation and divergence.

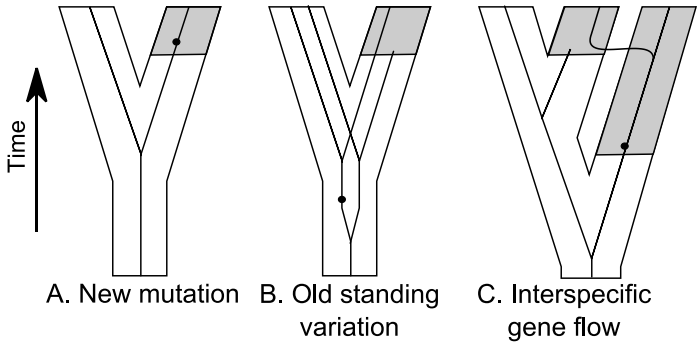

Figure 1. Three sources of adaptive genetic variation, shown as gene trees superimposed on the species tree, with the new environment shaded. A) A new mutation in the adapting lineage. B) The mutation occurs prior to speciation. C) Gene migration introduces the derived allele to the adapting lineage.

This project will apply the framework described above to understand the sources of adaptive variation in woodland *Eucalyptus*. The outcomes will include the identification of genes under selection during speciation and local adaptation to environmental variation. **Improving our understanding of adaptation and genetic variation in woodland eucalypts will make a significant contribution to their conservation, management and restoration.** We will address the following specific questions:

1. **How independent are the gene pools of woodland eucalypt species, and what is the contribution of ancient standing variants and introgressed alleles to speciation?** Using whole-genome resequencing and windowed analyses, we will determine the extent of allele sharing and identify regions of the genome that have undergone divergent selection between species. The molecular divergence and taxonomic spread of related alleles will allow us to determine the ages of such regions and thereby distinguish between the two mechanisms.
2. **Is new, introgressed or standing genetic variation involved in local adaptation within species?** We will use associations between alleles and environmental variables to identify genetic variants involved in local adaptation in two species. An optimised low-coverage whole-genome sequencing strategy will identify causal mutations with genomic resolution that is unprecedented in landscape genomics studies. The sources of adaptive alleles will then be determined from the resequencing data and gene trees produced in Aim 1. The importance of gene flow to local adaptation will be further unpacked by examining the contributions of environmental, spatial and reproductive barriers to allele distributions in an innovative application of generalised dissimiliarity modeling (GDM).
3. **Are key trait differences between species also controlled by ancient genetic variation?** We will employ natural hybrids to determine the number and location of loci affecting seedling morphology and physiology variation among species. We will then examine the geographic and taxonomic ranges of the QTL alleles to understand the source of the variation and to understand the relationship between local and species-wide adaptation.

Eucalyptus background

Eucalyptus is a highly diverse genus of trees with over 700 species. Many of the landscapes dominated by these species have been heavily modified by human activity and are predicted to undergo further changes in climate, CO_2 concentration and interactions with other species. **Eucalypt seedlings respond to environmental gradients with variation in many traits**, such as growth and leaf morphology structure [7-9]. The underlying genetic basis of this variation is not well known, but could be important determinants of successful adaptation to changing environments.

An important feature of *Eucalyptus* is the widespread co-occurrence of related species**. Hybridization occurs among eucalypt species despite the presence of numerous reproductive barriers** [10] and is best studied in Section *Maidenaria*, where the phenomenon of "chloroplast capture" been shown to occur as a result of asymmetrical hybridisation related to flower size [11-13]. The breakdown of species barriers, due in part to habitat fragmentation, has also been found in *E. aggregata* and its relatives [14]. Widespread overlap in flowering times [15] can facilitate hybridisation between eucalypts and shared polymorphisms between species are common in both molecular [16] and phytochemical data [17-19]. Despite their implications for mechanisms of speciation, the evolutionary significance of hybridisation versus incomplete lineage sorting in *Eucalyptus* has not been addressed systematically.

In a time when **enhancing habitat restoration and managing adaptation to global change are important goals** in Australian research, the group of eucalypts that dominate many eastern Australian woodland communities, are a valuable target for ecological genetics and speciation research. Section *Adnataria* is a moderately speciose group of eucalypts, including most boxes and ironbarks (approx. 100 species), and is approximately 20 million years old [20, 21]. Species commonly co-occur over broad areas, but retain habitat associations at a medium scale and differ substantially in seedling traits. Hybridisation is well-documented [22] and in several cases, hybridising species are in different subsections of *Adnataria* and therefore unlikely to be sibling species. This may explain why the best attempts to elucidate relationships within the section Adnataria have achieved little resolution of the backbone of the tree [23]. This suggests that regions of the genome have heterogeneous evolutionary histories [24, 25], which can now be resolved with genomic methods.

RESEARCH PROJECT

Significance and Innovation

This project makes an internationally significant contribution to speciation biology as it tests emerging ideas on the mechanisms for adaptive divergence in a powerful study system (woodland *Eucalyptus*) and a novel context (plants that are broadly sympatric across shared environmental gradients). This project will be executed with unprecedented genomic and geographic resolution. It is timely to bring these new ideas and modern tools to bear on one of the greatest challenges of Australian botany, the nature of *Eucalyptus* species, as woodland communities face increasing stresses [26]. An important and innovative feature of the project is that it does not just look for evidence of these processes, but also validates them in an ecological context while exploring their impact on the critical early stages of eucalypt growth.

An innovative sequencing strategy will provide unprecedented coverage of the genome, making it possible for our landscape genomics analysis to overcome the low linkage disequilibrium that impedes most association studies of *Eucalyptus*. Low coverage of a larger number of individuals can improve estimates of allele frequencies, which are the basis of allele-environment associations, over high coverage of a smaller number of individuals [27]. The downstream analysis of this data is also innovative, as current methodologies do not offer the means to study landscape genomics across species boundaries. By adopting generalised dissimilarity modeling, a tool of community genetics, we can compare the importance of species boundaries, isolation by distance and environmental variation in shaping the distribution of alleles, which is a completely new application of the method.

High-throughput, automated phenotyping in climate chambers is another innovation in this project, as it allows QTL mapping of adaptive traits on large sets of progeny. Further, the study design utilises offspring of natural hybrids, a general approach exploiting a single generation of recombination and population variation.

Adaptation to environmental variation is crucial to the long-term persistence of Australia's biodiversity throughout the past and into the future, as human activities will continue to exert their influence on most, if not all, Australian habitats. This research falls under the Strategic Research Priority 'Living in a changing world', with the specific goal to 'identify vulnerabilities and boundaries to the adaptability of changing natural and human systems'. This study will not only identify genomic regions that are important for the adaptation of *Eucalyptus* species to climate heterogeneity, it will also determine the distribution of beneficial alleles. Since the study organisms are the dominant foundation species of endangered ecological communities, this information has implications for the adaptability of entire ecosystems. Substantial variation within populations for traits and genes associated with climate and soil variation is favourable for adaptation to changing environments; in contrast, there will be cause for concern if little variation is found within populations.

Conceptual framework

Here we outline the theoretical rationale for the proposal and then below in the Methods and Experimental Design outline three projects to experimentally identify adaptive loci using phylogenomics, landscape genomics, and family based association studies.

Ancient alleles: a new context for an old question

The importance of new mutations versus standing variation in adaptation is a fundamental evolutionary question [3, 28] that has been addressed in the context of human evolution [29], weed evolution [30], and domestication [31]. Now, it is becoming clear that not only standing variation, but ancient alleles from related species can play major roles in both the adaptation and speciation processes. This phenomenon first became apparent in studies of parallel adaptation, where adaptive alleles in have been shown to share ancestry among environments, while the rest of the genome is distinct. The two possible explanations for this are independent evolution from standing variation and gene flow between the populations. Support for the latter comes from *Heliconius* butterflies, where colour patterning genes appear to have crossed species boundaries and led to new reproductively isolated races within formerly panmictic species. Further, in sticklebacks, where direct migration between freshwater populations is implausible, ongoing hybridisation with marine populations can supply standing variation to enable new populations to adapt.

This concept is likely to be useful for interpreting patterns of allele sharing in *Eucalyptus*, as it is a clear adaptive role for gene flow between species with apparently strong reproductive barriers or disjunct populations. **Standing variation may allow faster adaptation than new mutations** for several reasons: they may be present at higher frequencies and in a greater variety of genetic backgrounds (soft sweep, Figure 2B), and they must have survived selection against negative pleiotropic effects. **An extrinsic source of genetic variation offers an even more efficient route to adaptation**, whether through direct adaptive introgression (Figure 2C) or indirect gene migration. Together, these forms of interspecific gene flow are a compelling alternative to the persistence of standing variation or repeated new mutations (Figure 2A). An important feature of the work proposed here is that our experimental design will allow us to identify the source and adaptive value of shared variation among species. We will then explore the geographic and environmental distributions of these ancient variants, and assess their contribution to phenotypic divergence between species.

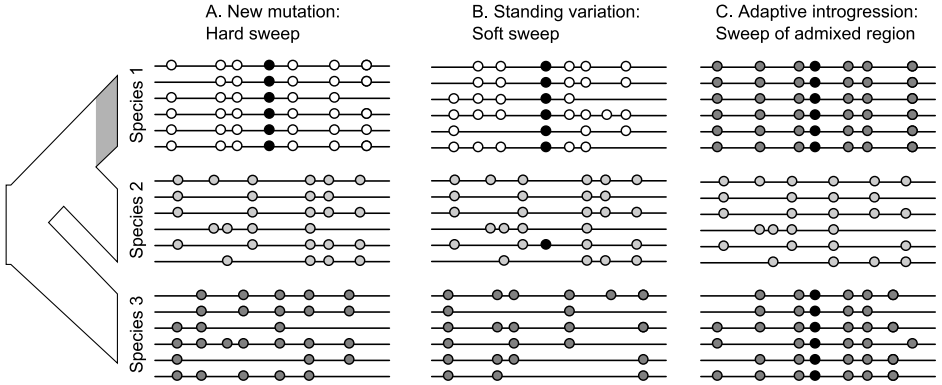

Figure 2. Genomic signatures of adaptation. A) The hard sweep of a new mutation reduces diversity of neighbouring loci in Species 1. B) Standing variation is utilized, resulting in a soft sweep with reduced effect on neighbouring sites. C) Adaptive introgression introduces large admixed regions with high similarity to other species. The adaptive allele is shown in black and neighbouring neutral variants are shaded according the species in which they are common.

The *Eucalyptus* in section *Adnataria* represent an attractive study system in which to examine the importance of these forces, as multiple species have adapted to numerous environmental gradients in a highly structured landscape. Similar to the recently-radiated African cichlids, there are greater numbers of potential donor species; however, these eucalypts co-occur over broad geographic ranges, creating replicated natural hybrid zones under different local conditions. Large effective population sizes required to maintain neutral alleles over long periods. *Eucalyptus* species can have effective population sizes above 10^7 [32] and are largely outcrossing, providing the opportunity for hybridisation. However, historic local adaptation and the current highly fragmented forest/agriculture system can limit gene flow. Nevertheless, the documented occurrence of hybridisation among woodland species implies that the re-use of adaptive variation among species could be facilitated by interspecific gene flow. This will be investigated in three Projects outlined in the Methods section.

Origins and range limits of shared variation

Landscape genomics is a growing field that merges spatial analysis methods with genomic approaches. Association tests are made between specific alleles and environmental variation to identify genes involved in local adaptation. To distinguish between neutral demographic processes, which affect the whole genome, and selection at adaptive loci, landscape genomic association studies include a term that controls for the underlying genetic structure when scanning for adaptive loci. Soft selective sweeps on standing genetic variation, including variation in polygenic adaptive traits

can thus be identified. Such methods have been used extensively to identify genes involved in adaptation to climatic variation, especially in plants [33, 34]. By performing a comparative landscape genomics analysis in closely-related species in concert with a broader phylogenomic study, we will be able to not only ask if the same genetic mechanism confers local adaptation to a given environmental gradient, but we will also be able to determine the age and source of this variation. This approach will help us understand the reason for persistence of shared polymorphism, as spatially heterogeneous selection can maintain adaptive alleles within species. Further it will give insight into whether the sources of **variation underlying local adaptation are similar to those involved in species divergence.**

Generalised dissimilarity modeling (GDM) is a tool developed in community ecology to understand dissimilarity in species composition through space [35]. It is an extension of basic Mantel and partial Mantel correlation tests. GDM produces predictive models and compares the importance of different predictors. It is particularly appropriate for landscape genomics applications because it can treat allelic dissimilarity at a locus as the response. Environmental, spatial geographic and background genomic dissimilarity matrices are used as predictors [36]. We will use the method in a novel way to assess the relative importance of species boundaries, isolation by distance and habitat filtering, in shaping the distribution of alleles to go beyond previous studies on single species [34]. Quantifying the proportions of alleles whose ranges are driven by each of these factors will allow us to **infer the major factors maintaining species divergence**.

We expect to see that species is the most important predictor of allele distributions throughout the genome, but also identify many exceptions at outlier loci. When neutral introgression is the cause of allele sharing, geographic distance is likely to be the only other significant predictor. Where environmental variables are also strong predictors, it would support local adaptation. Such support for adaptive introgression will reveal an underappreciated mechanism of evolution. It is likely that each of these outcomes will be observed in different parts of the genome and in different populations across the range. **This will have major implications for both conservation and restoration of these foundation species under rapid environmental change.** As described further below, Project 2 will apply comparative landscape genomics in two new species. We will combine the results with a parallel data set generated through ongoing work in *E. melliodora* (LP130100455). The project will draw upon the multi-species data set of Project 1 to understand the geographic patterns and fine-scale evolutionary history of alleles associated with both speciation and local adaptation.

Rapid analysis of the genetic architecture of species differences

Trait differences are central to niche separation between species, and consequently species coexistence. By investigating the genetic architecture of interspecific trait differences, researchers have gained insight into the number of quantitative trait loci (QTL) that control functional trait differences between species, their effect sizes and their degree of colocalisation within the genome [37]. The source of this variation has been examined in only a few model study systems [6, 38]. There is little understanding of how environmental heterogeneity affects traits underlying species differences, despite its importance in shaping historical and future adaptation of functional traits. Replicated natural hybrid zones are ideal for asking whether these traits are due to phenotypic plasticity or local adaptation.

Typical genetic mapping studies require more than one generation of crosses to provide sufficient recombination among divergent loci. In long-lived trees, however, it is possible to overcome this limitation by using natural hybrids and genomic technologies. We will make use of **open-pollinated seed and high density genotyping by sequencing to assay large progeny sets** in order to determine the functions of haplotypes inherited from the divergent genomes present in the maternal hybrid tree.[39]. Due to the linkage disequilibrium inherent in maternal families, high density genotyping will be suitable, rather than whole-genome sequencing, making this experiment very cost-effective.

QTL mapping will be carried out in Project 3, focusing on a pair of frequently-hybridising species based on the results of Project 1. *E. sideroxylon* and *E. albens* are the most likely targets, as one hybrid zone has been well-documented [22] and numerous herbarium collections have been identified as hybrids on the North-Western slopes and New England Tablelands of NSW. The concordance of the identified QTL will be compared between hybrid zones and the geographic spread of these alleles will be examined using the landscape genomics data set generated in Project 2.

Methods and Experimental Design

The research questions will be addressed through three interconnected projects. The first project will generate a reference data set of resequenced genomes, to determine the overall evolutionary relationships among species and the genomic extent of species distinctness. The second project will examine the relative importance of species boundaries and environmental factors using landscape genomic scans for adaptive loci. The third project will utilise natural interspecific hybrids to dissect the genetic architecture of species traits in multiple contact zones and examine the source of this variation with reference to the first two projects.

Project 1: Characterisation of the genomic distinctiveness of species

Goals

1. To examine the genetic distinctness of *Adnataria* species: what proportion of the genome is differentiated among species?
2. To examine the extent of allele sharing and the importance of ancient allelic variation
3. To identify genomic regions that have undergone positive and balancing selection during their shared evolutionary history.

This work will be conducted primarily by a postdoctoral fellow under the supervision of CI Andrew.

Approach

- **Sample 10 widely-distributed species, including pairs of species that are known to hybridise**, such that species within both major anther groups are represented. Likely *Adnataria* species include *E. melliodora, E. sideroxylon, E. albens, E. polyanthemos, E. microcarpa, E. odorata, E. leucoxylon,* and *E. caleyi*. In addition, we will include more distantly-related species that co-occur with *Adnataria* woodland species, such as *E. blakeleyi* (Blakely's red gum, Subgenus *Symphyomyrtus*: Section *Exsertaria*) and *E. macrorhyncha* (Subgenus *Eucalyptus*). We will use the Australian Virtual Herbarium (avh.ala.org.au) and herbarium specimens to plan sampling locations. At least 20 individuals will be collected from different pure stands throughout the range of each species, aiming for 10 samples whose classification can be confirmed. Care will be taken to correctly identify each specimen in the field and herbarium specimens will be sent to both the UNE Beadle Herbarium and the Australian National Herbarium for further corroboration.
- **Whole genome shotgun (WGS) resequencing** of 10 individuals per population at 10x coverage, using 100bp paired-end sequencing on Illumina HiSeq 1T. This will provide both deep 100x coverage per species but also high confidence in individual sequence variation within species to identify highly diverged and admixed regions among individuals. Although this approach is expected to miss some rare variants and therefore slightly underestimate allele sharing, it will provide unprecedented resolution to identify portions of the genome that are cohesive within each species and divergent between them.
- **Align reads to the annotated *E. grandis* reference genome** (www.phytozome.net). Initially the high quality reference genome will be used. Though a more distant relative, *E. grandis* is equidistant from the Adnataria species and any bias should therefore affect these species equally. We will compare results using a composite *Adnataria* reference derived from samples in this study. The depth of coverage obtained for this project will enable the read-backed phasing of haplotypes, which is important for downstream analysis. Whole genome *de novo* assembly will also be considered as tools and computational resources are now available (2Tb RAM *de novo* assembly nodes at ANU Genome Discovery Unit).
- **Outlier analyses and tests of selection** will be conducted in to identify loci displaying signatures of divergent selection. These will include divergence outlier analyses, such as BAMOVA [40]; haplotype-based methods, such as expected haplotype homozygosity [41] and the cross-population composite likelihood test [42]; and methods based on synonymous and non-synonymous mutations, such as the McDonald-Kreitman test. We will identify variants that are unique to a species or shared between multiple species and use the annotated genome to distinguish coding and non-coding sites. The proportion of amino acid substitutions fixed due to divergent selection, produced via the McDonald-Kreitman procedure, will be compared between species pairs [43].
- **Identify sources of variation from gene trees in windows within the genome**, using *E. grandis* as an outgroup. We will aim to identify non-recombining loci, which will be much more achievable at the broad geographic scale of the study and use phased haplotypes within the blocks where possible. However, appropriate window widths will need to be determined from the data. Across all *Adnataria* species, we will characterise the range of topologies obtained using the self-organising maps/hidden Markov model approach that enabled the identification of ancient adaptive variation in sticklebacks [44]. We anticipate that reciprocal monophyly of species is likely to be rare within the genome, owing to both incomplete lineage sorting and interspecific gene flow. Recovery of clades corresponding to species for

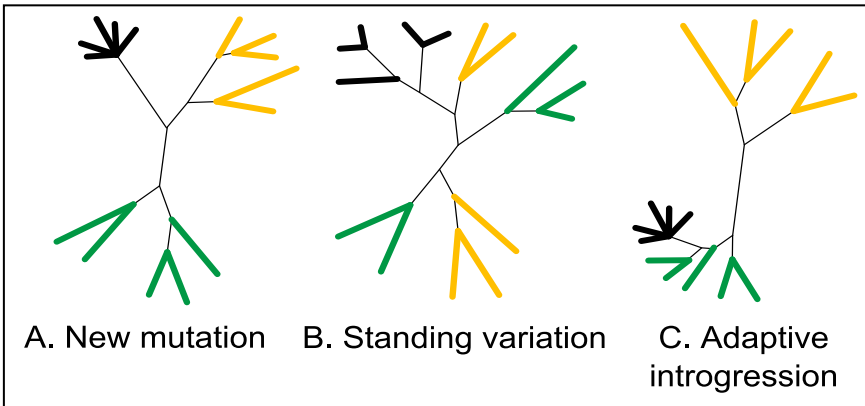

Figure 3. Simplified examples of gene trees expected under adaptation via A) new mutation, B) standing variation (with incomplete lineage sorting), and C) adaptive introgression. Tips are coloured by species.

particular windows would suggest cohesion of species, which may be driven by species-wide selective sweeps [45]. The branch lengths and distributions of coalescent times will be used to distinguish between incomplete lineage sorting due to balancing selection or neutral processes [5] and the tree topologies will indicate the extent of allele sharing, both inside and outside Section *Adnataria* (Figure 3).

- **Estimate species trees using statistical approaches**, such as *BEAST [46], which combines gene trees, and SNAPP, which takes genome-wide SNP data [47]. Following Carstens et al. [48], we will apply multiple methods and interpret the results conservatively. We may need to subsample loci in order to use some of these methods, but this is an active area of statistical development, particularly regarding the effects of gene flow on species tree estimation [49]. We will compare trees obtained using randomly-selected loci vs those that are monophyletic or fixed for at least one species. While incomplete taxon sampling may influence species tree accuracy, we anticipate that gene flow is likely to be a more serious challenge. Estimates of population sizes and divergence times can contain signatures of migration events [49], and will be used along with established tests [50] to identify specific genomic windows that show evidence of interspecific gene migration. The delimitation of species and estimation of their evolutionary relationships will indicate the breadth of allele sharing in this group, and will inform the selection of focal species for Projects 2 and 3.

Additional outcomes

1. Genome resequencing data that will fuel collaborations and student projects, as well as forming a public resource for researchers.
2. Resolving the species tree of these select species will help to inform future high-resolution efforts to understand the evolutionary history of the entire Section *Adnataria*.

Project 2: Geographic and environmental distribution of genetic variation across species

Goals

1. To identify genetic variants and genomic regions associated with environmental variation and make comparisons among species
2. To examine the role of ancient variation and recent interspecific gene flow in shaping allele sharing between species
3. To dissect the relative importance of species boundaries, spatial distance and environmental variation in shaping the range limits of alleles

This work will build on Project 1 and will be led by CI Andrew.

Approach

- **Range-wide sampling of two species that share large proportions of their allelic variation**, based on the results of Project 1. This will ensure that the ranges of polymorphisms can be compared between the species. We will collect leaf material from at least 20 individuals from 40 populations of each species. Putative hybrids will be collected in preparation for Project 3, but will not be included in the analysis for Project 2.
- **Low-coverage WGS** to characterise allele frequency distributions across the ranges of the two species. This approach, sometimes called "genome skimming," will produce up to 1x coverage of the nuclear genome, as well as an estimated 10x coverage of the mitochondrial genome and >200x coverage of the chloroplast genome. With 10-20 individuals sequenced per population, this will yield approx. 15x coverage per population. This design will efficiently assay the nuclear genome for allele frequencies at all SNPs in regions of the nuclear genome that are amenable to assembly. Compared to reduced representation sequencing, this approach provides the opportunity to **detect the most extreme outlier SNPs in the entire genome**, rather than those that are outliers merely by virtue of linkage to variants that are under selection. Strong selective sweeps will be readily detected, although the causal variants may be obscured, while there will also be power to detect selection on individual genes in soft sweeps. Each individual will be uniquely barcoded so that any aberrant samples or undetected recent hybrids can be identified and excluded from landscape genomic analysis, which would not be possible with typical pooled sequencing. If library construction and sequencing costs continue to decrease, as we anticipate they will do, the number of individuals sequenced per population will be increased.
- **Align to composite reference sequence** that will be generated by consensus from samples of both species in the resequencing data set of Project 1. This will avoid reference bias, while still providing a better reference for these species than the *E. grandis* reference genome.
- **Collect relevant environmental data** using GIS. Contemporary and historical climate data are available for the study area and we will collaborate with GIS specialists to select the most appropriate layers. Since regional soil characteristics are of potentially greater importance to filtering eucalypts species than climate, soil layers will also be included. We will use diverse data sources, including Geoscience Australia and WorldClim. For example, the Atlas of Australian Soils describes considerable variation in soil landscapes over *Eucalyptus* woodlands in eastern Australia

and the National Soil Grid supplies useful soil variables on a 250 m grid. While microsite climate and soil variation is very important for trees, the spatial scale of adaptation that we can detect is much greater (10-100km) due to historic gene flow across the landscape prior to habitat fragmentation.

- **Analysis of population structure** will be carried out on nuclear data, in order to explore the population structure within species. Hierarchical analysis of molecular variance (AMOVA) will be carried out using software designed to handle the genotype uncertainty inherent in low-coverage sequence data, such as BAMOVA [40]. We will test for range expansions using the directionality index of Peter and Slatkin [51], which is based on the frequencies of derived alleles as determined via interspecific comparison.
- **Comparative Bayesian landscape genomics will be used to identify allelic variation/genomic regions associated with climate and soil gradients within species**. The *BayEnV* software [52] uses a genetic covariance matrix to control for background genetic structure while testing for associations between loci and environmental variables. Unlike divergence outlier methods, this approach is robust to a variety of demographic scenarios. Since this method for identifying adaptive loci relies on the availability of the relevant environmental data, it will be used in concert with conventional outlier analysis and model-based spatial analysis [53] to assess variation that may be associated with unmeasured environmental variables. We will make comparisons between species (Figure 4). Parallel adaptations may be observed at the level of the genomic region or individual SNP. Derived alleles at selected loci will also be identified using the *E. grandis* genome as an outgroup and the geographic extent of such alleles will be compared between species, as well landscape genomics studies of other tree species.

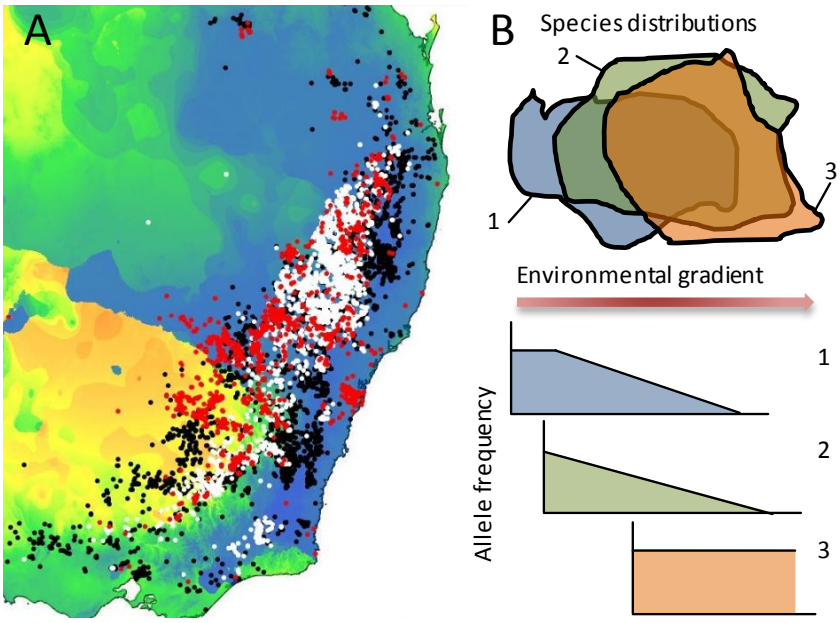

Figure 4. A) Distribution of *E. albens*, *E. sideroxylon* and *E. melliodora*, overlain on a climate variable (the temperature of the driest quarter). B) Schematic of comparative landscape genomics analysis of three species with overlapping distributions along an environmental gradient. An allele shows parallel environment associations in species 1 and 2, but not in species 3.

In addition to the two focal species considered here, we will compare the loci and genomic regions identified in this study with the results of a current ARC Linkage Project (LP130100455) that aims to use landscape genomics, high-throughput phenomics and association mapping to improve prospects for land restoration using *E. melliodora* in eastern Australia and *E. marginata* in Western Australia. CI Borevitz is lead investigator on that project and CI Andrew is a collaborator on the landscape genomics component. While the genomic resolution of the current proposal is much greater and the focus is on shared variation, the Linkage Project involves evidence from phenotypic variation and plasticity to support tests of local adaptation. In addition, if funded, the Future Fellowship proposal by CI Andrew (FT140101197) will provide landscape genomics and phenotypic data on additional species for comparison.

- **Examine the factors shaping the distributions of alleles using generalised dissimilarity modeling (GDM)**. This method, developed originally for community ecology applications, can compare the importance of species boundaries, environmental variation and geography in predictive models of the allelic composition of each locality. The ability of GDM to incorporate spatial distance or other dissimilarity measures as potential predictors is a strength of this method for genomics, as the genomic background covariance can be controlled for in the model (much like kinship is controlled for in association studies [54]). We will test outliers against empirical distributions of the GDM function parameters.

Additional outcomes

1. Predictions of genotypes suitable for existing and future conditions and the value of interspecific hybrids as sources of useful alleles
2. The forces shaping cytoplasmic introgression will be better understood as a result of this project. An adaptive role for plastid and mitochondrial adaptation is plausible, based on recent results [55]; however, we have until now lacked genomic tools to address this issue in *Eucalyptus*.

Project 3: Genetic architecture of species differences in seedling traits and phenotypic plasticity

Goals

1. To investigate the genetic architecture of species differences in seedling traits and their responses to climate variation, in order to determine whether ancient variants or introgressed genes are involved.
2. To explore the geographic distributions of QTL variants in each species and ask whether the different hybrid zones display concordant genetic architectures of species traits.
3. To examine macro-synteny within Adnataria and with the sequenced reference genomes, and identify putative regions controlling incompatibilities through their effects on segregation distortion.

This work will focus on the species used in Project 2 and will be undertaken by a PhD student to be co-supervised by CI Andrew and CI Borevitz, with phenotyping and genotyping carried out at ANU.

Approach

- **Reduced representation genotyping by sequencing** will be used for this part of the study, using restriction-associated DNA sequencing (GBS/RAD) method [56, 57]. This approach sequences the same subset of the genome in many individuals at with high confidence (>20x coverage), allowing high levels of multiplexing (384 per lane) using 100bp paired-end sequencing on Illumina HiSeq 1T. The protocol can be tuned to yield >20k reliable markers, which is sufficient to capture all recombination events, with high coverage at each site to enable confident genotyping. Much of the development work has already been done already for *E. melliodora* as part of ARC Linkage Project (LP130100455) by Dr Justin Borevitz's group, who have already demonstrated successful multiplexing at this level. We will **extract annotated, high quality sequence tags** after alignment, using existing pipelines.
- **Collect large maternal seed families from pure and hybrid populations** early in Year 2 or Year 3, depending on conditions for successful seed set. The pure seedlots will each comprise 300 seeds sourced from two unrelated mother trees from a single pure population. The hybrid seedlots will comprise 600 seeds from hybrid individuals in three distinct hybrid zones or swarms. Hybrid individuals will be identified using GBS/RAD sequencing of morphologically intermediate adult trees collected during field work for Projects 1 and 2.
- **High-throughput phenotyping of hybrid progenies in two climate regimes** will be carried out in the ANU Climatron phenomics facility (LE130100081). This facility is run by CI Borevitz's group and can measure seedling growth rates, photosynthetic rates and leaf morphology. In each set of 600 samples, two environmental treatments will be applied to approximate the climate differences between the hybrid zones.
- **Genetic mapping of QTL for seedling traits.** First we will construct high-quality genetic maps from pure progeny using GBS/RAD sequence tags. This approach will yield multi-allelic markers, which, unlike biallelic SNPs, will be able to distinguish between the two maternal alleles and the pollen donor's allele for a high proportion of markers, owing to the high nucleotide diversity of *Eucalyptus* (Figure 5). The most efficient approach to mapping using dense markers is to increase sample sizes while decreasing coverage, as imputation can be used to fill in missing genotypes. The statistical design will depend on the number of mothers per hybrid zone and the structure of the pollen pool for each one.
- **Compare genetic map structure** for the two species and examine patterns of **segregation distortion in the hybrid progenies, which may indicate intrinsic reproductive barriers.** We will have power to detect even subtle deviations (>60/40) in parental genome contribution to progeny set as large genetic regions will be well covered with genetic markers. We will then **examine the concordance of hybrid zones and the geographic ranges of the QTL** using the data generated for Project 2.

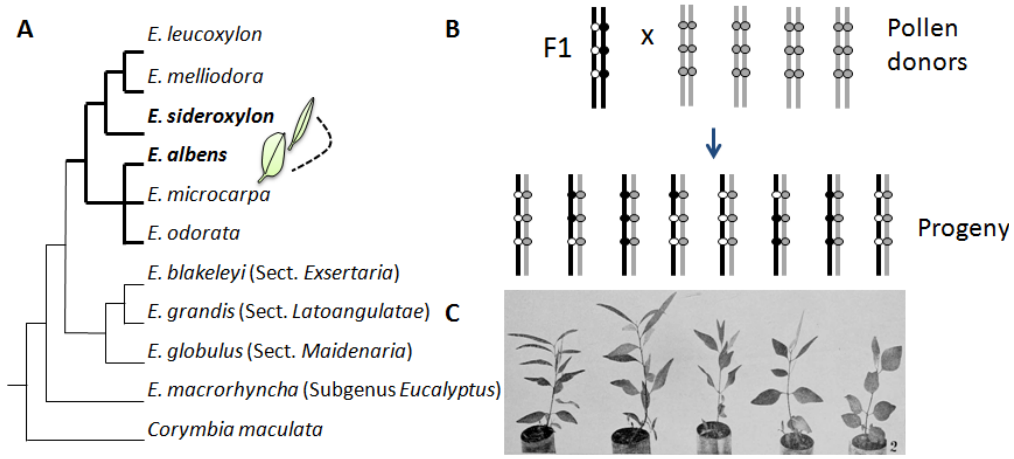

Figure 5. A) Juvenile leaf shape of a pair of species in section *Adnataria* with well-documented hybrid zones. B) Mapping design in natural hybrids, simplified for one chromosome and a few markers. A natural F1 hybrid has undergone natural open pollination by unknown pollen donors. Informative sequence loci will distinguish both maternal alleles (white or black) from the paternal allele (grey). C) Variation in seedling morphology among open-pollinated progeny from a single *E. albens* x *sideroxylon* mother.

Additional outcomes

1. High quality genetic maps of two previously unstudied *Eucalyptus*

species. These will be used to improve analysis of the genomic structure of divergence and gene flow between species.

2. Measurements of the plasticity of early seedling growth and other speciation related traits, which will inform the predictability of seedling genetic effects in restoration projects across different environments.

Timeline

This project is budgeted to run over four years in order to integrate the projects effectively. Preliminary results for Project 1 are needed to select the target species for Projects 2 and 3. Due to recent drought in NSW, we may require two years of sampling to obtain seed for Project 3.

Project/Person	Year 1	Year 1	Year 2	Year 2	Year 3	Year 3	Year 4	Year 4
Phylogenomics								
Landscape Genomics								
Phenotype Association								
Research assistant 0.5FTE								
Postdoc								

ROLE OF PERSONNEL

CI Andrew will participate in all of the field collections and supervise the molecular work. She will supervise the phylogenomic analysis in Project 1 and conduct the landscape genomic analysis in Project 2.

CI Borevitz will oversee the sequencing for all projects and provide support for the phenotyping, which will take place in his facilities at ANU. He will conduct the QTL analysis for Project 3.

The Postdoctoral Researcher, who will have a PhD in phylogenomics or comparative genomics, will perform the bulk of the analysis for Project 1 based at UNE. This will involve the use of standard pipelines for sequence processing and alignment. They will be responsible for phylogenomic analysis under the supervision of CI Andrew.

The Research Assistant will process field samples, extract DNA and construct WGS and ddRAD libraries at UNE. They will visit ANU for training in ddRAD and WGS library construction.

RESEARCH ENVIRONMENT

The proposed research combines the strengths of UNE, which has a strong long-term focus on environmental science and ecosystem management, with those of ANU, a major research institution with world class facilities for sequencing and phenotyping.

The strategy of UNE, as outlined in the 2012-2015 Research Plan, is to focus on further improving research excellence in its areas of specialisation. The School of Environmental and Rural Sciences is central to several of the Univerisity's research specialisations and hosts a number of Cooperative Research Centres. Of the five research themes of the University, this project aligns with the theme of "Climate change and environmental sustainability; protecting biodiversity, effective policies". The recent hiring of CI Andrew is part of a push to bring modern molecular methods to bear on the sustainability challenges facing Australia. In a changing climate, the protection of biodiversity relies on understanding its climate specificity. Determining the distribution of useful genetic variation in woodland eucalypts lays the groundwork for more informed management of these heavily fragmented ecosystems. The CRC for Spatial Information, which is partially hosted by the School, has direct relevance to this theme and brings critical mass in spatial analysis that will be help create a successful environment for this project. Through innovative use of high-throughput genomics technologies, the research proposed here also contributes to the University's "smart science, smart technology" goal.

UNE is a member of Intersect Australia Ltd. Intersect Australia Ltd is a not-for-profit company limited by guarantee, owned and funded by its members, including the ten universities in NSW, the University of Canberra, state government departments, and other organisations undertaking research in NSW. Intersect has established a capacity and capability to develop, deploy and support substantial and complex eResearch infrastructure, that is unique in Australia. Intersect will provide data storage, High Performance Computing and systems support to the Project. These resources will be complimented by the computing facilities at ANU, which include 2Tb RAM de novo assembly nodes at ANU Genome Discovery Unit.

ANU/RBS has a long history of population genetic research and a tradition of working with partners in the field, and this will be facilitated in the future with the establishment of the Centre for Biodiversity Analysis (CBA, directed by Prof Moritz and CI Borevitz is a steering committee member). Recent investments and upgrades to plant growth facilities (NCRIS) have provided phenomics capacity (including Borevitzlab senior staff) across climate simulated chambers in the ANU Climatron facility. ANU has local next generation sequencing capacity (Illumna Hiseq) with

multiplex reduced representation methods already developed (Borevitzlab). The collaboration between UNE and ANU will be maintained through regular meetings at conferences and combined field trips.

COMMUNICATION OF DATA

We will continue to publish in high-ranking international journals. We anticipate at least one major integrative paper, as well as methods and/or data papers in more specialised journals to result from the work addressing each Project. In addition, a synthetic summary of this research and other related work will be worthwhile late in the project. The results will also be presented at international and domestic conferences, most likely the Plant Animals Genomes conference, and meetings of the European Society for Evolutionary Biology, the Society for Molecular Biology and Evolution, Australasian Evolution Society and the Australian Genomic Technologies Association. I expect that there will be sufficient interest for a special symposium at an international evolutionary biology meeting, which will likely result in a special issue or special issue of a journal such as *Evolution* or *Molecular Ecology*.

Eucalypts are of immense cultural importance in Australia. The public affection for the genus gives the research proposed here the potential to enhance public interest in science through outreach. We will seek opportunities for public lectures and use the data generated to enrich our undergraduate teaching. It is particularly important to acknowledge the value of public participation in science by communicating outcomes of the research.

MANAGEMENT OF DATA

Data, protocols and analysis scripts will be made publicly available to contribute to future primary research and meta-analysis. The extensive sequence data generated for this project will form a valuable resource for comparative and population genomics studies of *Eucalyptus*, as well as for tree breeders and conservation geneticists. Sequence data will be submitted to the NCBI Sequence Read Archive and *Molecular Ecology Resources*, and other data and scripts will be uploaded to the Dryad repository at publication.

Data generated from this proposal will be managed with the support of Intersect Australia services. All data and associated metadata will be stored securely on the Intersect RDSI node and will be accessible to project partners during and after the life of the Project. Data will be backed up and replicated, protecting the investment in collecting it. The CIs will ensure open access to all of the data stored on the RDSI after the life of the project. Data will be citable and discoverable Research Data Australia managed by Australian National Data Service (ANDS). A primary copy of the data will be retained for a minimum of 10 years after publication to ensure the data is available for future research and dissemination.

References

1. Faria, R., et al. (2014), Mol Ecol, 23(3): p. 513-521. **2.** Seehausen, O., et al. (2014), Nat Rev Genet, 15(3): p. 176-192. **3.** Barrett, R.D.H. and D. Schluter (2008), Trends Ecol Evol, 23(1): p. 38-44. **4.** Seehausen, O. (2004), Trends Ecol and Evol, 19(4): p. 198-207. **5.** Hedrick, P.W. (2013), Mol Ecol, 22(18): p. 4606-4618. **6.** Schluter, D. and G.L. Conte (2009), Proc. Natl. Acad. Sci. USA, 106(S1): p. 9955-9962. **7.** Li, C., et al. (2000), Funct Plant Biol, 27(3): p. 231-238. **8.** Andrew, R.L., et al. (2010), Ann. Bot., 105(5): p. 707-717. **9.** McLean, E.H., et al. (2014), Plant Cell Env: In press. **10.** Griffin, A.R., I.P. Burgess, and L. Wolf (1988), Aust. J. Bot., 36(1): p. 41-66. **11.** Jackson, H.D., et al. (1999), Mol Ecol, 8(5): p. 739-751. **12.** Field, D.L., et al. (2011), Heredity, 106(5): p. 841-853. **13.** Gore, P.L., et al. (1990), Aust. J. Bot., 38(4): p. 383-394. **14.** Field, D.L., et al. (2008), J Ecol, 96(6): p. 1198-1210. **15.** Keatley, M. and I. Hudson (2007), Environ Model Assess, 12(4): p. 279-292. **16.** Kulheim, C., et al. (2009), BMC Genomics, 10(1): p. 452. **17.** Andrew, R.L., A. Keszei, and W.J. Foley (2013), Phytochemistry, 94(0): p. 148-158. **18.** Padovan, A., et al. (2012), J. Chem. Ecol., 38(7): p. 914-923. **19.** Keszei, A., C.L. Brubaker, and W.J. Foley (2008), Aust. J. Bot., 56(3): p. 197-213. **20.** Steane, D.A., et al. (2002), Aust Syst Bot, 15(1): p. 49-62. **21.** Crisp, M., L. Cook, and D. Steane (2004), Phil Trans Roy Soc B-Biol Sci, 359(1450): p. 1551-1571. **22.** Pryor, L.D. (1953), Proc Linn Soc NSW, 78: p. 43-48. **23.** Woodhams, M., et al. (2013), Systematic Biology, 62(1): p. 62-77. **24.** McKinnon, G.E., et al. (2005), Aust. J. Bot., 53(8): p. 827-838. **25.** Poke, F.S., et al. (2006), Mol Phylogenet Evol, 39(1): p. 160-170. **26.** Dunlop, M., et al. (2012), *The Implications of Climate Change for Biodiversity Conservation and the National Reserve System: Final Synthesis.* , CSIRO Climate Adaptation Flagship: Canberra. **27.** Alex Buerkle, C. and Z. Gompert (2013), Mol Ecol, 22(11): p. 3028-3035. **28.** Ralph, P. and G. Coop (2010), Genetics, 186(2): p. 647-668. **29.** Pritchard, J.K., J.K. Pickrell, and G. Coop (2010), Curr Biol, 20(4): p. R208-R215. **30.** Kane, N.C. and L.H. Rieseberg (2008), Mol Ecol, 17(1): p. 384-394. **31.** Innan, H. and Y. Kim (2004), Proc. Natl. Acad. Sci. USA, 101(29): p. 10667-10672. **32.** Yeoh, S.H., et al. (2013), Mol Phylogenet Evol, 68(3): p. 498-501. **33.** Eckert, A.J., et al. (2010), Genetics, 185(3): p. 969-982. **34.** Hancock, A.M., et al. (2011), Science, 334(6052): p. 83-86. **35.** Ferrier, S., et al. (2007), Divers and Distrib, 13(3): p. 252-264. **36.** Freedman, A.H., et al. (2010), Mol Ecol, 19(17): p. 3773-3788. **37.** Kirst, M., et al. (2004), Plant Physiol, 135(4): p. 2368-2378. **38.** Rieseberg, L.H., et al. (2003), Science, 301(5637): p. 1211-1216. **39.** Grattapaglia, D. and R. Sederoff (1994), Genetics, 137(4): p. 1121-1137. **40.** Gompert, Z. and C.A. Buerkle (2011), Genetics, 187(3): p. 903-917. **41.** Sabeti, P.C., et al. (2007), Nature, 449(7164): p. 913-918. **42.** Chen, H., N. Patterson, and D. Reich (2010), Genome Research, 20(3): p. 393-402. **43.** Smith, N.G.C. and A. Eyre-Walker (2002), Nature, 415(6875): p. 1022-1024. **44.** Jones, F.C., et al. (2012), Nature, 484(7392): p. 55-61. **45.** Morjan, C.L. and L.H. Rieseberg (2004), Mol Ecol, 13(6): p. 1341-1356. **46.** Heled, J. and A.J. Drummond (2010), Mol Biol Evol, 27(3): p. 570-580. **47.** Bryant, D., et al. (2012), Mol Biol Evol, 29(8): p. 1917-1932. **48.** Carstens, B.C., et al. (2013), Mol Ecol, 22(17): p. 4369-4383. **49.** Leaché, A.D., et al. (2014), Syst Biol, 63(1): p. 17-30. **50.** Patterson, N., et al. (2012), Genetics, 192(3): p. 1065-1093. **51.** Peter, B.M. and M. Slatkin (2013), Evolution, 67(11): p. 3274-3289. **52.** Coop, G., et al. (2010), Genetics, 185(4): p. 1411-1423. **53.** Yang, W.-Y., et al. (2012), Nat Genet, 44(6): p. 725-731. **54.** Atwell, S., et al. (2010), Nature, 465(7298): p. 627-631. **55.** Leinonen, P.H., et al. (2013), Mol Ecol, 22(3): p. 709-723. **56.** Peterson, B.K., et al. (2012), PLoS ONE, 7(5): p. e37135. **57.** Elshire, R.J., et al. (2011), PLoS ONE, 6(5): p. e19379.