Australian Systematic Botany, 2016, **29**, 235–246
doi:10.1071/SB16019_AC

Supplementary material

Effective application of next-generation sequencing (NGS) approaches in systematics and population genetics: case studies in *Eucalyptus* and *Acacia*

Hugh Cross[A,B,C], *Ed Biffin*[A,B], *Kor-jent van Dijk*[B], *Andrew Lowe*[B] *and Michelle Waycott*[A,B]

[A]State Herbarium of South Australia, Department of Environment, Water and Natural Resources, PO Box 1047, Adelaide, SA 5001, Australia.

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

[C]Corresponding author. Present address: Norwegian Institute of Bioeconomy Research, Department of Forest Health, Postboks 115, N-1431 Aas, Norway. Email: hugh.cross@nibio.no; hughbcross@gmail.com

Contents

Table S1. Taxa and populations used in case studies

Species	Population name	Population location	Number in study
Case study 1			
A. pinguifolia J.M.Black	Finnis Natural	Near Finiss Township, Fleurieu Peninsula, SA	6
A. pinguifolia	Finnis Restauration	Mixed restoration plot on Brimarvi Road, Near Finiss Township, Fleurieu Peninsula, SA	41
A. pinguifolia	Eyre Peninsula Natural	Between townships of Cockaleechie and Koppio, Eyre Peninsula, SA	5
A. pinguifolia	Seedlings	Nursery	38
Case study 2			
A. argyrophylla Hook.	Mount Torrens	Mount Torrens-Tepko Rd, TungkilloSA	3
A. brachybotrya Benth.	Rocky Gully	Near Rocky Gully Township, SA	1
A. aff. *brachybotrya*	Dog Fence	Near Dog Fence, coastal area near Wahgunyah CP, SA	2
A. pycnantha	Black Hill (BH)	Black Hill Conservation Park (CP), SA	4
A. pycnantha	Charleston (CH)	Charleston CP, SA	3
A. pycnantha	Dutchmans Stern (DS)	Dutchmans Stern CP, SA	4
A. pycnantha	Hale (HL)	Hale CP, SA	4
A. pycnantha	Kaiser Stuhl (KS)	Kaiser Stuhl CP, SA	5
A. pycnantha	Mt. Remarkable (MR)	Mt. Remarkable National Park, SA	4
A. pycnantha	Scott Creek (SC)	Scott Creek CP, SA	4
A. pycnantha	Wilpena Pound (WP)	Flinders Ranges National Park, SA	4
A. rivalis J.M.Black	Flinders Ranges 1	Flinders Ranges National Park, SA	4
A. rivalis	Flinders Ranges 2	Flinders Ranges National Park, SA	4
A. spilleriana J.E.Br.	World's End	World's End, northern end of Hallelujah Hill's road, SA	3
A. aff. *spilleriana*	Ika	Ika south of Gladstone, SA	3
Case study 3			
Angophora costata Domin	ABG	Cultivar, Adelaide Botanic Garden	1
A. euryphylla L.A.S.Johnson & K.D.Hill	ABG	Cultivar, Adelaide Botanic Garden	1
Calytrix tetragona Labill.	Waraweena	Waraweena, Beltana, SA	1
C. tetragona	Hale	Hale CP, SA	1
Corymbia ficifolia (F.Muell.) K.D.Hill & L.A.S.Johnson	ABG	Cultivar, Adelaide Botanic Garden	1
C. maculata (Hook.) K.D.Hill & L.A.S.Johnson	ABG	Cultivar, Adelaide Botanic Garden	1
Eucalyptus baxteri (Benth.) Maiden & Blakely ex J.M.Black	Deep Creek	Deep Creek CP, SA	1
E. cladocalyx F.Muell.	Dutchmans	Dutchmans Stern CP	1
E. leucoxylon F.Muell.	Flinders	Flinders Ranges National Park, SA	3

Species	Population name	Population location	Number in study
Eucalyptus macrorhyncha F.Muell. ex Benth. *subsp. macrorhyncha*	Spring Gully	Spring Gully CP, SA	1
E. microcarpa (Maiden) Maiden	Waite	Waite Arboretum remnant flora, Adelaide, SA	1
E. oblique L'Her.	Deep Creek	Deep Creek CP, SA	1
Leptospermum myrsinoides Schltdl.	Deep Creek	Deep Creek CP, SA	1
Melaleuca orophila Craven	Tepko	Tepko, SA	1

Detailed laboratory methods

The protocol to generate our reduced-representation library follows a standard approach for amplified fragment-length polymorphism (AFLP; Vos *et al*. 1995), substituting Ion Torrent adaptor sequences for fluorescently labelled primers in the final amplification step.

Restriction digest

The restriction digest step is critical for the success of any complexity reduction method that utilises a specific base-cutting enzyme. The discovery of homologous regions between samples is the central goal of this approach; thus, any variation in treatment of samples will lower the number of successful matches between samples. If digestion is incomplete in some samples, then subsequent amplification will produce very different regions than in other samples. As with any method that utilises restriction enzymes, good quality and sufficient enzyme are the key. In some cases, we used a lower concentration of enzyme, usually with no appreciable effect. However, the concentration of DNA should be monitored closely when adjusting the quantity of enzyme, as samples with much greater amounts of DNA than others may not be completely digested.

For each sample, genomic DNA was digested with the restriction enzymes *Eco*RI HF (R3101T or R3101L, New England Biolabs, Beverly, MA, USA) and *Mse*I (R0525L, New England Biolabs) in a 20-μL reaction for 3–4 h at 37°C, followed by 65°C for 20 min to deactivate the enzyme. To each sample was added 20 μL of a mixture containing the annealed adapters (*Eco*RI forward CTCGTATACTGCGTACC and reverse AATTGGTACGCAGTA, and *Mse*I forward GACGATGAGTCCTGAG and reverse TACTCAGGACTCATC) along with T4 DNA ligase (M0202L or M0202S, New England Biolabs), and incubated overnight at 16°C or for 4 h at room temperature (Table S2).

Preselective amplification

The restriction–ligation reactions were diluted 1 : 10 and used as a template for an initial polymerase chain reaction (PCR). The first amplification of the digested fragments utilised a proofreading polymerase optimised for difficult templates (DyNAzyme EXT DNA polymerase F505L or F505S, Finnzyme, Thermo Fisher Scientific, Waltham, MA, USA). Other standard polymerases, such as Amplitaq Gold, (Life Technologies, Carlsbad, CA, USA) were tried, but did not perform as well. This amplification consisted of a 25-μL reaction for each sample containing DyNAzyme buffer, dNTPs, $MgCl_2$, adaptor primers (Tables S3, S4), DyNAzyme polymerase and diluted restriction–ligation reaction, run on a thermalcycler for 95°C for 5 min, then 20–30 cycles at 95°C for 30 s, 56°C for 30 s, and 72°C for 1 (to 2) min; followed by an extension step of 10 min at 60°C. The preselective PCR products were then diluted 1 : 20 in water.

Table S2. Reaction conditions used in restriction digest–adaptor ligation

Reagent	Volume used per sample (μL)	Stock concentration	Final concentration	Concentration tested
Restriction digest				
Molecular-grade water	6.3			
10× NEB buffer 2	2	10×	1×	
*Mse*I	1	10 units μL^{-1}	10 units	
EcoR1 HF	0.5 (0.1)	20 units μL^{-1} (100)	10 units	
BSA mol. biol. grade	0.2	1 mg mL^{-1}	200 pg	
DNA	10	~20 ng μL^{-1}	10 ng μL^{-1}	~1–100
Total	20			
Adaptor ligation				
Molecular-grade water	11			
T4 DNA ligase	4	10×	1×	
Reaction buffer (NEB)				
Adaptor *Eco*R1	2	5 μM	0.25 μM	
Adaptor *Mse*I	2	50 μM	2.5 μM	
T4 DNA ligase	1	400 U μL^{-1}	400 units	
Total	20			

Table S3. Reaction conditions used in preselective amplification

Reagent	Volume used per sample (μL)	Stock concentration	Final concentration
Molecular-grade water	12.25		
10× DyNAzyme buffer	2.5×	10×	1×
dNTPs	2	2.5 mM each	200 μM each
$MgCl_2$	1	50 mM	2 mM
*Eco*R1 +A primer	2.5	5 μM	0.5 μM
*Mse*I +C primer	2.5	5 μM	0.5 uM
DyNAzyme polymerase	0.25	1 unit μL^{-1}	0.25 units
Rest-ligation product	2	1 : 10	
Total	25		

Table S4. Primers used in polymerase chain reactions

Barcode sequences shown are example sequences. Multiple primers with different 7-bp and 6-bp sequences were used (see text)

Name	Ion Torrent key sequence (A or P1)	Barcode sequence	Adaptor sequence (selective bases in red)
*Eco*RI +A primer (preselective)			TACTGCGTACCAATTCA
*Mse*I +C primer (preselective)			GACGATGAGTCCTGAGTAAC
*Ion*A *Eco*R1 +1 (selective)	CCATCTCATCCCTGCG TGTCTCCGACTCAG	TGTAGTG	TACTGCGTACCAATTCA
*Ion*A *Eco*R1 +2 (selective)	CCATCTCATCCCTGCG TGTCTCCGACTCAG	TAGCTGC	TACTGCGTACCAATTCAC
*Ion*A *Eco*R1 +3 (selective)	CCATCTCATCCCTGCG TGTCTCCGACTCAG	AGACGTC	TACTGCGTACCAATTCACC
*Ion*P1 *Mse*I +1 (selective)	CCTCTCTATGGGCAGT CGGTGAT	AGCACG	GACGATGAGTCCTGAGTAAC
*Ion*P1 *Mse*I +2 (selective)	CCTCTCTATGGGCAGT CGGTGAT	TCTCTG	GACGATGAGTCCTGAGTAACA
*Ion*P1 *Mse*I +3 (selective)	CCTCTCTATGGGCAGT CGGTGAT	ACATCG	GACGATGAGTCCTGAGTAACAC
*Ion*P1 *Mse*I +4 (selective)	CCTCTCTATGGGCAGT CGGTGAT		GACGATGAGTCCTGAGTAACACA

Selective amplification

The second PCR utilised fusion primers, consisting, in addition to the adaptor sequence, a 7- or 8-bp barcode sequence and the *Ion*A key sequence (on the *Eco*R1 adaptor sequence end) and a 6-bp barcode and the *Ion*P1 key sequences (on the *Mse*I adaptor sequence end; Table S4). A different 7- or 8-bp barcode sequence on the *Eco*R1 primer was used for each individual in the experiment, whereas alternate 6-bp barcodes on the *Mse*I primers were primarily used to test for cross-contamination between and within experiments. Generally, the same *Mse*I primer was used in a single experiment. As a unique barcodes for each *Eco*R1 primer were needed, all of our experiments were performed with *Eco*R1 +AC selective primers only. We had 96 uniqely barcoded *Eco*R1 +AC primer to our availability; increments and reductions of selectivity were performde with the IonP1 *Mse*I primers only.

To reduce complexity (i.e. the number of different markers amplified in the procedure), primers had an additional 2–4 basepairs at the 3′ end (Table S4). Selective amplifications were run in 25-μL reactions with AmpliTaq Buffer, dNTPs, $MgCl_2$, fusion-selective primers, Amplitaq Gold DNA polymerase, and the diluted preselective-amplification product. Because a different *Ion*A *Eco*R1 primer was added to each sample, this primer was not included in the initial mixture. For efficiency for large experiments, a plate of barcode IonA *Eco*R1 primers was prepared ahead of time and added to the plate of reactions before the preselective product, using a multichannel pippetter. The primer plate was used a maximum of five times and then discarded. Care was taken to avoid contamination (e.g. filtered tips, changing gloves between steps), and alternate barcoded IonP1 *Mse*I primers for each separate experiment were used to monitor cross-contamination. The thermal-cycling parameters were split as

follows: 95°C for 9 min, followed by 5 cycles of 95°C for 30 s, 56°C for 30 s, and 72°C for 45 s, then 15 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s, ending with a 72°C extension for 7 min (Table S5).

Table S5. Reaction conditions used in selective amplification

Reagent	Volume per sample (μL)	Stock concentration	Final concentration
Molecular grade water	13.55		
GeneAmp 10× PCR Buffer I	2.5	10×	1×
dNTP	2	2.5 mM each	200 μM each
$MgCl_2$	2.5	25 mM	2.5 mM
*Ion*P1 *Mse*I primer	0.75	10 μM	300 nM
*Ion*A *Eco*R1 +AC primer	1.5	5 μM	300 nM
Amplitaq Gold polymerase	0.2	5 units μL^{-1}	1 unit
Diluted preselective product	2		
Total	25		

Purification and size selection

All the selective-amplification reactions were pooled (5–10 μL each) and purified using either AMPure XP (Agencourt, Beckman Coulter Inc., Brea, California, USA) or the QIAquick PCR Purification Kit (Qiagen, Venlo, Netherlands). The purified pooled products were quantified using a Qubit 2.0 Fluorometer (Life Technologies) to determine an appropriate dilution for size selection. The pooled reaction mixture was then size-selected using either a Pippin Prep (Sage Science, Beverly, Massachusetts, USA) or an E-Gel (Life Technologies) to maximise the products at the appropriate size range for Ion Torrent (Life Technologies) sequencing, namely, 150–250 bp for 200-bp chemistry, ~250–450 for 400-bp chemistry (once Ion Torrent keys are removed, the products would be ~60 bp shorter, and therefore within the appropriate range). Successful experiments were obtained with all of these alternatives; however, we found the most consistent results with the E-Gel system.

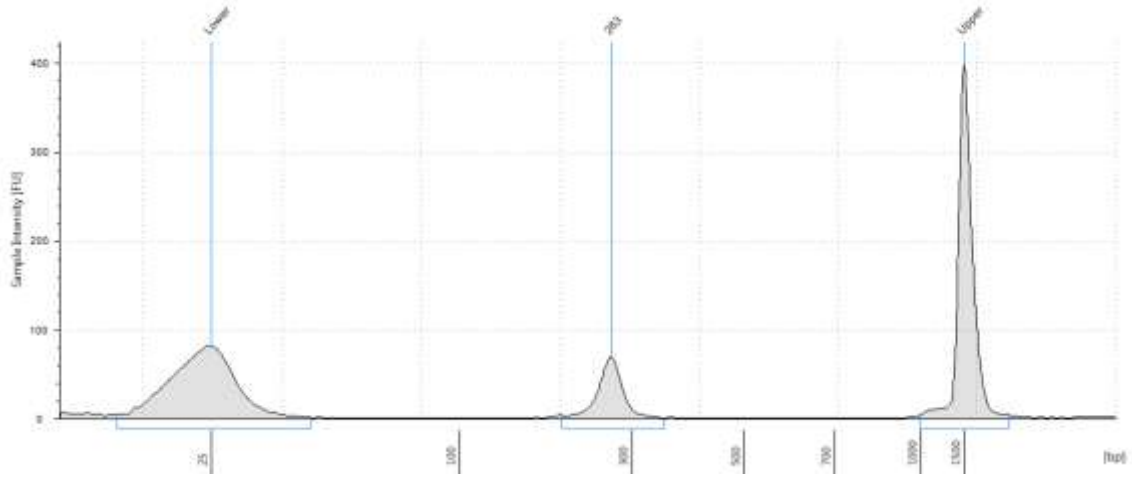
Because of loss of amplicons during the size selection step, especially with the E-Gel system, an additional re-amplification of 5–10 (15) cycles was sometimes performed afterwards, using the Ion Torrent key sequences (*Ion*A and *Ion*P1) as primers. We found that more consistent libraries were created with 15 cycles of selective amplification, followed by 5–10 additional cycles after size-selection. A 2200 TapeStation (Agilent, Santa Clara, California, USA) and a Qubit 2.0 were used to determine the concentration of the pooled and cleaned selective-amplification product, so as to estimate the appropriate dilution for size selection. If the concentration of the product was too high, then usually poor results obtained (too much carry-over of larger and smaller products). A balance was needed between sufficient product for downstream sequencing, but not too much for more accurate size selection. The re-amplification after this step helped strike a proper balance.

For deep sequencing, the size-selected amplification pool was then purified up to three times by using AMPure XP, and then quantified, usually by using both a Qubit 2.0 and a TapeStation with a D1000 high-sensitive ScreenTape (Fig. S1). Accurate quantifications and careful observation of the final size range of amplicons were critical factors for successful sequencing. Too much product could cause

failure of the emulsion PCR or a high number of polyclonal wells in the Ion Torrent sequencing. Likewise, an excessive quantity of small amplicons in the final library would result in highly inefficient sequencing runs dominated by small products and primers. A high number of large products would not be sequenced, but would affect the overall quantification.

The resulting quantifications were then used to calculate the dilution of the library, which was in the range of 9–14 ppm, approximately half to one-third of manufacturer's specifications. Emulsion PCR and enrichment were then conducted on the diluted mixture by using the Ion OneTouch 2 System (Life Technologies) according to manufacturer's specifications, and the enrichment was checked on using the IonSphere Quality Control Kit (Life Technologies) on a Qubit 2.0 system. Next-generation sequencing on the Ion Torrent PGM or Proton sequencing machines followed.

A.



B.

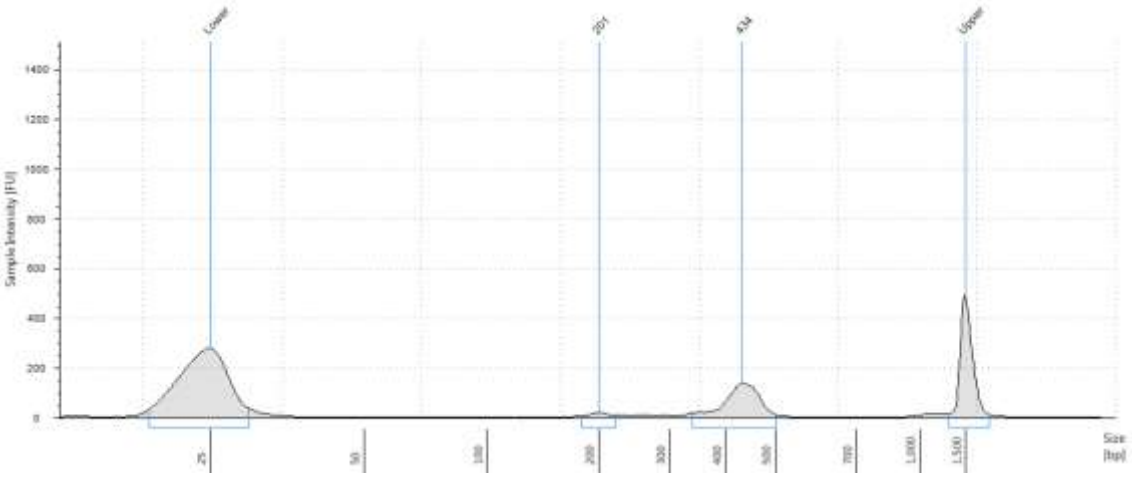


Fig. S1. Example of TapeStation measurements of AFLPseq libraries, indicating (A) ideal range for 200-bp chemistry, and (B) a sample prepared for 400-bp chemistry sequencing with excessive small fragments that must be re-purified.

Analysis of data

Case study 1: analysis of Acacia pinguifolia

For this study, selectivity was set to two additional base pairs (CA) on the *Eco*RI primer (+2) and four additional base pairs (CACA) on the *Mse* primer (+4). An eGel was used for size selection, retaining the fraction in the range of 350–400 bp. Sequencing was undertaken on the Ion Torrent PGM using a 316 v1 chip (Life Technologies). The deep-sequencing run yielded 3.8 million reads after removing ISPs that were polyconal and of low quality. The mean read length was 275 bp and the median length was 326 bp, including the fusion primer sites and barcoding sequences.

CLC Genomic Workbench (Qiagen) was used to demultiplex the samples, trim the primer sequences, and perform a *de novo* assembly. The assembly included only full-length reads (where both barcodes were present). From this assembly (provisional reference genome), 641 reference sequences were selected that had at least 100 reads. All sequence reads (including those for which only the forward barcode was present) were mapped onto the reference sequences in CLC Genomic Workbench. Contigs were extracted, calling ambiguous bases when reads conflicted. Ambiguity bases were called if the least frequent base was present in at least 25% of the reads, using a minimal read depth of 20 reads. The contigs were imported into Geneious (Kearse *et al*. 2012) and renamed (all reads had the name of the tested sample). Then all reads were mapped onto the reference sequences and mappings with fewer than 45 samples were discarded from the analysis. Subsequently, all loci were screened manually for potential single-nucleotide polymorphisms (SNPs). Loci that had a high frequency of SNPs or that had fixed heterozygotes were discarded from the analysis because these where considered signs of paralogy. In total, 55 loci were selected and concatenated into one alignment. SNPs with a frequency of more than 5% were selected for subsequent analyses, resulting in a total of 126 SNPs. Each SNP was subsequently phased into a binary genotype format, calling the ambiguous loci heterozygotes.

With the 126 selected SNPs, short Structure (Pritchard *et al*. 2000) analyses were performed (for K = 1, through to $K = n$; 10 000 burnin and 25 000 iterations), with 20 separate runs for each value of *K*. From these initial runs, the the *ad hoc* delta *K* method (Evanno *et al*. 2005) was used to establish the most likely number of clusters. As expected $K = 2$ was the most likely number of clusters; thus, for $K = 2$, a long run was performed with Structure using default settings (burnin 50 000 and 500 0000 iterations; Fig. S2.

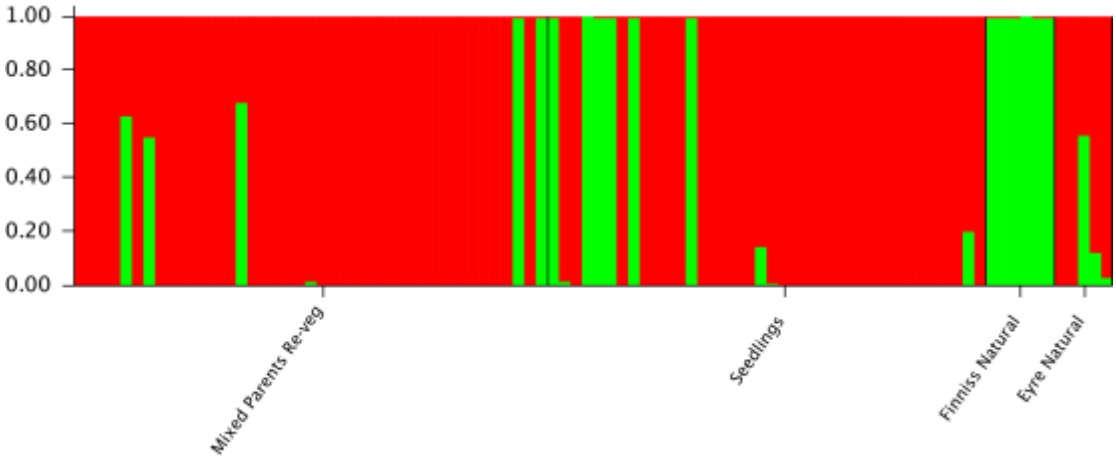


Fig. S2. Structure plot of *Acacia pinguifolia* natural and revegetaged populations, along with seedlings.

Case study 2: Acacia *at and above the species level*

For the *Acacia* species in Case study 2, two datasets were produced, with one using selective polymerase chain reaction of one additional base pair on the *Eco*RI primer (plus-1 dataset) and the other using two additional base pairs on *Eco*RI (plus-2 dataset); both datasets used two additional base pairs on the *Mse* primer. Size selection, purification of products, quantification and sequencing were as described earlier in Supplementary material and in the Materials and methods section of the main article. The datasets were run on two separate Ion Torrent 316 chips.

The FASTX Barcode Splitter (FastX Toolkit, Command Line version, http://hannonlab.cshl.edu/fastx_toolkit) was used to demultiplex the reads into separate files according to their barcode sequence. Then, the FastX toolkit and Cutadapt (Martin 2011) was utilised in a bash script to process the sequences with the following steps: from each read, forward primers were trimmed, keeping a minimum length of 50 base pairs, and each read was renamed; then each sequence was reverse-complemented and the reverse primers (now facing forward) were trimmed; only reads where both primers were identified were kept; reads were then reverse-complemented back to their original orientation, and a quality filter was applied, keeping only reads where at least 90% of the sequence had a *Q*-value of 20 or higher. After filtering, there were 1 280 830 sequences in the plus-1 dataset, with an average of 41 317 sequences (18 917–84 074) per sample, and 2 717 412 sequences, 87 658 (40 545–154 772) per sample, in the plus-2 dataset.

All quality- and length-filtered reads were then converted to fasta files, and the program Pyrad (Eaton 2014) was used to cluster and align the sequences, determine SNPs and output phyml files for downstream phylogenetic analyses. The settings used for Pyrad were as follows: Datatype = gbs (genotype by sequencing), Mindepth (minimum depth of locus) = 3 (with majority base call turned on for low depth loci), Clustering threshold = 0.85, MinCov (minimum number of samples in a final locus) = 3, MaxSH (maximum indels with a shared heterozygous site) = 3, maxH (maximum heterozygous sites in a consensus sequence) = 8, and min length = 50 bp. It is noted that Pyrad has primarily been

used for Illumina data and has not been thoroughly tested with Ion Torrent data. For this reason, a subset of one of the datasets (plus-1) was analysed as in Case study 1 (assembly and mapping with CLC, manual annotation of SNP regions), and the results were nearly identical to the results with Pyrad (albeit with a greater number of loci). Thus, the Pyrad method appears to be promising for Ion Torrent complexity reduction data, although more testing is needed for larger datasets.

A custom python script was then used to convert the RAxML file from DNA-sequence format to numerical representation of all bases and ambiguous calls. The concatanated alignments (alignment length: plus-1 = 704 673 characters, plus-2 = 237 352 characters) were then run with RAxML (Stamatakis 2014) as a multistate dataset with the MULTIGAMMA model, on a cluster with 15 threads and 60 GB of RAM. Twenty initial trees were calculated and 100 bootstrap replicates were run. The resulting phylogenies are shown in Fig. 3 of the main article and Figs S3 and S4.

So as to calculate the number of shared loci between any two samples, the Pyrad .loci output file was used in a custom python script to construct a table of all loci shared between any two samples. This was then plotted against the branch lengths from the RAxML best tree (extracted using the Geneious program ver. 6), using R (see Fig. 4 of the main article).

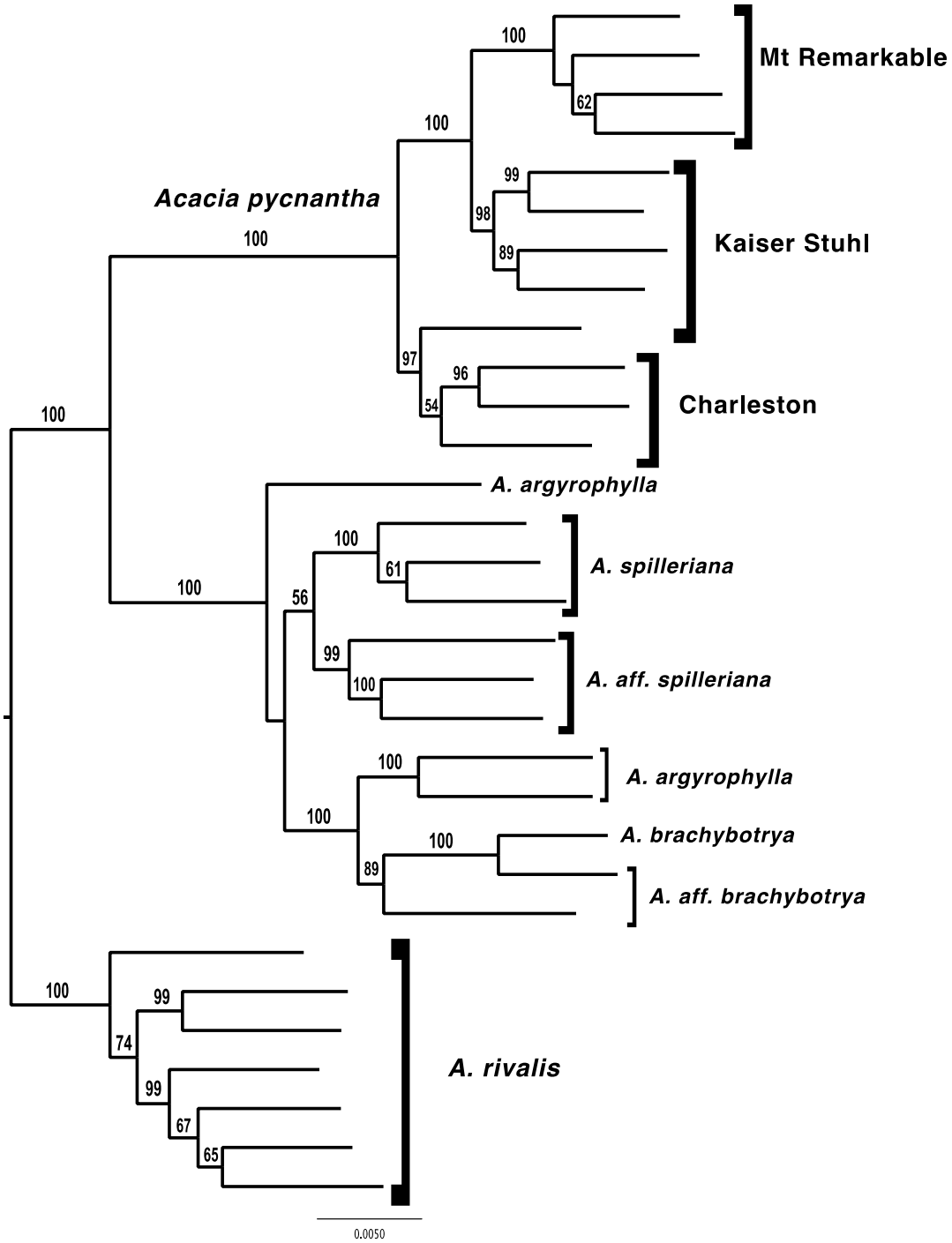


Fig. S3. Phylogeny of Case study 2 *Acacia* plus-1 selective-amplification data, using RAxML. Bootstrap supports of > 50 from 100 replicates are shown above branches.

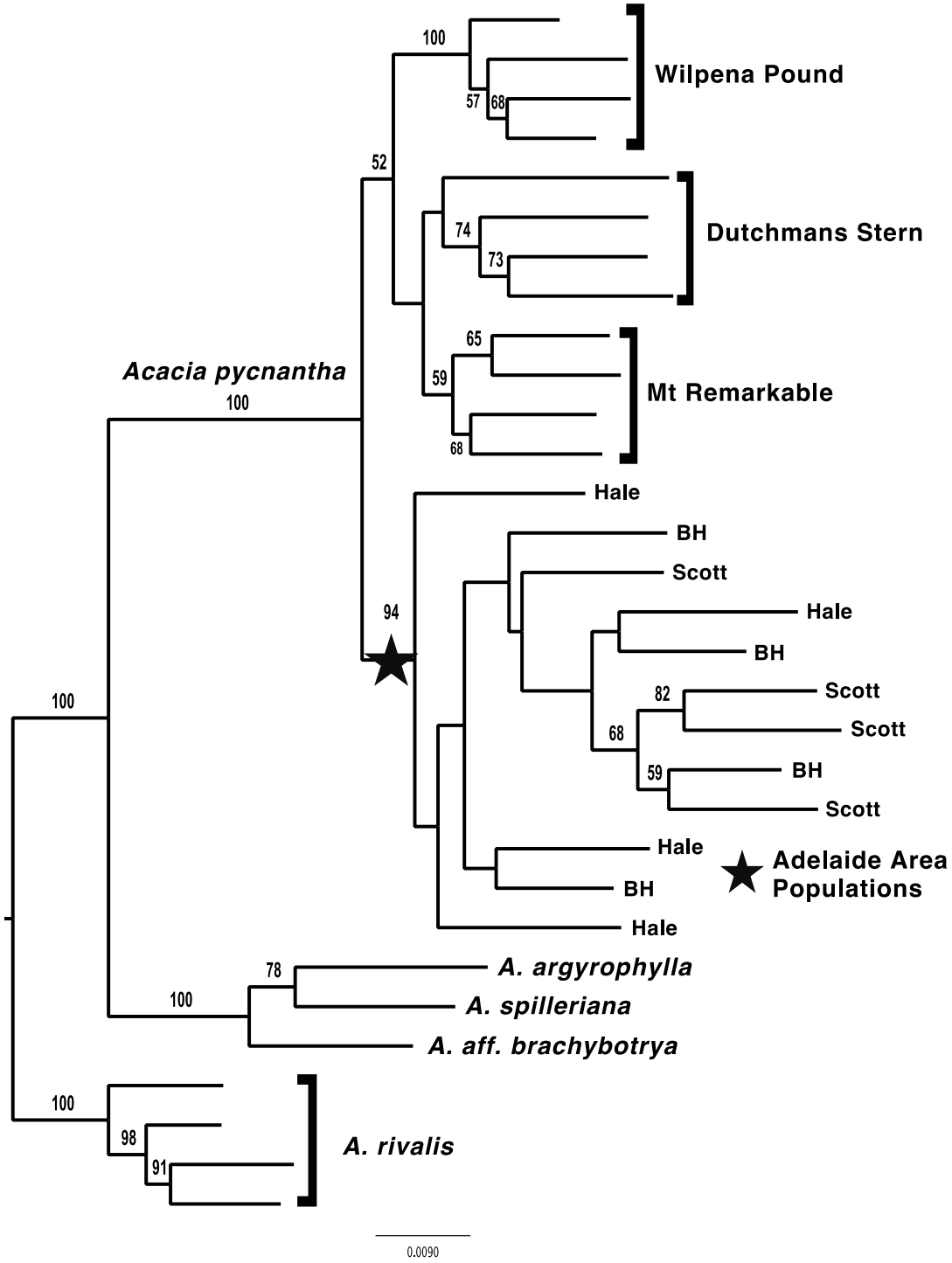


Fig. S4. Phylogeny of Case study 2 *Acacia* plus-2 selective-amplification data, using RAxML. Bootstrap support >50 from 100 replicates are shown above branches.

Case study 3: signal conservation in Myrtaceae

For the *Eucalyptus* case study, two additional base pairs were used on each primer (+2 on the EcoR1 primer and +2 on the Mse primer). Otherwise, the size selection (200 bp), sequencing (PGM library) and initial data processing were as described for Case study 1. The samples of *E. cosmophylla* complex (including *E. paludicola* and *E. ovata*) were analysed using CLC Genomics Workbench as described earlier (but also screened for paralogues by mapping back to draft *E. grandis* genome). Additionally, for all *Eucalyptus* and other Myrtaceae species, reciprocal mapping analyses were performed to determine the shared homology of the subsampled (reduced-representation) genomic sequences and gauge the utility of the technique for systematic analyses with an increasing genetic distance. Sequence data were generated for 11 divergent Myrtaceae species (Table S6) and were processed using CLC Genomics Workbench. For each species, we performed a *de novo* assembly (insertions, deletions, mismatch all set to 2 – equal weighting appears to reduce indel error; and 50% of the read has to be 80% similar) and extracted contigs with coverage greater than 15×. The sequencing reads from all other taxa were mapped back to each set of contigs (settings as above) and the percentage of mapped bases was recorded.

Table S6. Taxonomic classification of Myrtaceae taxa included in the present study

Tribal classification of Myrtaceae follows Wilson *et al*. (2001), whereas the classification of *Eucalyptus sens. str*. is according to Brooker (2000). Also shown are the approximate divergence times ('Div') relative to *E. grandis* (subg. *Symphyomyrtus*) according to Thornhill *et al*. (2015) and the proportion sequencing reads that mapped ('Map') to the *E. grandis* genome

Species	Tribe	Subgenus	Section	Div	Map
Eucalyptus microcarpa	Eucalypteae	*Symphyomyrtus*	*Adnataria*	<30	0.98
Eucalyptus leucoxylon	Eucalypteae	*Symphyomyrtus*	*Adnataria*	<30	0.94
Eucalyptus cladocalyx	Eucalypteae	*Symphyomyrtus*	*Sejunctae*	<30	0.99
Eucalyptus macrorhycha	Eucalypteae	*Eucalyptus*	*Capillulus*	35	0.98
Eucalyptus baxteri	Eucalypteae	*Eucalyptus*	*Capillulus*	35	0.97
Eucalyptus obliqua	Eucalypteae	*Eucalyptus*	*Eucalyptus*	35	0.98
Angophora floribunda	Eucalypteae			55	0.65
Corymbia maculata	Eucalypteae			55	0.7
Leptospermum continentale	Leptospermeae			65	0.59
Calytrix tetragona	Chamelaucieae			65	0.12
Melaleuca orophila	Melaleuceae			70	0.51

References

Brooker MIH (2000) A new classification of the genus *Eucalyptus* L'Her.(Myrtaceae). *Australian Systematic Botany* **13**, 79–148. doi:10.1071/SB98008

Eaton DA (2014) PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844–1849. doi:10.1093/bioinformatics/btu121

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. doi:10.1111/j.1365-294X.2005.02553.x

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 doi:10.1093/bioinformatics/bts199.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17**, 10–12. doi:10.14806/ej.17.1.200

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

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. doi:10.1093/bioinformatics/btu033

Thornhill AH, Ho SY, Külheim C, Crisp MD (2015) Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. *Molecular Phylogenetics and Evolution* **93**, 29–43. doi:10.1016/j.ympev.2015.07.007

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**, 4407–4414. doi:10.1093/nar/23.21.4407

Wilson PG, O'Brien MM, Gadek PA, Quinn CJ (2001) Myrtaceae revisited: a reassessment of infrafamilial groups. *American Journal of Botany* **88**, 2013–2025. doi:10.2307/3558428