Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular

TESE DE DOUTORADO

Desenvolvimento e aplicações de DArT (*Diversity Arrays Technology*) e genotipagem por sequenciamento (*Genotyping-by-Sequencing*) para análise genética em *Eucalyptus*

Carolina Paola Sansaloni

Brasília, Abril de 2012

Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular

Desenvolvimento e aplicações de DArT (*Diversity Arrays Technology*) e genotipagem por sequenciamento (*Genotyping-by-Sequencing*) para análise genética em *Eucalyptus*

Carolina Paola Sansaloni

Orientador: Dr. Dario Grattapaglia

Tese apresentada ao Departamento de Biologia Celular do Instituto de Biologia, da Universidade de Brasília, como requisito parcial para obtenção do grau de Doutor em Ciências Biológicas, Área de Concentração: Biologia Molecular

Brasília, Abril de 2012

TERMO DE APROVAÇÃO

Tese apresentada ao Departamento de Biologia Celular da Universidade de Brasília, como requisito parcial para obtenção de grau de Doutor em Ciências Biológicas, área de concentração Biologia Molecular.

Tese defendida e aprovada em 05/04/2012 por:

Dr. Dario Grattapaglia - Orientador
Embrapa Recursos Genéticos e Biotecnologia

Dr. Marcio Elias Ferreira
Embrapa Recursos Genéticos e Biotecnologia

Dr. Georgios Joannis Pappas Júnior
Embrapa Recursos Genéticos e Biotecnologia

Dr. Alexandre Siqueira Guedes Coelho
Universidade Federal de Goiás

Dr. Evandro Novaes
Universidade Federal de Goiás

Membro suplente:
Dr. Márcio de Carvalho Moretzsohn (Embrapa Recursos Genéticos e Biotecnologia)

A meus grandes amores

César e nossa princesa Isabella

DEDICO

AGRADECIMENTOS

A realização dessa tese somente foi possível pelo apoio recebido de diversas pessoas e instituições, mas gostaria de agradecer de maneira especial:

À Universidade de Brasília (UnB), e especialmente ao programa de pós-graduação em Biologia Molecular, representados por todos os professores e funcionários, pela oportunidade de realização deste curso.

À Coordenação de Aperfeiçoamento de Pessoal de Nivel superior (CAPES) pelo apoio financeiro que tornou possível a realização desse trabalho.

À EMBRAPA Recursos Genéticos e Biotecnologia, pela infra-estrutura de trabalho.

À todas as empresas florestais e Instituições que facilitaram as amostras e os recursos necessários para a realização deste trabalho.

Devo muito à orientação do Dr. Dario Grattapaglia, com quem estabeleci uma relação profissional e pessoal, a qual espero seja duradoura após a finalização deste trabalho. É em pessoas como ele que me espelho e busco inspiração para o meu próprio desenvolvimento profissional e pessoal.

Ao Dr. Andrzej Kilian, director da empresa *Diversity Arrays Technology*, por ter acreditado no projeto e aberto as portas do DArT para recebernos com muito carinho em Canberra, Australia. Pelos grandes conhecimentos transmitidos que são os pilares para o meu futuro profissional.

Aos membros da banca de avaliação desta tese, o Dr. Marcio Elias Ferreira, Dr. Georgios J. Pappas Júnior, Dr. Alexandre Siqueira Guedes Coelho e o Dr. Evandro Novaes pelos comentários, sugestões e questionamentos que foram essenciais para a melhoria e o estabelecimento do formato final.

Aos pesquisadores e técnicos do Laboratorio de Genética Vegetal da Embrapa Recursos Genéticos e Biotecnología, Marcão, Marcio, Vânia, Glaucia, Zilneide, Lorena, Peter e muito especialmente a minha grande amiga e conselheira Marília.

Aos amigos que tive o prazer de conhecer no Laboratorio de Genética Vegetal, Bruna (e Daniel), Thaisa, Marília (Georgios e meu príncipe Nicolas), Dione (Andrê e Ian), Ediene e Rodrigo, Tati (Bruno e Arthur), Natália, Mariana, Marco (Bia e Teo), Tulio (e Cecilia), Pedro e Dani pela amizade de tantos anos. A muitos de vocês considero irmãos do coração que

a vida me presenteu e tenho certeza que nem a distância nem o tempo vai destruir essa amizade tão linda e verdadeira.

A minha familia Australiana, Colleen, Michael, Kasia, Grzegorz, Jason, Ling, Eric, Damian, Puthick, Vanessa, Cina, Cleare, Frank, Gosia, Hang e Adriane, por todo o apoio e carinho durante minha estadía em Canberra, Australia, onde passei dois anos maravilhosos da minha vida, cheio de experiências inesquecíveis.

Aos meus pais, Norma e Juan, os principais responsáveis por minha educação e formação como pessoa. Sou eternamente agradecida pelos ensinamentos, apoio incondicional e incentivo para meu crescimento pessoal e profissional. Tudo o que eu sou eu devo a vocês.

Ao meu irmão Adrián e sua esposa Sole pelo apoio, carinho e principalmente por ter me presenteado com três princesas maravilhosas: Valen, Flor e Vicky, amos demais vocês!

À família que me acolheu como filha: Cacho, Elena, Gastón, David, Sonia, Sofi, Tizi, tios e primos, pelo carinho e apoio recebido sempre.

Ao César, meu grande amor e companheiro de vida, pelo seu apoio, amizade, compreensão e amor incondicional. Este trabalho é compartilhado 100% com você e é isso que o faz tão especial. Obrigada por ser o melhor esposo do mundo e por ter cumprido meu grande sonho de formar uma família maravilhosa juntos.

E finalmente a pessoa que conseguiu mudar completamente o sentido da minha vida, minha princessa Isabella. Nunca imaginei que meu coração tinha tanto amor para dar. Cada dia com um simples sorriso você me faz a mamãe mais feliz do mundo. Te amo demais Pipi!

ÍNDICE

LISTA DE FIGURAS..... x

LISTA DE TABELAS..... xi

RESUMO..... 1

ABSTRACT..... 2

1. INTRODUÇÃO..... 4

1.1. Biologia do gênero *Eucalyptus*..... 4

1.2. Importância econômica, ambiental e social do *Eucalyptus* no Brasil..... 5

1.3. Base genética das principais metodologias de marcadores moleculares 6

1.3.1. Polimorfismos de longitud de fragmentos de restrição ou RFLPs..... 6

1.3.2. DNA Polimórfico Amplificado ao Acaso ou RAPD 6

1.3.3. Polimorfismo de Comprimento de Fragmentos Amplificados ou AFLP..... 8

1.3.4. Microssatélites ou SSRs..... 9

1.3.5. Polimorfismo de base única ou SNPs..... 10

1.4. Aplicações de marcadores moleculares em *Eucalyptus*..... 11

1.4.1. Fingerprinting e analise de diversidade genética..... 11

1.4.2. Genética de populações e filogenia..... 12

1.4.3. Mapas genéticos 13

1.4.4. Mapeamento de QTLs e seleção assistida por marcadores..... 13

1.5. Novas metodologias de marcadores moleculares para análise “genome-wide” 14

1.5.1. DArT (Diversity Arrays Technology..... 15

1.5.2. Genotipagem por sequenciamento de representações DArT ou GbS..... 19

2. Objetivos..... 20

3. CAPÍTULO 1. DESENVOLVIMENTO E VALIDAÇÃO DE UM MICROARRANJO DArT PARA GENOTIPAGEM DE *Eucalyptus*..... 21

3.1. INTRODUÇÃO..... 21

3.2. MATERIAIS E MÉTODOS..... 22

3.2.1. Material biológico utilizado..... 22

3.2.2. Redução de complexidade genômica 22

3.2.3. Construção de bibliotecas piloto de clones DarT..... 24

3.2.4. Preparação de amostras ou “targets” para hibridização..... 25

3.2.5. Hibridização dos targets ao microarranjo..... 26
3.2.6. Lavagem e digitalização dos slides..... 27
3.2.7. Processamento das imagens e declaração dos genótipos..... 28
3.2.8. Expansão das representações genômicas, analise de redundância e sequenciamento de clones DarT..... 31
3.2.9. Analise de validação do microarranjo operacional..... 32
3.3 RESULTADOS E DISCUSSÃO..... 32
3.3.1 Redução da complexidade genômica..... 34
3.3.2 Triagem das baterias piloto de clones DArT 35
3.3.3 Triagem da primeira bateria de clones DArT para seleção de clones polimórficos..... 36
3.3.4 Triagem da segunda bateria de clones DArT para seleção de clones polimórficos e montagem do microarranjo DArT operacional 39
3.3.5. Validação da plataforma de genotipagem DArT para *Eucalyptus* em estudos de diversidade e filogenia..... 44
3.3.6. Validação da plataforma de genotipagem DArT para mapeamento genético em *Eucalyptus*..... 53
3.4. CONCLUSÕES..... 54
4. CAPÍTULO 2. AVALIAÇÃO DA METODOLOGIA DE GENOTIPAGEM POR SEQUENCIAMENTO DART-SEQ PARA MAPEAMENTO GENÉTICO EM *Eucalyptus*... 59
4.1. INTRODUÇÃO..... 59
4.2. MATERIAL E MÉTODOS..... 61
4.2.1. Material vegetal..... 61
4.2.2. Análise de locos microssatélites e verificação de parentesco..... 62
4.2.3. Genotipagem com o microarranjo operacional DarT..... 63
4.2.4. Genotipagem por seqüenciamento..... 63
4.2.4.1. Redução da complexidade genômica e ligação de adaptadores..... 65
4.2.4.2. Amplificação dos targets..... 67
4.2.4.3. Agrupamento, purificação e quantificação dos targets..... 67
4.2.4.4. Geração de clusters em sistema cBotTM (Illumina) e sequenciamento no Genome Analizer GAII Illumina..... 68

4.2.4.5. Processamento dos dados brutos e alinhamento das sequências de DNA. 69
4.2.4.6. Seleção de marcadores para análise de mapeamento genético..... 68
4.2.4.7. Construção do mapa de ligação de alta densidade..... 70
4.3. RESULTADOS E DISCUSSÃO 71
4.3.1. Genotipagem DArT baseada em microarranjo..... 71
4.3.2. Genotipagem DArT-Seq baseado em NGS..... 71
4.3.3. Detecção de polimorfismos de marcadores..... 72
4.3.3.1. Método de redução da complexidade *PstI_Taq_ad_HpaII*..... 76
4.3.3.2. Método de redução da complexidade *PstI_Taq_ad_HhaI*..... 76
4.3.4. Mapeamento genético..... 77
4.4. Conclusões e perspectivas..... 84
5. REFERÊNCIAS BIBLIOGRÁFICAS..... 86

ANEXO I. Cópia do artigo publicado:

Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D, Vaillancourt RE, Kilian A (2010) A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. Plant Methods 6:16

ANEXO II. Cópia do artigo publicado:

Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J, Kilian A, Myburg AA, Grattapaglia D, Vaillancourt RE (2011) Population genetic analysis and phylogeny reconstruction in Eucalyptus (Myrtaceae) using high-throughput, genome-wide genotyping. Mol Phylogenet Evol 59:206-224

ANEXO III. Cópia do artigo publicado:

Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A (2011) Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proceedings 5:P54

LISTA DE FIGURAS

Figura 1. Fluxograma esquemático da metodologia DArT..... 18

Figura 2. Representação esquemática do processo de redução da complexidade genômica..... 23

Figura 3. Representação esquemática da hibridização dos *targets*..... 26

Figura 4. Visão geral da rotina do proceso de hibridização para uma placa de 96 poços contendo amostras de DNA genômico..... 27

Figura 5. Slide do microarranjo de *Eucalyptus* detectado com diferentes combinações laser/filtro..... 28

Figura 6. Interpretação de dados calculado pelo programa *DArTSoft* mostrando a diferença de intensidade entre genótipos..... 29

Figura 7. Exemplo de tabela de Microsoft Excel exportada pelo programa *DArTSoft*..... 31

Figura 8. Fluxograma das etapas de desenvolvimento do microarranjo de genotipagem DArT para *Eucalyptus*..... 33

Figura 9. Resultado das sete combinações de enzimas de restrição testadas para a redução da complexidade genômica..... 34

Figura 10. Dendrograma Neighbor Joining construído com 7.960 marcadores polimórficos, mostrando o posicionamento de. *E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis e E. cladocalyx* do gênero *Eucalyptus* e *Corymbia variegata* do género *Corymbia*..... 47

Figura 11. Dendrograma Neighbor Joining construído com 7.043 marcadores DArT polimórficos, mostrando o posicionamento de sete espécies do gênero *Eucalyptus* (*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis e E. cladocalyx)*..... 48

Figura 12. Dendrograma Neighbor Joining construído com 5.033 marcadores polimórficos, mostrando o posicionamento de 58 amostras de cinco espécies do subgênero *Symphyomyrtus* (*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis)*..... 49

Figura 13. Dendrograma Neighbor Joining construído com 4.302 marcadores polimórficos, mostrando o posicionamento de 24 amostras das espécies *E. urophylla* e *E. grandis*, ambas pertencentes ao subgênero *Symphyomyrtus* e mesma seção

Latoangulatae..... 50
Figura 14. Dendrograma Neighbor Joining construído com 3.565 marcadores polimórficos, mostrando identificação individual de 12 amostras de *E. grandis*..... 51
Figura 15. Dendograma Neighbor Joining representando as relações filogenéticas entre 62 espécies de *Eucalyptus*..... 52
Figura 16. Fluxograma do procedimento DArT-Seq baseado em NGS (Next Generation Sequencing)..... 64
Figura 17. Sequências indexadoras ou barcodes. (a) sequência *forward* como adaptador (vermelho), o *barcode* de 4 a 8 pb (azul) e o sitio *PstI* (verde). (b) sequência *reverse* incluindo o adaptador (vermelho) e o *barcode* (azul). (c) sequência adaptador-*barcode* depois do anelamento, mostrando o iniciador de PCR no alinhamento (negro)..... 65
Figura 18. Sequências do adaptador comum. (a) sequência *forward* do adaptador (vermelho); (b) sequência *reverse*do adaptador (vermelho) e duas bases complementares ao sitio de reconhecimento da enzima *HpaII* (azul). (c) sequência do adaptador comum depois do anelamento, mostrando o iniciador de PCR no alinhamento (preto)..... 66
Figura 19. Representação esquemática da *flowcell* de uma corrida de sequenciamento de GAIIX..... 69
Figura 20. Controle de qualidade de *targets*..... 72
Figura 21. Controle de qualidade das representações genômicas após a purificação..... 73
Figura 22. Distribuição do número de marcadores mapeados segundo a segregação dos parentais em cada um dos 11 grupos de ligação..... 80
Figura 23. Mapa de ligação da família BRASUZ1..... 82

LISTA DE TABELAS

Tabela 1. Bibliotecas construídas e utilizadas para a triagem do primeiro microarranjo protótipo de marcadores DArT..... 36

Tabela 2. Espécies de *Eucalyptus* e número de indivíduos de cada espécie (total de 284) utilizados como *targets* para a triagem do primeiro microarranjo protótipo visando a seleção de marcadores polimórficos..... 37

Tabela 3. Distribuição de clones DArT polimórficos dentro de cada classe de bin no primeiro microarranjo protótipo de triagem..... 38

Tabela 4. Resultado da análise de redundância de clones DArT com base no sequenciamento de clones amostrados em nove *bins* não redundantes declarados com base em distâncias *Hamming*..... 39

Tabela 5. Origem dos clones DArT amostrados para a segunda bateria de triagem..... 40

Tabela 6. Populações de *Eucalyptus* e o correspondente número de indivíduos utilizados como *target* para a triagem da segunda bateria de 7.680 clones..... 41

Tabela 7. Distribuição de clones DArT polimórficos dentro de cada classe de bin da segunda bateria de clones submetidos à triagem..... 42

Tabela 8. Distribuição e origem de *bins* a partir dos quais clones DArT foram selecionados para integrar o microarranjo operacional DarT..... 43

Tabela 9. Resultados do desenvolvimento do microarranjo DArT em *Eucalyptus* na fase protótipo e operacional incluindo a triagem de marcadores polimórficos e análise de redundância..... 44

Tabela 10. Análise de polimorfismo em painéis de diversidade decrescente de espécies..... 45

Tabela 11. Número de marcadores polimórficos identificados em cada população de mapeamento de *Eucalyptus* (diagonal) e entre populações de mapeamento (acima da diagonal). 54

Tabela 12. Poder informativo dos marcadores DArT do microarranjo operacional para mapeamento genético, baseado na amostragem de seis populações de mapeamento (informação dos pedigrees na Tabela 10)..... 54

Tabela 13. Resumo dos filtros utilizados para a seleção de reads para análise. Para a região do *barcode* o parâmetro de filtro utilizado foi: 75% da sequência com um valor Q Phred ≥30. Para a qualidade da sequência completa: 50% da sequência com um valor Q Phred ≥10..... 74

Tabela 14. Distribuição do número de *tags* alinhados com ambos os métodos de redução de complexidade sobre o genoma de referência de *Eucalyptus*..... 75

Tabela 15. Distribuição do número de posições onde um *tag* único alinha-se no genoma de referência de *Eucalyptus* com os dois métodos de redução de complexidade..... 76

Tabela 16. Distribuição no genoma de referência de *Eucalyptus* dos marcadores polimórficos

DArTs e *insilico* DArTs de ambos os parentais selecionados para mapeamento..... 78

Tabela 17. Estatística do mapa de ligação da família BRASUZ1 construido com o software da DArT com base em distância Hamming..... 79

RESUMO

Espécies de *Eucalyptus* tem sido utilizadas com sucesso para plantios florestais devido ao seu rápido crescimento, sua capacidade de adaptação às diversas condições edafo-climáticas e pelo seu potencial econômico na produção de energia, fibra e madeira sólida, reduzindo assim a pressão sobre as florestas tropicais e a biodiversidade associada. Marcadores moleculares tais como RAPD, AFLP, microssatélites, e mais recentemente SFP e SNPs têm contribuído para a caracterização e conservação dos recursos genéticos do gênero *Eucalyptus*, auxiliando também na compreensão da evolução do gênero, além de permitir a construção de mapas de ligação e identificação de QTLs (*Quantitative trait loci*). Entretanto, estas técnicas são lentas, laboriosas, apresentam limitações de cobertura genômica e envolvem custos elevados para a análise de muitos indivíduos. *Diversity Arrays Technology* (DArT) é um método baseado em hibridização que permite genotipar centenas a milhares de marcadores num simples ensaio. Esta tecnologia gera um perfil genômico com um alto rendimento e um grande poder de transferibilidade entre espécies. Neste trabalho é apresentado o desenvolvimento da primeira plataforma de genotipagem de alto desempenho de marcadores DArTs em microarranjo para o gênero *Eucalyptus*, e demonstrada sua eficiência em estudos de diversidade genética, filogenia e mapeamento genético. Foram desenvolvidas 18 bibliotecas genômicas de complexidade reduzida a partir de 64 espécies diferentes do gênero. Um total de 23.808 fragmentos de DNA foram avaliados para revelação de polimorfismos DArT, e 13.300 (56%) destes declarados polimórficos entre um painel de triagem composto por 284 indivíduos. Destes, 7.680 marcadores foram selecionados para a construção de um microarranjo de genotipagem para uso em rotina. Em um estudo de diversidade intra-específica, 4.752 marcadores foram polimórficos e 5.013 mostraram segregação mendeliana em seis populações segregantes não relacionadas, com uma média de 2.211 marcadores polimórficos por população. Na etapa seguinte do trabalho, foi otimizada a tecnologia de genotipagem por sequenciamento DArT-seq para a construção de um mapa genético de alta densidade para uma população segregante do cruzamento entre árvores elite de *E. grandis* (BRASUZ1 x M4D31). A população foi genotipada com o microarranjo DArT e com a técnica DArT-seq. Enquanto o microarranjo DArT forneceu 1.088 marcadores, a genotipagem DArT-seq forneceu 2.449. No total, um mapa de ligação integrado por 564 marcadores DArT, 1.930 marcadores DArT-seq e 29 microssatélites foi construído. Além destes marcadores, mais de 1.500 SNPs derivados da metodologia DArT-seq foram obtidos proporcionando uma vantagem adicional pela inclusão de marcadores co-dominantes no mapa. O desenvolvimento de metodologias de genotipagem por sequenciamento (GbS) via enzimas de restrição como DArT-seq ou captura com sondas, representa uma aplicação adicional das tecnologias de "*next generation sequencing*" além do sequenciamento,

potencializando a análise genética com marcadores moleculares. A combinação do elevado número de marcadores, baixo custo, metodologia relativamente accessível e uso de reagentes universais, aponta para um uso crescente de GbS nos próximos anos nas mais diversas aplicações em estudos de genética de populações, investigações evolutivas e em apoio ao melhoramento acelerando e aumentando a precisão da seleção direcional de características multifatoriais complexas.

ABSTRACT

Species of Eucalyptus have been successfully used for forest plantations due to its rapid growth, its ability to adapt to various soil and climatic conditions and its economic use in energy, fiber and solid wood, reducing pressure on tropical forests and associated biodiversity. Molecular markers such as RAPD, AFLP, microsatellites, and more recently SFP and SNPs have contributed to the characterization and conservation of genetic resources of the genus, to the understanding of the evolution of the genus, and allowing the construction of linkage and QTLs maps. However, these techniques are slow, laborious, provide limited genome coverage and are costly for the analysis of large sample sizes. Diversity Arrays Technology (DArT) is a hybridization-based method that allows genotyping hundreds to thousands of markers in a single assay. This technology generates a genomic profile with high throughput and transferability between species. This study presents the development of the first high throughput genotyping platform for species of *Eucalyptus* based on a DArT microarray and demonstrates its use for diversity, phylogeny and mapping studies. A total of 18 reduced representation genomic libraries from 64 different species of the genus were developed. A total of 23,808 DNA fragments were screened for polymorphism and 13,300 (56%) of them declared polymorphic in a panel of 284 individuals. Out of these, 7,680 markers were selected to populate a routine DArT genotyping microarray. In an inter-specific diversity study, 4,752 were deemed polymorphic while 5,013 showed Mendelian segregation when assessed in six inter-specific mapping pedigrees, with an average of 2,211 polymorphic markers per pedigree. The subsequent step of the study, involved the optimization of the genotyping-by-sequencing technology called DArT-seq to construct a high density genetic map for a segregating population derived from the *E. grandis* elite trees (BRASUZ1 x M4D31). The population was genotyped both with the DArT microarray and with DArT-seq. While the DArT microarray yielded 1,088 markers, the DArT-seq method supplied 2,449 markers. In total, an integrated linkage map with 564 DArT markers, 1,930 DArT-NGS markers and 29 microsatellites was built. Besides these mapped markers, an additional set of over 1,500 SNPs derived from DArT-seq were scored providing an additional advantage by the inclusion of co-dominant markers on the map. The development of genotyping by sequencing (GBS)

methods via restriction enzymes as DART-seq or capture probes, represents a further application of the technologies of "next generation sequencing" beyond sequencing, empowering the genetic analysis with molecular markers. The combination of the large number of markers, low cost, relatively accessible methodology and use of universal reagents, points to an increased use of GBS in the coming years in several applications in studies of population genetics, evolutionary investigations and in support to plant breeding accelerating and increasing accuracy of directional selection for complex multi-factorial traits.

1. INTRODUÇÃO

1.1. Biologia do gênero *Eucalyptus*

O gênero *Eucalyptus* L´Herit pertence à família *Myrtaceae*. Possui mais de 700 espécies distribuídas em oito subgêneros, sendo o principal deles o *Symphyomyrtus*, com mais de 300 espécies, dentre as quais estão as seis espécies mais plantadas para fins comerciais *E.grandis, E. globulus, E. urophylla, E. camaldulensis, E. saligna* e *E. tereticornis* (Eldridge, Davidson *et al.*, 1994; Brooker, 2000; Poke, Vaillancourt *et al.*, 2005). O gênero é endêmico da Austrália, com exceção de *E. urophylla* e *E. deglupta,* que são naturais do Timor e Papua Nova Guiné, respectivamente (Keane, Kile *et al.*, 2000; Potts e Pederick, 2000). As extremas diferenças climáticas existentes de norte a sul deste vasto continente, com as variações de altitude e solo, tem resultado em uma imensa diversidade de habitats para as quais o *Eucalyptus* tem-se adaptado muito bem. Este gênero mostra uma excepcional diferenciação e ampla variabilidade genética. Esta variação fornece a base para programas de melhoramento de espécies adaptadas a um amplo espectro de ambientes, inclusive no Brasil.

Florestas plantadas de *Eucalyptus* são conhecidas por seu rápido crescimento, forma reta, qualidade superior da madeira para múltiplas aplicações, ampla adaptabilidade a solos e climas e facilidade de manejo por plantio direto e rebrota (Eldridge, Davidson *et al.*, 1994; Potts, 2004). Espécies de *Eucalyptus* são atualmente plantadas em mais de 90 países onde são utilizadas para diversos produtos florestais tais como madeira sólida, postes, energia, celulose, carvão vegetal, óleos essenciais, mel e tanino, bem como para sombra em parques e jardins (Doughty, 2000).

A partir do século 18, diversas espécies do gênero *Eucalyptus* foram introduzidas em países como Índia, França, Chile, Brasil, África do Sul e Portugal onde apresentaram excelente adaptação climática, sucesso reprodutivo e índices produtivos elevados (Potts, 2004). A partir da década de 60, com o desenvolvimento de métodos industriais de processamento das fibras curtas do *Eucalyptus*, as plantações de espécies do gênero começaram a apresentar importância comercial crescente e os programas de melhoramento tiveram início em países como Estados Unidos (*E. grandis*) e Portugal (*E. globulus*) (Eldridge, Davidson *et al.*, 1994). Em 1970 iniciaram-se as plantações clonais no Congo e no Brasil com destaque para a espécie *E. urophylla,* principalmente em função da maior facilidade de propagação vegetativa desta espécie. A partir de 1980, populações base de melhoramento foram formadas, envolvendo principalmente as espécies *E. grandis*, *E. tereticornis* e *E. viminalis* (Eldridge,

Davidson *et al.*, 1994). No começo da década de 90, o Brasil já era o país com maior acervo genético de *Eucalyptus*, atrás somente da Austrália e Indonésia (Ferreira, 1992). Isto evidencia o amplo interesse da indústria de base florestal pelos recursos oferecidos por este gênero, envolvendo grandes esforços no planejamento de programas de melhoramento que permitam obter com grande eficácia produtos derivados da madeira para abastecer a crescente demanda nacional e internacional.

1.2. Importância econômica, ambiental e social do *Eucalyptus* no Brasil

Plantios de florestas sustentáveis de espécies do gênero *Eucalyptus* de rápido crescimento suprem, de modo racional e eficaz, a demanda por biomassa lenhosa de alta qualidade nas mais diversas condições ambientais constituindo uma das principais fontes de matéria prima florestal sustentável no planeta. Apesar da forte vocação florestal do Brasil, somente cerca de cinco milhões de hectares de floresta são plantados hoje (Abraf, 2011). Porém, em relação a 2009, a área de plantios florestais aumentou um 3,2 %. No período 2005‑2010, o crescimento acumulado foi de 23,0%. A expansão dos plantios florestais em 2010 pode ser considerada modesta, quando comparada com o período 2005‑2009 (4,5% a.a.) (Abraf, 2011). Este crescimento da área plantada com este gênero é consequência da necessidade de suprir a demanda de madeira para produção de celulose e carvão vegetal para indústria siderúrgica, assim como para a construção civil, produção de móveis, papelão, óleos e outros (Mora e Garcia, 2000).

O papel central das florestas plantadas na economia de base florestal é reconhecido mundialmente. Somente no ano 2010 as exportações brasileiras de produtos de florestas plantadas atingiram US$ 7,5 bilhões, um aumento de 34% em relação ao ano anterior e de mais de 40% comparado com 2005. No Brasil, 37,5% de toda a madeira produzida é utilizada para a extração de celulose. Nos últimos dez anos observou-se um crescimento anual de 5,9% na produção deste material, atingindo mais de 14 milhões de toneladas no ano 2010. Estes valores colocaram o Brasil no quarto lugar do ranking mundial entre os produtores de celulose. De acordo com a Bracelpa (Associação Brasileira de Celulose e Papel), as empresas brasileiras produziram 9,8 milhões de toneladas de papel no ano 2010, aumentando em 5,4% a produção em relação ao ano anterior; 5,4 milhões de toneladas abasteceram o consumo do mercado doméstico e as exportações de papel chegaram a um total de 2,07 milhões de toneladas, um aumento do 3,3% desde 2009. As florestas plantadas colaboram também com a produção de energia renovável através do carvão vegetal, que hoje é obtido de espécies nativas em 55,0% e é direcionado principalmente para a indústria

siderúrgica. Atualmente o setor de florestas gera 4,7 milhões de empregos, incluindo empregos diretos, empregos indiretos e empregos resultantes do efeito‑renda, gerando 600.000 novos postos de trabalho nos últimos cinco anos, o que manifesta a importância de este mercado como um grande suporte na criação de fontes de emprego no Brasil (Abraf, 2011).

Produtividades florestais crescentes e refinamentos na qualidade dos produtos de madeira por meio de melhoramento genético tornar-se-ão cada vez mais estratégicos para a indústria florestal, independentemente do uso final da madeira ser para energia, fibra, celulose ou produtos estruturais de madeira sólida. Ferramentas moleculares baseadas na identificação de polimorfismos no DNA, envolvidos no controle genético de fenótipos de interesse, prometem fornecer novas oportunidades para a seleção de características de crescimento, adaptabilidade a novas condições climáticas e propriedades da madeira de árvores cultivadas (Grattapaglia, Plomion *et al.*, 2009).

1.3. Base genética das principais metodologias de marcadores moleculares

Marcadores moleculares são definidos como qualquer fenótipo molecular oriundo de um gene expresso ou um segmento qualquer de DNA identificado (podendo ou não ser expresso). Abrangem desde as aloenzimas e isoenzimas (frequentemente referidos como marcadores bioquímicos), até as técnicas que estimam a variação diretamente no DNA (Ferreira e Grattapaglia, 1998). Os marcadores moleculares refletem diferenças hereditárias (i.e. polimorfismos) em sequências de DNA homólogas entre indivíduos. Estas diferenças podem ser devido ao polimorfismo de nucleotídeo simples (SNPs), inserções ou deleções (INDELs) ou rearranjos (translocações ou inversões). Os métodos de detecção de polimorfismo envolvem a utilização de: (1) endonucleases de restrição, (2) hibridização de ácidos nucléicos, (3) amplificação de sequências de DNA via PCR ou (4) sequenciamento. A decisão de qual marcador é o mais apropriado para uso depende da espécie, do objetivo do trabalho e dos recursos disponíveis. A seguir são brevemente apresentadas as tecnologias de marcadores moleculares mais amplamente utilizadas em plantas.

1.3.1. Polimorfismos de longitud de fragmentos de restrição ou RFLPs

Define-se RFLP (*Restriction Fragment Length Polymorphism*) como as diferenças no comprimento de fragmentos de restrição causados por SNPs ou INDELs que criam ou destroem sítios de reconhecimento de endonucleases de restrição. Tanto as bases teóricas como a técnica de RFLPs (Botstein, White *et al.*, 1980) no mapeamento do genoma de plantas têm sido extensivamente

analisadas (Tanksley, Young *et al.*, 1989). Na técnica, o DNA é digerido com enzimas de restrição, que clivam a fita dupla de DNA em sequências específicas, gerando um grande número de fragmentos de diferentes tamanhos. Os mesmos são separados por eletroforese, posteriormente desnaturados e transferidos para uma membrana de nitrocelulose (processo denominado Southern blot). A membrana é exposta a uma solução contendo sondas radioativas ou quimioluminescentes, que hibridizam com a região complementar de DNA, permitindo a visualização destes fragmentos através do processo de autoradiografia ou captação de luz emitida. O polimorfismo ocorre devido à variação na distribuição dos sítios de restrição na fita de DNA, gerando fragmentos de diferentes tamanhos, que resultam das diferenças na posição das bandas no gel.

A tecnologia de marcadores RFLP permitiu a construção dos primeiros mapas de ligação em plantas na década de 80 (Bernatzky e Tanksley, 1986; Helentjaris, Weber *et al.*, 1986), e iniciou uma rápida evolução no campo da genômica comparativa (Gale e Devos, 1998; Paterson, Bowers *et al.*, 2000). Os RFLPs são marcadores que têm a vantagem de cobrir, potencialmente, todo o genoma. Possuem expressão co-dominante, permitindo identificar genótipos heterozigotos e homozigotos. O número de marcadores é praticamente ilimitado; e apresentam alta consistência e reprodutibilidade dos resultados (Ferreira e Grattapaglia, 1998). Atualmente, raros são os casos nos quais a utilização dessa tecnologia se justifica devido ao fato de ser um processo demorado que requer trabalho intensivo com protocolos demorados demandando uma grande quantidade de DNA de alta qualidade. Além disso, é uma técnica difícil de automatizar e/ou multiplexar, tendo assim um rendimento baixo na genotipagem.

1.3.2. DNA Polimórfico Amplificado ao Acaso ou RAPD

O método se baseia na amplificação simultânea de segmentos de DNA com um iniciador de sequência arbitrária em uma PCR (*Polymerase Chain Reaction*) de baixa estringência (35-45°C). Para que haja amplificação de marcadores os sítios de anelamento do iniciador devem se encontrar a distâncias máximas em geral < 2kpb em sentido inverso. O polimorfismo RAPD (*Random Amplified Polymorphic DNAs*) deriva de diferenças de sequência e/ou indels entre indivíduos de forma que o anelamento do oligonucleotídeo iniciador varie gerando a presença ou ausência do produto amplificado (Welsh e Mcclelland, 1990; Williams, Kubelik *et al.*, 1990). Os segmentos amplificados, também chamados amplicons, são separados em géis de agarose, e marcadores RAPDs são assim constituídos pela presença ou ausência de um segmento particular. Quando comparada às demais

técnicas moleculares, é considerada simples, rápida, com custo relativamente baixo e geradora de um número elevado de marcadores com alto nível de polimorfismo.

Estes marcadores têm contribuído significativamente no desenvolvimento rápido de mapas genéticos, análise de variabilidade e *fingerprinting* (ou "impressão digital" do DNA) em espécies florestais (Grattapaglia e Sederoff, 1994; O' Malley, Grattapaglia *et al.*, 1996). As principais limitações da técnica RAPD consistem na reprodutibilidade variável dos resultados e uma informação genética limitada por loco, devido ao seu comportamento "dominante" ou seja, que não permite distinguir genótipos heterozigotos dos homozigotos. Isto impede a estimativa direta de frequências alélicas a partir dos dados genotípicos e consequentemente dos diversos parâmetros que determinam a estrutura genética de populações (Williams, Kubelik *et al.*, 1990; Ferreira e Grattapaglia, 1998).

1.3.3. Polimorfismo de Comprimento de Fragmentos Amplificados ou AFLP

AFLP (*Amplified Fragment Length Polymorphism*) é uma técnica de *fingerprinting* multiloco baseada em PCR que identifica eficientemente polimorfismo de DNA sem a necessidade de informação de sequência prévia (Vos, Hogers *et al.*, 1995; Mueller e Wolfenbarger, 1999). Na aplicação da técnica o DNA genômico é clivado com duas enzimas de restrição, uma de corte raro e outra de corte frequente e, posteriormente, ligado a adaptadores específicos que possuem terminais complementares às extremidades resultantes da clivagem pelas enzimas de restrição. Em seguida, é realizada a reação de PCR para a amplificação seletiva de fragmentos com iniciadores específicos que contém a sequência complementar aos adaptadores e bases arbitrárias adicionais que proporcionam a capacidade seletiva reduzindo a população de fragmentos visualizados em eletroforese em gel de poliacrilamida (Ferreira e Grattapaglia, 1998; Costa, Pot *et al.*, 2000).

Marcadores AFLP tem como vantagens o grande número de marcadores analisados em um único gel, com alto poder de detecção de variabilidade genética e a possibilidade de maior robustez dos resultados quando comparados com a técnica RAPD em função da etapa de redução de complexidade genômica proporcionada pelo corte com enzimas de restrição e amplificação seletiva. Entretanto, a principal limitação dos marcadores AFLP, é o baixo conteúdo de informação genética por loco, pois, assim como os marcadores RAPD, são de natureza dominante. Além disso, a análise AFLP envolve um maior número de etapas, necessitando de maior quantidade de reagentes e equipamentos, incrementando o custo das análises (Ferreira e Grattapaglia, 1998; Hoelzel e Green,

1998; Zhivotovsky, 1999; Costa, Pot *et al.*, 2000). A maior limitação dos marcadores dominantes, entretanto, sejam eles RAPD ou AFLP, é o fato deles apresentarem baixa transferibilidade mesmo entre indivíduos da mesma espécie. Portanto, mapas genéticos construídos com esses marcadores têm uma utilidade limitada, que se restringe quase que exclusivamente ao próprio pedigree utilizado para a sua construção, impossibilitando, por exemplo, experimentos de mapeamento comparativo e validação de QTLs.

1.3.4. Microssatélites ou SSRs

Dentre as diversas classes de marcadores moleculares disponíveis para pesquisa genética, os microssatélites ou SSR (*Simple Sequence Repeats*) destacam-se quanto à sua versatilidade como ferramentas moleculares (Chambers e Macavoy, 2000). São definidos como repetições em *tandem* de pequenos motivos de DNA de 1 a 6 pb de comprimento que exibem variação no número de repetições num determinado loco (Litt e Luty, 1989; Tautz, 1989; Weber e May, 1989). Os marcadores microssatélites apresentam uma grande abundância no genoma e um multialelismo derivado de sua elevada taxa de mutação. Assim, eles têm sido um recurso de grande difusão em estudos genéticos que fazem uso de marcadores moleculares. Segundo Chambers e MacAvoy (2000) o fator principal para sua popularidade tem sido o poder fornecido por esta classe de marcadores para solução de problemas biológicos. Além disso, pelo fato de serem co-dominantes, o que possibilita a discriminação de genótipos heterozigotos, e por seu caráter multi-alélico, marcadores microssatélites são os que possuem níveis mais elevados de informação de polimorfismo, ou PIC (*Polymorphism Information Content*).

O surgimento de locos microssatélites ocorre em regiões onde motivos simples de DNA repetitivo já estão representadas. Durante a síntese de DNA, a ocorrência de *slippage* numa região que contém um motivo microssatélite pode resultar no ganho ou perda de uma ou mais unidades de repetição (Chambers e Macavoy, 2000). Atualmente é consenso que o mecanismo principal de mutação e, por consequência, de evolução destes locos é o fenômeno de *slipped-strand mispairing* ou *DNA slippage* (Levinson e Gutman, 1987); (Tautz, 1989; Messier, Li *et al.*, 1996).

Metodologias de detecção e determinação de tamanho de alelos de locos microssatélites abrangem desde eletroforese em gel de agarose corada com brometo de etídio, passando por corridas em géis de poliacrilamida com detecção por auto-radiografia ou coloração com nitrato de prata. Finalmente, a detecção automatizada de alelos durante a eletroforese por meio de iniciadores

marcados com fluorocromos é hoje a metodologia padrão que surgiu apoiada no desenvolvimento de sequenciadores automáticos de DNA (Edwards, Civitello *et al.*, 1991; Ziegle, Su *et al.*, 1992; Schlotterer, 1998).

1.3.5. Polimorfismo de base única ou SNP

Nos últimos anos com a crescente disponibilidade de sequências em bancos de dados, principalmente de EST (*Expressed Sequence Tag*), cada vez mais tem crescido o interesse no desenvolvimento e utilização de marcadores baseados em polimorfismo de base individual ou SNP (*Single Nucleotide Polymorphism*). As variações de um único nucleotídeo caracterizadas como substituições, deleções e inserções são as formas de variação mais frequentes em diferentes genomas. Embora muito tenha sido feito e publicado sobre as vantagens destes marcadores para análise genética, a sua utilização em escala ainda se restringe a poucas espécies, especialmente seres humanos (Kruglyak e Nickerson, 2001; Stephens, Smith *et al.*, 2001; Van Uum, Stevens *et al.*, 2012), animais modelo (Grosse, Kappes *et al.*, 1999; Heaton, Harhay *et al.*, 2002; Kijas, Lenstra *et al.*, 2012) e, no caso de plantas cultivadas, *Arabidopsis* (Schmid, Sorensen *et al.*, 2003; Horton, Hancock *et al.*, 2012), soja (Zhu, Song *et al.*, 2003), e milho (Bhattramakki, Dolan *et al.*, 2002) entre outras, embora nos últimos anos SNPs tenham sido desenvolvidos também para espécies florestais com *Pinus* (Eckert, Pande *et al.*, 2009) e *Eucalyptus* (Grattapaglia, Silva-Junior *et al.*, 2011).

O custo de desenvolvimento e validação de marcadores SNPs é um processo demorado e caro, demandando o re-sequenciamento de painéis de indivíduos representativos da espécie ou das várias espécies alvo. Uma vez que os SNPs são validados existem várias metodologias para genotipá-los. Entretanto, uma grande distância separa a genotipagem de alguns poucos SNPs em muitos indivíduos, o que pode ser feito mesmo via PCR e eletroforese em agarose, e a genotipagem de milhares de SNPs em milhares de indivíduos. Plataformas de genotipagem de SNPs utilizando microarranjos de oligonucleotídeos ou contas (“beads”) são disponíveis, mas ainda a custos elevados para a grande maioria das espécies e laboratórios. O fato dos SNPs estarem envolvidos na predição de predisposição a doenças em humanos (Menendez, Krysiak *et al.*, 2006) ou como marcadores para características desejáveis em plantas (Thornsberry, Goodman *et al.*, 2001) e animais domésticos (Winter, Krämer *et al.*, 2002), tem resultado na criação de uma indústria crescente de tecnologias e plataformas que podem genotipar centenas de milhares de SNPs simultaneamente em indivíduos. Enquanto que para seres humanos e organismos modelo a questão de custo não é problema na

maioria das vezes, para plantas cultivadas de menor expressão a questão de custo passa a ser um fator limitante para a adoção da tecnologia.

1.4. Aplicações de marcadores moleculares em *Eucalyptus*

Um grande número de trabalhos tem sido publicado sobre marcadores moleculares e suas aplicações em espécies florestais desde os anos 80, quando isoenzimas permitiram realizar os primeiros estudos de sistemas de cruzamento em pomares de sementes e produzir as primeiras versões de mapas genéticos de coníferas (Moran e Bell, 1983). Desde então, a análise genética de espécies florestais progrediu essencialmente em razão do desenvolvimento de novas técnicas moleculares, começando com marcadores RFLP no final dos anos 80 e início dos anos 90 (Devey, Jermstad *et al.*, 1991; Byrne, Murrell *et al.*, 1995), seguidos de marcadores RAPD (Carlson, Tulsieram *et al.*, 1991; Grattapaglia e Sederoff, 1994) e AFLP (Gaiotto, Bramucci *et al.*, 1997; Marques, Araújo *et al.*, 1998), microssatélites brondani 1998 (Byrne, Marquezgarcia *et al.*, 1996; Brondani, Brondani *et al.*, 1998) e, mais recentemente, polimorfismos de base individual (SNP) (Brown, Gill *et al.*, 2004; Gonzalez-Martinez, Ersoz *et al.*, 2006; Grattapaglia, Silva-Junior *et al.*, 2011). Todos estes tipos de marcadores moleculares são de grande utilidade para diversos estudos genéticos em *Eucalyptus.*

1.4.1. *Fingerprinting* e análise de diversidade genética

Em áreas como melhoramento genético de *Eucalyptus* e produção florestal em larga escala, a correta identificação de clones é uma etapa crucial e muito importante (Grattapaglia e Kirst, 2008). O planejamento e a propagação de plantios florestais ocorrem vários anos antes da produção e consumo da madeira. Portanto, erros na identidade dos clones podem ter grandes efeitos em todo o processo produtivo. Várias são as tecnologias utilizadas para assistir no processo de identificação e proteção varietal em *Eucalyptus*, tais como RAPD (Keil e Griffin, 1994), AFLP (Gaiotto, Bramucci *et al.*, 1997) e microssatélites (Kirst, Cordeiro *et al.*, 2005; Ottewell, Donnellan *et al.*, 2005). Centenas de microssatélites dinucleotídeos têm sido publicadas, sendo a maioria desenvolvida a partir de E. *grandis* e *E. urophylla* (Brondani, Brondani *et al.*, 2002; Brondani, Williams *et al.*, 2006). Algumas dezenas, também foram geradas a partir de outras espécies como *E. globulus, E. nitens, E. sieberi* e *E. leucoxyon* (Byrne, Marquezgarcia *et al.*, 1996; Glaubitz, Emebiri *et al.*, 2001; Steane, Vaillancourt *et al.*, 2001; Ottewell, Donnellan *et al.*, 2005). Mais recentemente, têm sido desenvolvidos microssatélites baseados em tetra e pentanucleotídeos. Estes marcadores fornecem perfis de genótipos multiloco, são co-dominantes e também significativamente mais robustos que os

dinucleotídeos, além de uma alta possibilidade de multiplexagem (Sansaloni, 2008). Vários estudos têm demonstrado o grande poder de discriminação dos marcadores moleculares e seu potencial para aplicações em proteção varietal de clones, além da sua eficácia na avaliação da variabilidade genética existente em populações naturais, cálculo de distância genética e determinação de parentesco entre indivíduos de populações de melhoramento (Nesbitt, 1995; Baril, Verhaegen *et al.*, 1997; Marcucci Poltri, Zelener *et al.*, 2003; Cupertino, Leal *et al.*, 2011).

1.4.2. Genética de populações e filogenia

A primeira fonte de marcadores moleculares em *Eucalyptus* foi a aloenzimas (Brown, Matheson *et al.*, 1975). Estas foram utilizadas principalmente para resolver questões no nível de populações, como sistemas de cruzamento, diversidade genética e diferenciação de populações (Moran, 1992). Mais tarde, estudos baseados em DNA de *Eucalyptus* começaram com análises de RFLP em DNA de cloroplastos (Steane, West *et al.*, 1992). Devido à escassez de marcadores e do trabalho laborioso envolvido, apenas um pequeno número de amostras e marcadores foram utilizados, proporcionando uma baixa resolução das relações filogenéticas entre gêneros e subgêneros (Sale, Potts *et al.*, 1993). Em contraste, estudos utilizando esta mesma tecnologia de RFLP em DNA genômico provaram ser efetivo para muitas análises genéticos dentro e entre espécies relacionadas (Byrne, Parrish *et al.*, 1998; Butcher, Otero *et al.*, 2002; Wheeler, Byrne *et al.*, 2003; Elliott e Byrne, 2004). Posteriormente, o desenvolvimento de marcadores microssatélites para *Eucalyptus* (Byrne, Marquezgarcia *et al.*, 1996; Glaubitz, Emebiri *et al.*, 2001; Steane, Vaillancourt *et al.*, 2001; Thamarus, Groom *et al.*, 2002; Ottewell, Donnellan *et al.*, 2005; Brondani, Williams *et al.*, 2006) abriu a possibilidade de uma genotipagem mais ampla do genoma em um número relativamente maior de amostras. Estes marcadores proporcionaram o poder de examinar relações genéticas dentro e entre populações (Elliott e Byrne, 2004; Steane, Conod *et al.*, 2006; Jones, Vaillancourt *et al.*, 2007; Payn, Dvorak *et al.*, 2008; Butcher, Mcdonald *et al.*, 2009).

Assim, enquanto os marcadores microssatélites têm o potencial de fornecer resolução filogenética em níveis taxonômicos de espécies próximas e são muito úteis para estudos populacionais dentro de espécies, eles não são adequados para reconstrução filogenética em níveis taxonômicos mais distantes em função do seu modo de evolução muito rápida. Tecnologias mais robustas e com alto desempenho baseadas em polimorfismos de sequência do tipo RFLP ou SNPs de evolução mais lenta são necessários para estudos filogenéticos.

1.4.3. Mapas genéticos

Progressos significativos têm sido feitos nos últimos quinze anos no desenvolvimento de mapas genéticos em espécies florestais, com destaque para *Eucalyptus* e *Pinus* (Neale, 2007; Grattapaglia e Kirst, 2008). Os primeiros mapas de ligação em *Eucalyptus* foram construídos com algumas centenas de marcadores RAPD e AFLP (Grattapaglia e Sederoff, 1994; Verhaegen e Plomion, 1996; Myburg, Griffin *et al.*, 2003) e posteriormente com combinações de marcadores RAPD, RFLP, isoenzimas e ESTs (*Expressed Sequences Tags*) (Byrne, Murrell *et al.*, 1995; Bundock, Hayden *et al.*, 2000; Gion, Rech *et al.*, 2000). Com o advento de novas classes de marcadores como microssatélites e SNPs, mais informativos e transferíveis, os marcadores RAPD, RFLP, AFLP e isoenzimas foram cada vez menos utilizados. Em 2006, foi construído um mapa genético utilizando 234 marcadores microssatélites, dispostos em 11 grupos de ligação e cobrindo aproximadamente 90% do genoma de *Eucalyptus* (Brondani, Williams *et al.*, 2006). Recentemente, um mapa consenso foi construído a partir de duas populações de *Eucalyptus* utilizando uma combinação de 320 microssatélites e 304 SNPs (Lima, Silva-Junior *et al.*, 2011).

1.4.4. Mapeamento de QTLs e seleção assistida por marcadores

Vários trabalhos descreveram o sucesso na identificação de locos controladores de características quantitativas ou QTL (*Quantitative Trait Locu*) de maior efeito em componentes de produtividade (crescimento volumétrico e forma), qualidade da madeira (densidade básica, teor de lignina, rendimento em celulose), resistência a estresses abióticos (tolerância ao frio e à seca) e resistência a patógenos, principalmente fungos (Junghans, Alfenas *et al.*, 2003; Freeman, O'reilly-Wapstra *et al.*, 2008; Grattapaglia, Plomion *et al.*, 2009; Mamani, Bueno *et al.*, 2010; Gion, Carouche *et al.*, 2011). Entretanto, apesar de dezenas ou mesmo centenas de QTLs terem sido mapeados, a informação gerada tem sido pouco útil para a seleção assistida por marcadores (SAM). As principais razões para isso foram discutidas (Grattapaglia, 2004; Grattapaglia e Kirst, 2008; Grattapaglia, Plomion *et al.*, 2009; Grattapaglia e Resende, 2010) e podem ser resumidamente listadas: (1) Em geral alguns poucos QTLs são detectados capturando proporções limitadas da variação, uma vez que apenas a variação alélica dos parentais da população é amostrada; (2) As populações utilizadas para detecção de QTLs são relativamente pequenas, a magnitude de efeito dos QTLs é superestimada dificultando a predição de ganho esperado; (3) O comportamento imprevisível da interação entre alelos favoráveis aos QTLs e diferentes background genéticos, diferentes locais e diferentes idades da populações envolvidas.

Resultados práticos do melhoramento assistido pela genômica em espécies florestais ainda são raros. Com exceção de aplicações para seleção de parentais (Grattapaglia, Ribeiro *et al.*, 2004) e seleção individual dentro de famílias expandidas (Grattapaglia e Kirst, 2008), métodos de aplicação mais generalizada ainda não existem. A perspectiva de integração da genômica em um programa operacional de melhoramento vai depender em grande parte de aliar novas tecnologias de genotipagem que permitam amostrar o genoma de forma muito mais ampla, densa e economicamente viável, com novos conceitos em melhoramento genético. A proposta de uma metodologia preditiva denominada Seleção Genômica (Meuwissen, Hayes *et al.*, 2001) têm fornecido resultados animadores para a prática de seleção assistida. Esta metodologia foca exclusivamente nos aspectos de eficiência operacional e ganho genético e não visa a identificação de QTLs ou genes (Grattapaglia, Plomion *et al.*, 2009; Grattapaglia, Sansaloni *et al.*, 2010). Para a aplicação desta metodologia é necessária a utilização de uma tecnologia de genotipagem com capacidade de cobrir o genoma inteiro através de milhares de marcadores moleculares. Assim, grande parte dos alelos de interesse estarão em desequilíbrio de ligação com pelo menos um ou mais marcadores e, portanto, devidamente capturados nos modelos preditivos (Grattapaglia e Resende, 2010). No melhoramento florestal, testes de progênie derivados do intercruzamento de algumas dezenas de parentais elites constituem uma condição favorável para a implementação da Seleção Genômica. Um atrativo importante da utilização da Seleção Genômica no melhoramento de espécies florestais é a possibilidade de realizar seleção precoce, e com isso aumentar o ganho genético por unidade de tempo. Esta oportunidade é rara em espécies anuais, mas é evidente em espécies perenes com longo ciclo de vida. Ou seja, o fato de que a seleção possa ser praticada vários anos antes do fenótipo se expressar adequadamente (ex. densidade da madeira), representa uma vantagem potencial significativa.

1.5. Novas metodologias de marcadores moleculares para análise "*genome-wide*"

Marcadores moleculares tem sido utilizados de forma crescente em vários programas de melhoramento no mundo visando auxiliar na identificação de fenótipos desejáveis. No caso de aplicações que demandam uma análise ampla do genoma, a tecnologia ideal deve oferecer não somente milhares de marcadores moleculares cobrindo todo o genoma, mas estes devem ser obtidos preferencialmente em um experimento único, simples e de baixo custo. Embora muitas tecnologias de marcadores moleculares tenham sido desenvolvidas para *Eucalyptus* nas últimas décadas (Grattapaglia e Sederoff, 1994; Gaiotto, Bramucci *et al.*, 1997; Brondani, Brondani *et al.*,

2002; Brondani, Williams *et al.*, 2006; Grattapaglia, Silva-Junior *et al.*, 2011; Neves, Mamani *et al.*, 2011), todas elas apresentam algumas limitações. Por exemplo, no caso de microssatélites, a descoberta e validação de um grande número de marcadores que cobrem e genoma inteiro é lenta porque envolve várias etapas. No caso de microssatélites e SNPs a maioria dos marcadores baseia-se em informação de sequência e o custo de genotipagem em grandes populações é alto. O desenvolvimento de técnicas robustas que permitam a genotipagem de milhares de marcadores em milhares de amostras em um simples experimento resolveria grande parte das limitações mencionadas.

1.5.1. DArT (Diversity Arrays Technology)

A tecnologia DArT (*Diversity Arrays Technology*) foi desenvolvida e otimizada por DArT P/L em Canberra - Austrália (www.diversityarrays.com), a partir de 2011 com o objetivo de resolver várias das limitações das tecnologias existentes. Esta é uma técnica de marcadores moleculares baseados em hibridização de microarranjos para descoberta de polimorfismo ao longo do genoma que resulta na presença versus ausência de fragmentos individuais em representações genômicas (Jaccoud, Peng *et al.*, 2001). Os marcadores podem ser gerados com alto desempenho reduzindo os custos por amostra e por "data-point". Um simples ensaio DArT consegue genotipar simultaneamente centenas a milhares de polimorfismos SNPs e inserções/deleções (Indel) através do genoma. Por esta tecnologia não depender do conhecimento prévio de sequência, ela é de especial interesse para espécies que não possuem dados moleculares ou quando os recursos são limitados. Este método é equivalente ao RFLP porém com muito maior desempenho e de forma reversa (*Reverse Southern Blotting*), uma vez que as sondas são imobilizadas no microarranjo e o DNA a ser analisado é hibridizado.

O método DArT foi desenvolvido inicialmente em arroz (Jaccoud, Peng *et al.*, 2001), posteriormente validado em cevada (Wenzl, Carling *et al.*, 2004) e na espécie modelo *Arabidopsis thaliana* (Wittenberg, Lee *et al.*, 2005). Devido à eficiência mostrada por esta técnica na descoberta de marcadores nessas espécies, DArT foi amplamente utilizado em mais de 60 espécies de plantas como arroz (Xie, Mcnally *et al.*, 2006), mandioca (Xia, Peng *et al.*, 2005), cevada (Wenzl, Carling *et al.*, 2004), trigo (Akbari, Wenzl *et al.*, 2006; Niedziela, Bednarek *et al.*, 2012), ervilha (Yang, Saxena *et al.*, 2011), centeio (Milczarski, Bolibok-Bragoszewska *et al.*, 2011), aveia (Tinker, Kilian *et al.*, 2009; He e Bjørnstad, 2012), banana (Hippolyte, Bakry *et al.*, 2010), beterraba (Simko, Eujayl *et al.*, 2012), tomate (Van Schalkwyk, Wenzl *et al.*, 2012), grão de bico (Thudi, Bohra *et al.*, 2011), entre outras.

Esta tecnologia tem sido testada em algumas espécies de animais como camundongo, rã, mosquito, gado e ovelha (http://www.diversityarrays.com). Também foi utilizada para detectar variações de comunidades microbianas (Sessitsch, Hackl *et al.*, 2006) e em alguns organismos haplóides como o fungo patógeno da cana de açúcar *Ustilago scitaminea* (Braithwaite, 2005, dados não publicados) ou do trigo *Mycosphaerella graminicola* (Wittenberg, 2007).

As aplicações de DArT variam desde a obtenção de perfis genéticos para a identificação individual até à analise de diversidade (Wenzl, Carling *et al.*, 2004; White, Law *et al.*, 2008; Steane, Nicolle *et al.*, 2011; Zhang, Liu *et al.*, 2011; He e Bjørnstad, 2012). Além disso, é possível a construção rápida de mapas de ligação de alta densidade (Wenzl, Li *et al.*, 2006; Hippolyte, Bakry *et al.*, 2010; Milczarski, Bolibok-Bragoszewska *et al.*, 2011; Oliver, Jellen *et al.*, 2011; Thudi, Bohra *et al.*, 2011), mapeamento físico, em projetos que envolvem sequenciamento dos marcadores (Paux, Sourdille *et al.*, 2008; Rodríguez-Suárez, Giménez *et al.*, 2012), identificação de QTLs (Bedo, Wenzl *et al.*, 2008; Huynh, Wallwork *et al.*, 2008), seleção assistida por marcadores (Mccartney, Stonehouse *et al.*, 2011) e seleção genômica ampla (Crossa, Campos *et al.*, 2010; Grattapaglia, Sansaloni *et al.*, 2010; Resende, Resende *et al.*, 2012).

Resumidamente, a metodologia envolve a reunião de um grupo de DNA genômico total de diversos indivíduos que representem o germoplasma de interesse. Este DNA é submetido a um processo de redução da complexidade através de corte com enzimas de restrição que reconhecem e cortam preferencialmente em regiões hipometiladas do DNA, ricas em DNA de alta complexidade. As representações genômicas obtidas a partir do pool de DNA são clonadas criando bibliotecas de insertos individuais os quais são imobilizados sobre um microarranjo geralmente denominado de "descoberta", pois tem por objetivo permitir a seleção de fragmentos que revelem polimorfismo de sequência. Posteriormente, o DNA de amostras individuais é cortado com as mesmas enzimas de restrição com o objetivo de gerar a mesma redução de complexidade utilizada anteriormente para o desenvolvimento das bibliotecas. Os fragmentos obtidos são marcados com fluorescência e finalmente hibridizados sobre o arranjo de descoberta. Os fragmentos polimórficos detectados mostram sinais de hibridização variáveis entre diferentes indivíduos, onde a presença de sinal é representada por “1” e a ausência por “0” (Huttner, Wenzl *et al.*, 2004). Estas etapas da técnica estão padronizadas, embora possam precisar de pequenas modificações dependendo da espécie alvo.

A técnica pode ser resumida em três passos principais (Figura 1):

1) Desenvolvimento do Arranjo

- Construção de uma biblioteca (representação genômica) por meio da aplicação de um método de redução da complexidade do DNA genômico.
- Impressão dessa biblioteca genômica sobre o microarranjo.

2) Genotipagem

- Redução da complexidade genômica das amostras ou *targets*.
- Marcação com fluorescência das representações genômicas (*targets*) a serem genotipadas.
- Hibridização dos *targets* com o microarranjo.
- Lavagem e digitalização dos slides.

3) Análise dos dados

- Extração de dados do microarranjo e detecção dos polimorfismos.
- Aplicação dos marcadores polimórficos.

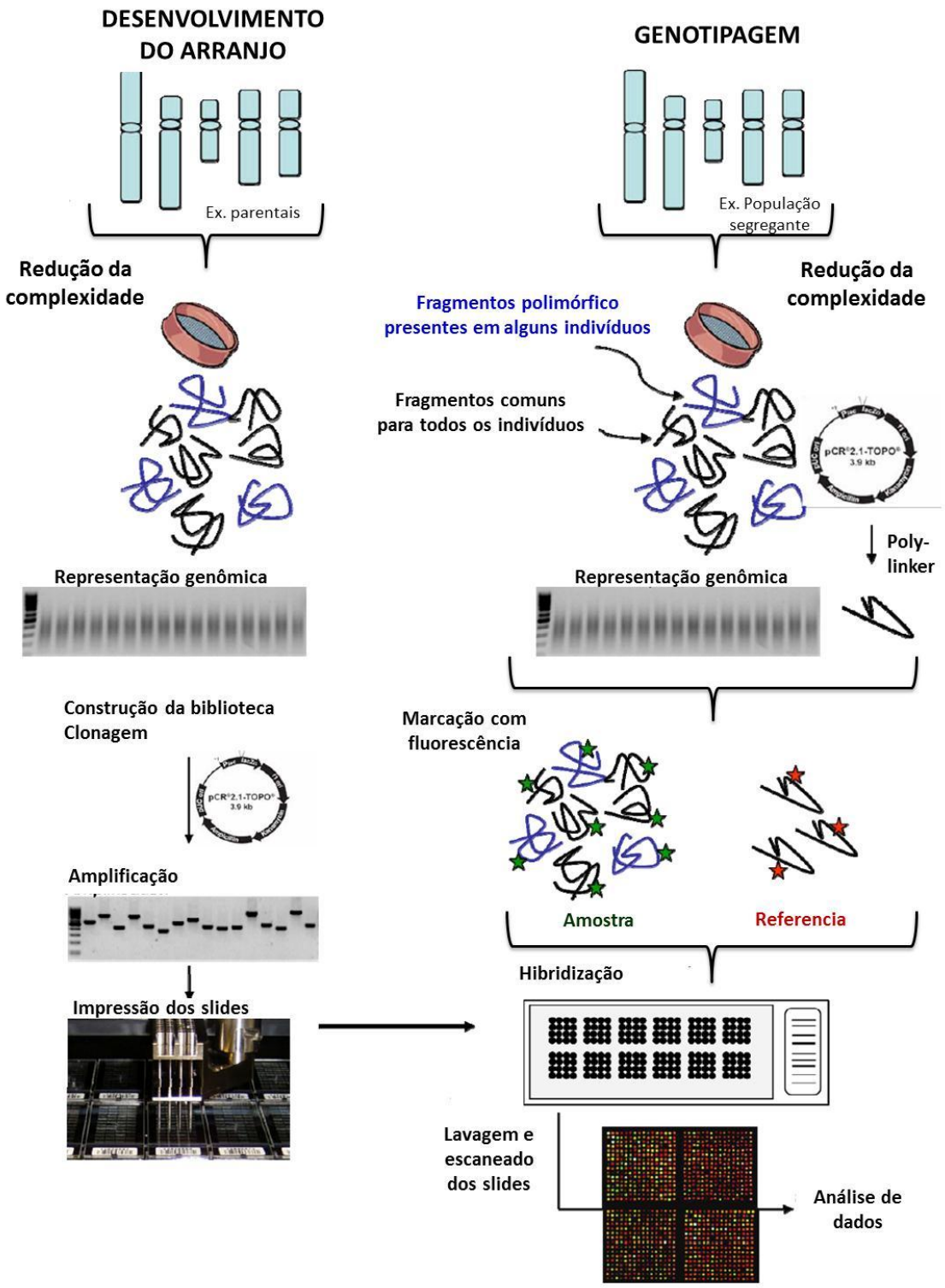


Figura 1. Fluxograma esquemático da metodologia DArT. Na primeira etapa (Desenvolvimento do Arranjo - esquerda) são construídas as bibliotecas genômicas de complexidade reduzida que posteriormente são impressas no microarranjo. Na segunda etapa (Genotipagem - direita), os indivíduos de uma população específica são genotipados por meio da hibridação das suas representações genômicas (i.e. amostra ou *target*) sobre os slides (i.e. microarranjo). O *Software*

(*DArTSoft*) identifica insertos informativos que revelam polimorfismos do tipo presença (1) ou ausência (0) na representação genômica.

Embora esta tecnologia seja muito eficiente e operacional para espécies de plantas, desde 2009 DArT Pty. Ltd. vem trabalhando na adaptação do método de redução de complexidade genômica e detecção de polimorfismos, não mais via hibridização, mas sim via sequenciamento do tipo Next Generation Sequencing pela geração de milhões de sequências curtas.

1.5.2. Genotipagem por sequenciamento de representações DArT

Diferentes abordagens de genotipagem por sequenciamento em plantas começaram a ser desenvolvidas em 2010 nos EUA por alguns poucos grupos em Cornell e Brigham Young University (Maughan, Yourstone *et al.*, 2010; Myles, Chia *et al.*, 2010), e na Austrália na DArT Pty (A. Kilian Com. pess.). Esta abordagem é hoje denominada genericamente de GbS (*Genotyping-by-Sequencing*) (Elshire, Glaubitz *et al.*, 2011), a qual apresenta um excelente potencial de genotipar milhares de amostras para milhares ou dezenas de milhares de marcadores a custos de poucas dezenas de dólares.

A tecnologia baseia-se na redução de complexidade genômica da amostra de DNA total utilizando combinações específicas de enzimas de restrição (ER), otimizadas para cada espécie. Os fragmentos selecionados resultantes são submetidos ao sequenciamento para a geração de dezenas de milhões de sequências curtas (em geral 77 bases) em plataformas de sequenciamento Illumina de NGS (*Next Generation Sequencing*), nas quais os polimorfismos são detectados. Após a redução de complexidade via enzimas de restrição e antes do sequenciamento, cada amostra de DNA a ser genotipada recebe adaptadores com sequências indexadoras (barcodes) que permitem mais tarde rastrear as sequências geradas para cada amostra. Desta forma, 96 amostras podem ser sequenciadas conjuntamente em cada canaleta de sequenciamento otimizando significativamente os custos. Assim, são detectados polimorfismos de presença ou ausência entre amostras derivados da variabilidade na distribuição dos sítios de restrição e de SNPs entre sequências em comum entre todas as amostras.

Em cada corrida de sequenciamento em plataforma Illumina podem ser genotipadas 768 amostras e gerados milhares ou dezenas de milhares de marcadores a depender da diversidade nucleotídica da espécie. Até o momento existem poucos trabalhos publicados referentes a esta nova

tecnologia. Em 2010 dois trabalhos (Maughan, Yourstone *et al.*, 2010; Myles, Chia *et al.*, 2010) demonstraram a viabilidade do método, e recentemente foram publicados outros dois trabalhos mostrando o sucesso da técnica em milho, trigo e cevada (Elshire, Glaubitz *et al.*, 2011; Poland, Brown *et al.*, 2012). O laboratório DArT vem trabalhando nesta metodologia desde 2009 e já possui um sistema operacional para várias espécies de plantas (dados não publicados). Trata-se, portanto, de uma metodologia muito inovadora que deverá certamente ser utilizada em diversas espécies de plantas nos próximos anos.

2. OBJETIVOS

1) Desenvolvimento de uma plataforma de microarranjo DArT para genotipagem de polimorfismos de sequência com alta performance e cobertura genômica com 7.000 a 8.000 marcadores polimórficos dentro e entre as principais espécies plantadas de *Eucalyptus*.

2) Teste e análise comparativa da eficiência de redução da complexidade genômica com diferentes combinações de enzimas de restrição, na revelação de fragmentos capazes de detectar polimorfismos de sequência entre e dentro espécies de *Eucalyptus.*

3) Utilização do microarranjo DArT em estudos de diversidade genética, identificação individual dentro e entre espécies de *Eucalyptus*, identificação de híbridos, reconstrução filogenética e mapeamento genético.

4) Avaliar a metodologia de genotipagem por sequenciamento baseada na redução de complexidade genômica DArT por meio da construção de um mapa genético para uma população segregante derivada da árvore BRASUZ1.

3. CAPÍTULO 1

DESENVOLVIMENTO E VALIDAÇÃO DE UM MICROARRANJO DArT PARA GENOTIPAGEM DE *Eucalyptus*

3.1. INTRODUÇÃO

Uma série de tecnologias de marcadores moleculares têm sido desenvolvidas e utilizadas para espécies de *Eucalyptus* nos últimos 20 anos (Grattapaglia e Kirst, 2008). Cada uma dessas tecnologias permitiu importantes avanços na compreensão da variabilidade genética, estudos evolutivos e no melhoramento deste vasto gênero que inclui mais de 700 espécies, algumas das quais são plantadas no mundo inteiro (Myburg, Potts *et al.*, 2007). Marcadores moleculares têm sido utilizados para resolver problemas filogenéticos, para descrever a estrutura genética de populações naturais, resolver questões relacionadas com a manipulação da variação genética em populações de melhoramento e construir mapas de ligação que, por sua vez, levaram à identificação de QTLs para diversas características de interesse como produtividade, resistência a doenças, e qualidade da madeira (Grattapaglia e Kirst, 2008; Grattapaglia, Plomion *et al.*, 2009). No entanto, a densidade de genotipagem alcançada, mesmo com tecnologias tais como AFLP, permanece em algumas centenas de marcadores por amostra e sendo esta técnica baseada em gel requer um procedimento relativamente laborioso. Em estudos com microssatélites, a multiplexagem têm permitido o aumento de eficiência de genotipagem. No entanto, a transferibilidade de microssatélites através de espécies é notoriamente deficiente e precisa ser investigada e otimizada antes dos microssatélites serem utilizados em novas espécies (Grattapaglia e Kirst, 2008). Portanto, métodos de genotipagem de alta performance e com ampla cobertura genômica são necessários para incrementar a resolução e a velocidade de análise em uma variedade de aplicações.

Diversity Arrays Technology (DArT) (Jaccoud, Peng *et al.*, 2001) fornece uma alternativa interessante para satisfazer os requisitos de performance, cobertura genômica e transferibilidade entre espécies, outro aspecto de grande relevância especialmente quando se trata de um gênero tão diverso. Esta tecnologia de genotipagem baseia-se na redução da complexidade genômica utilizando enzimas de restrição, seguido por uma hibridização em microarranjos que permite analisar, simultaneamente, centenas a milhares de marcadores no genoma.

Embora tenha sido desenvolvida há mais de 10 anos, esta tecnologia tem ganhado muito interesse principalmente em plantas (Wenzl, Carling *et al.*, 2004; Wittenberg, Lee *et al.*, 2005; Xia, Peng *et al.*, 2005; Akbari, Wenzl *et al.*, 2006; Tinker, Kilian *et al.*, 2009; Hippolyte, Bakry *et al.*, 2010; Oliver, Jellen *et al.*, 2011; He e Bjørnstad, 2012; Niedziela, Bednarek *et al.*, 2012). Neste capítulo, é apresentado o desenvolvimento da primeira plataforma de genotipagem de marcadores DArT para Eucalyptus envolvendo mais de 7000 marcadores selecionados. Além disso, são apresentadas suas potenciais aplicações em estudos de diversidade, filogenia e mapeamento de ligação em espécies de *Eucalyptus.*

3.2. MATERIAL E MÉTODOS

3.2.1. Material biológico utilizado

O desenvolvimento da plataforma de genotipagem DArT foi realizado em duas etapas. A primeira envolveu a triagem de duas amplas baterias de fragmentos DArT visando a seleção de clones reveladores de marcadores polimórficos. A segunda etapa envolveu a montagem e validação do microarranjo operacional. Amostras de um conjunto de 12 indivíduos geneticamente não relacionados de cada uma das sete principais espécies de *Eucalyptus* (*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis e E. cladocalyx*) foram utilizadas durante a primeira etapa de desenvolvimento (Tabela 1). Na segunda etapa, além das amostras utilizadas na primeira etapa, foi também empregado um painel de diversidade envolvendo um indivíduo de cada uma de 62 espécies diferentes de *Eucalyptus* para a validação do microarranjo operacional (Tabela 2). O DNA dos indivíduos foi extraído a partir de tecido fresco de folhas e casca usando o método CTAB 2% (Doyle e Doyle, 1987). O equipamento TissueLyser da Qiagen foi usado para a maceração das folhas. Após a extração, o precipitado foi ressuspendido em uma solução de Tris/EDTA (TE) a pH 8,0 com Ribonuclease A (RNaseA) e incubado à temperatura de 37 ºC por 20 minutos para ação da enzima. O DNA foi então quantificado em géis de agarose 1% corados com brometo de etídeo, bem como no equipamento NanoDrop para obter uma avaliação da pureza do DNA.

3.2.2. Redução de complexidade genômica

O primeiro passo no procedimento DArT envolve a redução do número de fragmentos presentes na representação genômica. Este processo é chamado de redução da complexidade. Se o número de fragmentos genômicos únicos (complexidade) nos *targets* aumenta, a possibilidade de hibridação cruzada e obtenção de intensidade de sinal baixo ou não específico aumenta. Portanto, é

essencial gerar um subconjunto de fragmentos do genoma para que o sinal seja específico e intenso. Como DArT é um método baseado em hibridização, a representação genômica pode ser muito mais complexa em comparação com sistemas baseados em gel, tais como RAPD e AFLP. No entanto, para descobrir polimorfismo baseado em SNPs, Indels ou mesmo diferenças de metilação, e obter uma marcação suficiente de todos os fragmentos presentes na representação, a redução da complexidade é necessária.

O número de marcadores que podem ser obtidos com a técnica DArT depende da diversidade presente nas amostras utilizadas na construção da biblioteca, o método de redução da complexidade, e o número de clones que são impressos nos slides. A redução da complexidade é gerada como produto da digestão do DNA com uma combinação de enzimas de restrição, sendo uma delas uma enzima de corte raro e sensível à metilação (em geral PstI) de forma a cortar preferencialmente em regiões hipometiladas ricas em genes ou sequências de alta complexidade. A outra enzima utilizada em geral é de corte frequente com o objetivo específico de remover fragmentos *PstI/PstI* mais longos da população de fragmentos enriquecendo assim para fragmentos *PstI/PstI* de tamanho mais curto e portanto dentro do tamanho adequado para clonagem e PCR. Adaptadores são ligados à sequência complementar do sítio de reconhecimento da enzima *PstI* somente, de forma que somente são amplificados os fragmentos *PstI/PstI* produzindo assim a representação genômica desejada via PCR (Figura 2).

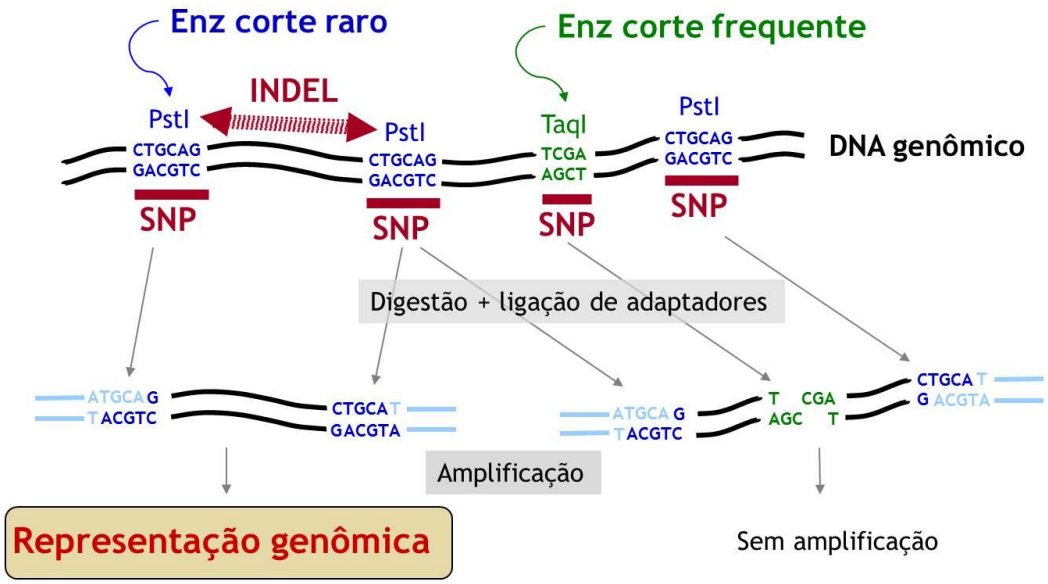


Figura 2. Representação esquemática do processo de redução da complexidade genômica produzida por uma combinação de enzimas de corte raro (*PstI* em azul) e corte frequente (*TaqI* em verde). Adaptadores são ligados aos extremos *PstI* (azul claro) e unicamente são amplificados os fragmentos com extremos *PstI/PstI*.

Foram testados sete métodos de redução de complexidade para identificar o mais adequado para *Eucalyptus*. O teste envolveu a digestão de duas amostras, uma de *E. grandis* e outra de *E. urophylla*, com várias combinações de enzimas, sendo utilizada em todos os casos *PstI* como enzima primária de corte raro, enquanto que *TaqI, BstNI, MspI, HpaII, BanII, MseI, AluI* como enzimas secundárias de corte frequente. Foram realizados simultaneamente a digestão e ligação de adaptadores em 75 ng de DNA genômico em uma solução aquosa de 10 µL contendo 2 unidades de cada enzima de restrição, 80 Unidades de DNA Ligase T4 e 0,05 mM do adaptador (5'-CACGATGGAGCATCCAGT-3' anelado com 5'-CTGGATCCATCGTGCA-3'). As reações foram incubadas a 37°C por 2 horas, seguido por 2 horas a 60°C conforme exigido pelas combinações de enzimas.

Do produto da reação de digestão/ligação foi utilizado 1 µL como molde para a amplificação via PCR em uma reação de 50 µL utilizando os iniciadores que reconhecem os sítios *PstI* (5'-GATGGATCCAGTGCAG-3') com os seguintes parâmetros: 94°C por 1 min, seguido por 30 ciclos de 94°C por 20 segundos, 58°C por 40 segundos, 72°C por 1 min, e finalizando com uma extensão de 72°C por 7 min. Para comprovar o sucesso da amplificação, foram corridos 5 µL do produto de amplificação em um gel de agarose 1,2% corado com brometo de etídio. A melhor combinação de enzimas de restrição selecionada foi a que apresentou um rastro homogêneo de fragmentos. Padrões de bandas indicam a presença de fragmentos repetitivos ou multicópias que serão redundantes na biblioteca e na subsequente descoberta de marcadores (Kilian, Huttner *et al.*, 2005).

3.2.3. Construção de bibliotecas piloto de clones DArT

Foram escolhidos os dois métodos de redução de complexidade que apresentaram um rastro mais homogêneo de fragmentos para a construção de bibliotecas piloto contendo 1.536 clones de cada método. As representações genômicas de cada método de redução de complexidade foram clonadas usando o *TOPO TA Cloning Kit* (Invitrogen), seguindo as instruções do fabricante. Colônias brancas individuais foram depositadas em placas de 384 poços contendo médio LB, com 4,4% de glicerol e 100 µg/mL de ampicilina. Posteriormente as placas foram incubadas a 37°C por 18 horas. A amplificação por PCR foi realizada utilizando 0,5µL de cultura bacteriana como molde e 0,2

µM de iniciadores "M13 Forward" e "M13 Reverse" (Invitrogen). O programa de PCR utilizado pelo termociclador foi o seguinte: 95°C por 4 min., 57°C por 35 seg., 72°C por 1 min., seguido por 35 ciclos de 94°C por 35 seg., 52°C por 35 seg., 72°C por 1 min. e um passo final de 72°C por 7 min. Uma pequena alíquota de 1µL do produto de PCR foi analisado em gel de agarose 1,2% para confirmar o sucesso da amplificação. Os produtos de PCR restantes foram secados a 37°C e lavados com etanol 70% antes de ser dissolvidos com 25µL de buffer "DArTspotter", desenhado para usar com slides de microarranjo revestidos com poli-L-lisina (P. Wenzl não publicado - disponível de DART Pty Ltd).

O microarranjo foi impresso utilizando um arrayer MicrogridII (Biorobotics) em slides de vidro revestidos com poli-L-lisina (Erie Scientific). Imediatamente foi colocado em todos os slides um código de barra para identificação individual, e estes deixados secando sobre a bancada por 24 horas para facilitar a aderência do DNA impresso. Passadas as 24 horas, os slides foram imersos em agua Milli-Q a 95 C por 2 min, e em seguida, em água Milli-Q com 0,1 mM DTT e 0,1 mM EDTA a 20 C e, para eliminar o excesso de DNA sobre os slides. Finalmente, foram secados por centrifugação a 500 × g por 7 min e através de vácuo por 30 min.

3.2.4. Preparação de amostras ou "*target*" para hibridização

Representações genômicas das 12 amostras de *E. grandis* e *E. globulus* foram preparadas como descrito anteriormente na construção da biblioteca, com o fim de gerar "*targets*" para hibridizá-los ao microarranjo. Os produtos de amplificação foram precipitados individualmente com isopropanol, lavados com etanol 70% e secos ao ar em temperatura ambiente por 12 horas. As 12 amostras geradas para cada espécie foram testadas com réplicas completas. Os *targets* foram marcados com fluorescência em uma reação de 10µL contendo 2,5 nM de Cy3-dUTP (de cor verde) ou Cy5-dUTP (de cor vermelha) (Amersham Bioscience), 2,5 unidades de fragmentos Klenow exo de *E. coli* Polimerase I (New England Biolabs) e 25µM de decâmeros arbitrários em 1x de tampão NEB 2 (New England Biolabs). Finalmente, as reações marcadas foram incubadas a 37°C por um período de 3 horas.

O sinal de hibridização do fragmento polylinker do vetor de clonagem foi utilizado como referência de qualidade, ou seja, a intensidade do sinal da referência é analisada pelo programa *DArTSoft* determinando para cada clone a quantidade de DNA impresso no arranjo. A referência foi marcada com a fluorescência carboxi-6-FAM (de cor azul) (Invitrogen), seguindo o mesmo procedimento utilizado para marcar os *targets*. Para isso foram utilizados aproximadamente 500ng

de produto de PCR do fragmento polylinker amplificado a partir do vetor vazio utilizado para a construção da biblioteca. Esta marcação da referência posteriormente foi misturada com o tampão de hibridação (1:1000) (Figura 3).

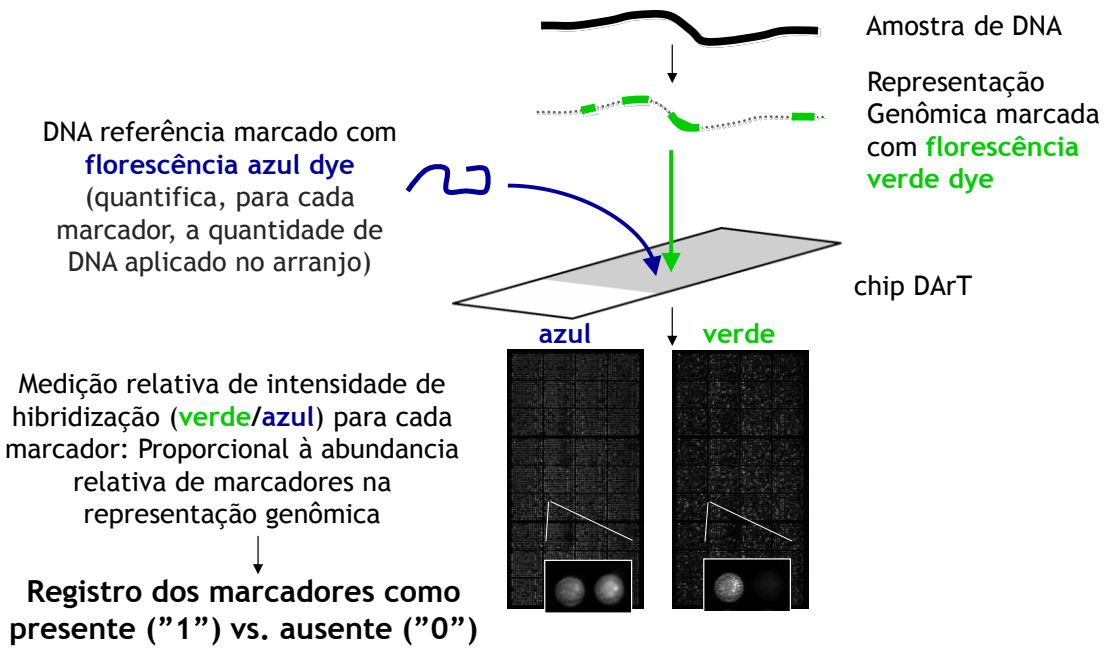


Figura 3. Representação esquemática da hibridização dos *targets* (verde) e a região polylinker do vetor (azul), sobre o slide que apresenta todos os clones do microarranjo impressos.

3.2.5. Hibridização dos *targets* ao microarranjo

Os *targets* marcados com fluorescência foram misturados com um tampão de hibridização contendo uma proporção de 50:5:1 de Express Hyb (Clonetech), DNA de esperma de salmão (Promega) e a região polylinker do vetor pCR 2,1 TOPO marcada com FAM (Invitrogen), além de 2 mM de EDTA a pH 8,0. Em rotina, a mistura de hibridização é preparada em conjuntos de 8 tubos arranjados em formato de placa de 96 poços (Figura 4). Aos 5µL de *target* marcado são acrescentados 60µL do tampão de hibridização, e estes misturados gentilmente e desnaturados a 95°C por 2 min. Em seguida, a temperatura é diminuída até 55°C para posteriormente aplicar a mistura sobre os slides localizados dentro das câmaras de hibridização. Estas são então fechadas e mantidas em banho-maria a 62,5°C por um período de 18 horas (ou *overnight*).

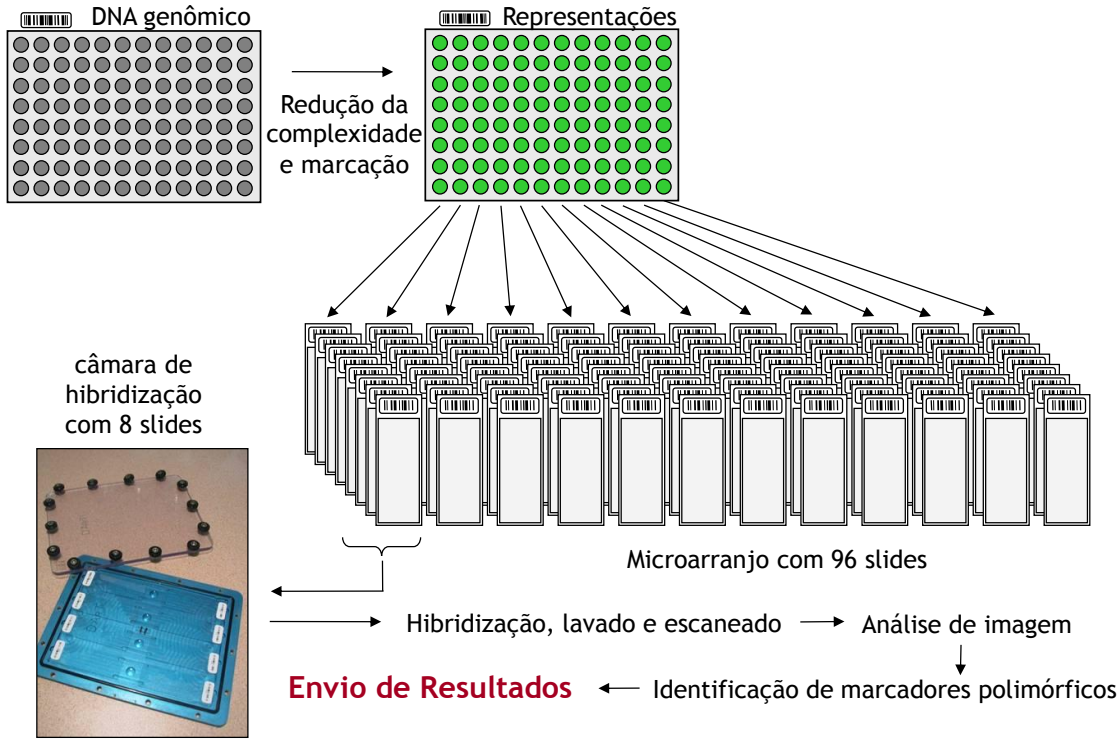


Figura 4. Visão geral da rotina do proceso de hibridização para uma placa de 96 poços contendo amostras de DNA genômico.

3.2.6. Lavagem e digitalização dos slides

Posteriormente à hibridização, os slides do microarranjo são lavados em quatro soluções de estringência crescente (1x SSC, SDS 0,1% por 4 min; 1x SSC por 4 min; 0,2x SSC por 1 min; 0,02x SSC por 30 segundos), secados por centrifugação a 500 × g por 7 min. e através de vácuo por 30 min.. Uma vez secos, os slides são colocados em um scanner para microarranjos com laser confocal TECAN LS300, e uma leitura da intensidade da fluorescência realizada com uma resolução de 20 µm por pixel. Este scanner digitaliza sequencialmente três imagens para cada slide, utilizando as seguintes combinações de laser/filtro: 488 nm/520 nm (para detectar o sinal fluorescente azul emitido pelo polylinker do vetor TOPO pCR 2,1 marcado com FAM); 543 nm/590 nm (para geração das imagens a partir do sinal fluorescente verde das amostras marcadas com Cy-3); 633 nm/670 nm (para geração das imagens a partir do sinal fluorescente vermelho das amostras marcadas com Cy-5) (Figura 5).

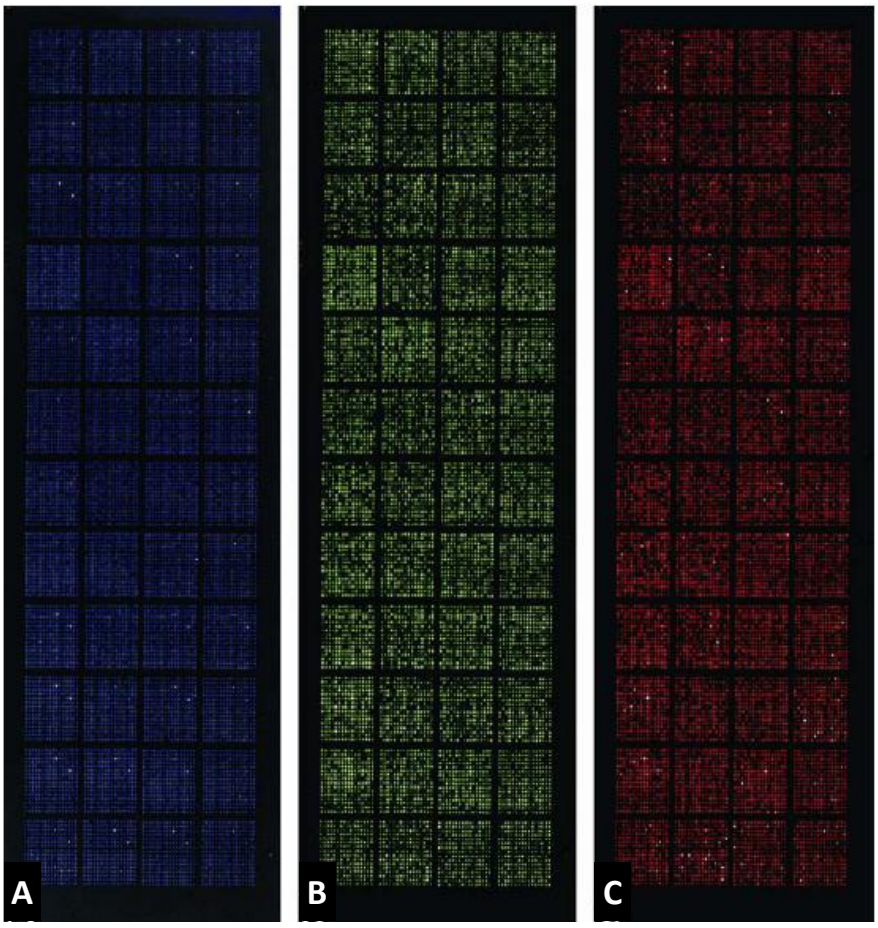


Figura 5. Slide do microarranjo de *Eucalyptus* detectado com diferentes combinações laser/filtro. Fragmento do polylinker do vetor de clonagem marcado com FAM, que serve como uma referência de qualidade (A), e duas amostras diferentes marcadas com Cy3 (B) e Cy5 (C).

O uso de uma terceira marcação fluorescente não é absolutamente necessário, tanto que a técnica DArT pode ser realizada utilizando qualquer scanner de duas cores. No entanto, a terceira marcação fluorescente fornece um rendimento das amostras significativamente maior refletindo em um menor custo por ensaio, já que duas amostras diferentes podem ser processadas em um único slide ao invés de apenas uma.

3.2.7. Processamento das imagens e declaração dos genótipos

As imagens resultantes foram analisadas utilizando o programa *DarTSoft* versão 7.44, desenvolvido pela empresa Diversity Arrays Technology Pty Ltd para extração de dados desde o microarranjo, detecção de polimorfismo, e genotipado das amostras (Cayla et al. não publicado). *DArTSoft* localizou automaticamente os "spots" individuais sobre o microarranjo, levando em consideração o diâmetro dos "spots" (em pixels) e a resolução com a qual os slides foram

digitalizados. Esta localização foi realizada a partir das imagens TIFF de 16 bits geradas pelo scanner. O programa rejeitou os "spots" que apresentaram um sinal baixo para a sequência de referência, indicando baixa qualidade da hibridização. Foram calculadas as intensidades relativas do sinal para cada "spot" aceito do microarranjo como log [*target*/referência], ou seja, log [sinal de Cy-3/ sinal de FAM] para os *targets* marcados com Cy-3, e log [sinal de Cy-5 / sinal de FAM] para os *targets* marcados com Cy-5. Se estas relações (*target*/referência) eram similares em todos os slides, os fragmentos foram considerados monomórficos, enquanto que se dois clusters (alelos) eram distinguidos e a variância da intensidade relativa entre eles era de pelo menos 80% da variância total, os clones foram considerados polimórficos e genotipados de forma binaria como "0" ou "1". Os clones foram incorporados à tabela de genótipos quando possuíam uma probabilidade ≥ 0,95 em pelo menos 90% dos slides. E se a probabilidade era menor a 0,95, estes clones eram situados fora dos clusters, e por tanto considerados como dados faltantes ou "-"(Figura 6).

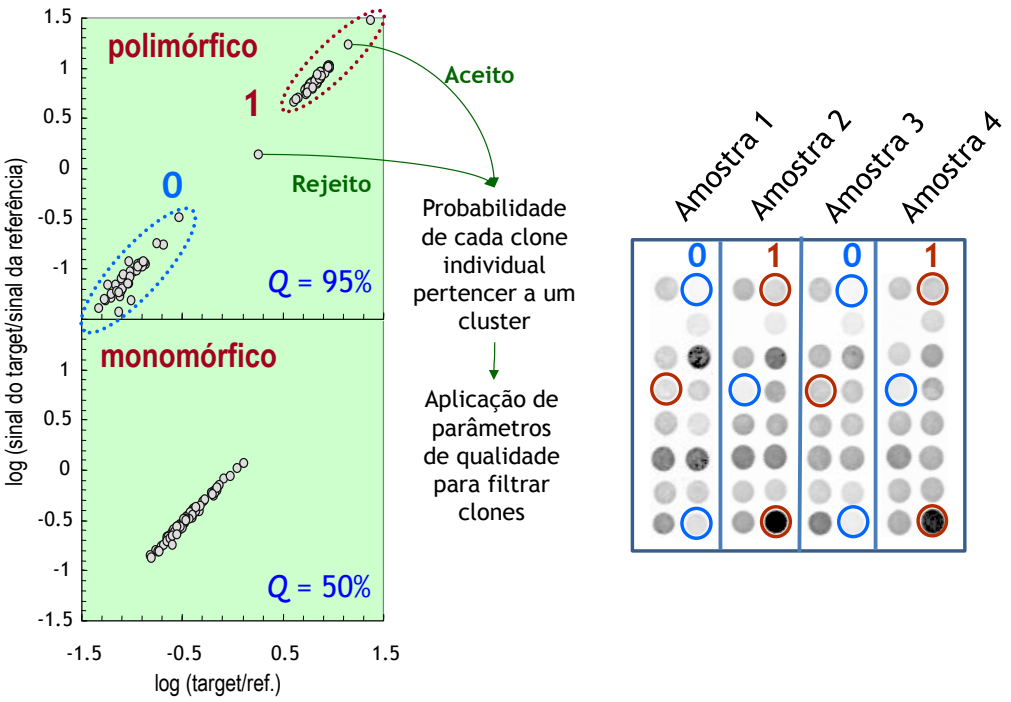


Figura 6. Interpretação de dados calculado pelo programa *DArTSoft* mostrando a diferença de intensidade entre genótipos. À esquerda, os gráficos mostram o log (sinal do *taget*/sinal da referência), quanto maior é a diferença de intensidade entre clusters (Q=95%), maior é o polimorfismo dos marcadores. Se a intensidade é homogênea (Q<50%), os marcadores são monomórficos. À direita está exemplificada a diferença de intensidades entre amostras representadas como ausência (0) e presença (1).

Para cada um destes clones, o programa gerou uma serie de parâmetros de qualidade para auxiliar na seleção de clones polimórficos. Os parâmetros utilizados neste trabalho foram:

- **Valor de Reprodutibilidade** (0-100%): em cada slide foram impressas duas réplicas de cada um dos 7.680 clones, totalizando 15.360 fragmentos impressos aleatoriamente no microarranjo. Para uma maior confiança, além destas réplicas de clones, foram hibridizadas também réplicas de todos os *targets*, ou seja todas as amostras foram genotipadas em duplicata. A Reprodutibilidade foi então calculada a partir dos indivíduos que possuem réplicas de hibridização no mesmo slide (i.e. clones do microarranjo) ou em slides diferentes (i.e. *target*). Os valores (0 ou 1) para estas réplicas devem pertencer ao mesmo cluster. Se um valor é comparado com um valor "perdido", não é considerado como uma diferença. Neste trabalho foram aceitos marcadores com um máximo de três erros entre as réplicas.

- **Valor de Call Rate** (0-100%): representa a porcentagem de genótipos que foram chamados de “0” ou “1” para cada clone em todos os *targets*. Foi aceito um valor mínimo de 80% para o marcador ser considerado polimórfico.

- **Valor Q** (0-100%): mede a fração da variância total entre os dois clusters em todos os indivíduos. Um valor Q alto indica que os dois clusters correspondendo aos dois possíveis genótipos são bem diferenciados. Foram considerados valores de Q> 50% para esta análise.

- **Valor de conteúdo de informação polimórfica ou PIC** (*Polymorphism information content*) (0-0,5): este valor mostra como estão distribuídas as frequências relativas de presença versus ausência de sinal. O valor máximo de 0,5 resulta de um marcador cujas frequências de presença e ausência são iguais a 0,5. Vale ressaltar que este valor de PIC não corresponde ao parâmetro PIC proposto originalmente para avaliar o conteúdo informativo de marcadores co-dominantes especificamente para mapeamento genético (Botstein, White *et al.*, 1980).

Uma vez finalizada a análise de polimorfismo, o programa *DArTSoft* exporta um arquivo que pode ser aberto utilizando Microsoft Excel, no qual todos os genótipos e parâmetros podem ser observados e manipulados (Figura 7).

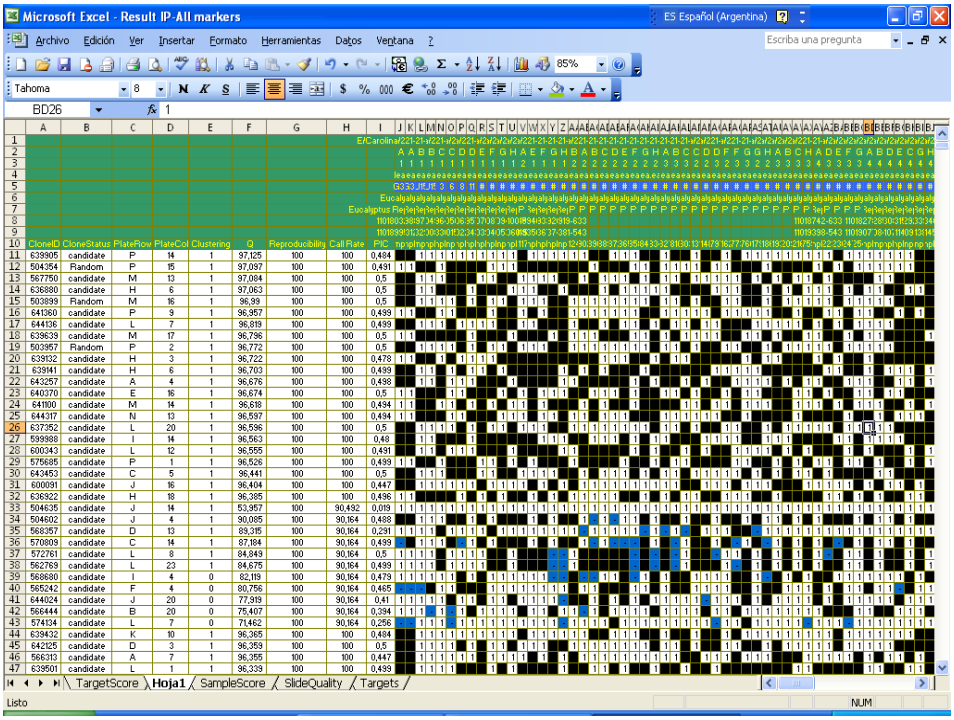


Figura 7. Exemplo de tabela de Microsoft Excel exportada pelo programa *DArTSoft*. Linhas são marcadores e colunas da esquerda para a direita são os números dos marcadores, seguidos pelas coordenadas do clone de DNA correspondente na biblioteca e os vários parâmetros de qualidade Q, Reprodutibilidade, Call Rate e PIC. Os genótipos são representados como "1" (em branco), "0" (em preto) e "-" ou dados perdidos (em azul).

3.2.8. Expansão das representações genômicas, análise de redundância e sequenciamento de clones DArT

Após a análise das pequenas representações genômicas no painel de descoberta, composto por 24 indivíduos, foram estimadas as proporções de fragmentos declarados polimórficos. O método de redução da complexidade que apresentou maior quantidade de marcadores polimórficos foi escolhido para expandir as bibliotecas seguindo o mesmo protocolo anteriormente descrito. Assim, foram gerados 16.128 clones no primeiro microarranjo protótipo e 7.680 no segundo microarranjo, totalizando 23.808 clones.

O desenvolvimento do arranjo envolve a seleção aleatória de clones a partir das bibliotecas. Esta prática carrega certo nível de redundância de clones (i.e. fragmentos de DNA com igual ou similar sobreposição de sequências), o qual produz uma superestimação do polimorfismo. A redundância dos clones DArT polimórficos foi analisada utilizando o pacote de programas *DArT ToolBox* (http://www.diversityarrays.com/). Este programa constrói uma matriz de distância

Hamming entre todos os clones DArT, dois a dois utilizando os dados genotípicos dos indivíduos analisados. Uma distância Hamming entre duas séries de mesmo tamanho é o número de posições nas quais os símbolos diferem (Hamming, 1950), no caso, o número de indivíduos para os quais os dois marcadores comparados diferem quanto à presença ou ausência do sinal. Clones com distância Hamming igual a zero são considerados redundantes e alocados a um mesmo grupo ou *Bin* considerado, assim, não redundante.

3.2.9. Análise de validação do microarranjo operacional

Para validar a utilidade do microarranjo operacional DArT para múltiplas aplicações em *Eucalyptus* foram realizadas três análises genéticas específicas. A primeira visou validar a utilização dos marcadores DArT em estudos de identificação individual e diversidade de espécies e transferibilidade de marcadores. Para isso foi preparado um painel de 96 amostras envolvendo um conjunto de 12 indivíduos geneticamente não relacionados de cada uma das sete principais espécies de *Eucalyptus* (*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis e E. cladocalyx)* e mais 12 indivíduos de *Corymbia variegata*. A segunda análise foi realizada para testar o comportamento dos marcadores em estudos de filogenia, onde foi utilizando um segundo painel chamado "filogenia", formado por amostras individuais de 62 espécies diferentes de *Eucalyptus*, sendo 56 de espécies pertencentes ao subgênero *Symphyomyrtus* e as outras espécies de outros três subgêneros (*Alveolata, Eucalyptus* e *Minutifructus*). Em ambos os estudos foram construídos dendrogramas utilizando o programa *DARwin5* (Perrier e Jacquemoud-Collet, 2006). O terceiro teste de validação envolveu uma amostragem de cinco populações de mapeamento, com a finalidade de observar o potencial dos marcadores DArT para estudos de mapeamento genético em populações interespecíficas. Para isso foi utilizado o programa *DArTSoft* e o aplicativo *Score Merge* do mesmo programa.

3.3. RESULTADOS E DISCUSSÃO

Os resultados deste capítulo descrevem as várias etapas do desenvolvimento do microarranjo para *Eucalyptus.* O primeiro passo foi encontrar o método mais apropriado de redução da complexidade genômica. Posteriormente foi desenvolvido e testado o microarranjo "Protótipo" utilizando a combinação de enzimas selecionada. Uma segunda etapa de expansão foi realizada criando o microarranjo "Operacional", o qual foi finalmente validado para genotipagem de espécies de *Eucalyptus* (Figura 8).

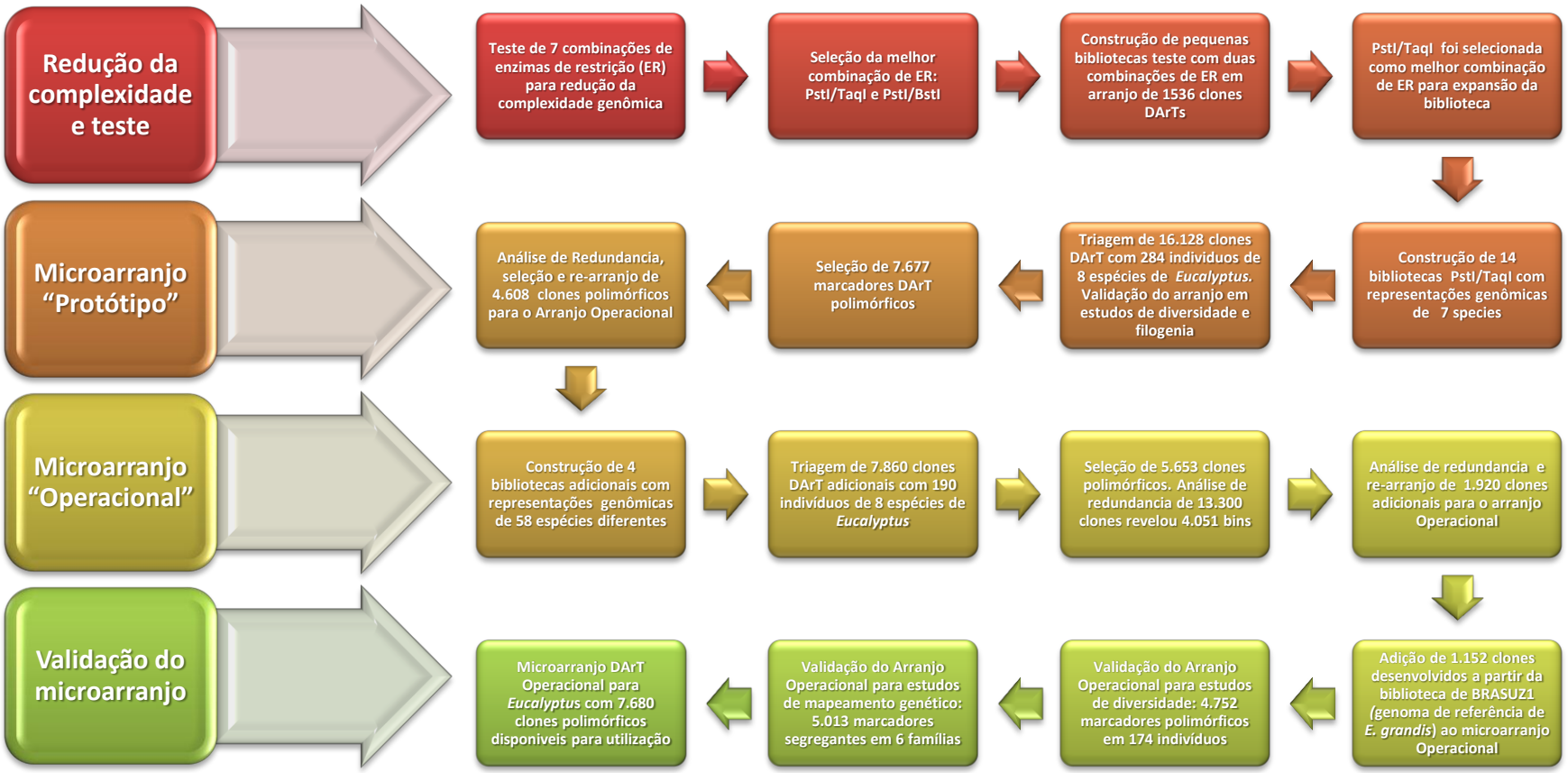


Figura 8. Fluxograma das etapas de desenvolvimento do microarranjo de genotipagem DArT para *Eucalyptus.*

3.3.1. Redução da complexidade genômica

A primeira etapa no desenvolvimento do microarranjo DArT foi a seleção do método de redução de complexidade genômica. Foram testadas sete combinações de enzimas de restrição (ER) em duas amostras de *Eucalyptus (E. grandis* e *E. urophylla)*. A ER de corte raro *PstI* foi utilizada em todos os casos como enzima primaria em combinação com enzimas de corte frequente (*TaqI, BstNI, MspI, HpaII, BanII, MseI ou AluI*) como enzimas secundarias (Figura 9). A enzima *PstI* é sensível à metilação CpG, pelo que exclui muito DNA repetitivo metilado da representação. A representação genômica produzida pela digestão com *PstI* em combinação com *TaqI (PstI/TaqI)* e *BstNI (PstI/BstNI)* foram consideradas as mais adequadas para *Eucalyptus* para avançar nas subsequentes etapas de desenvolvimento.

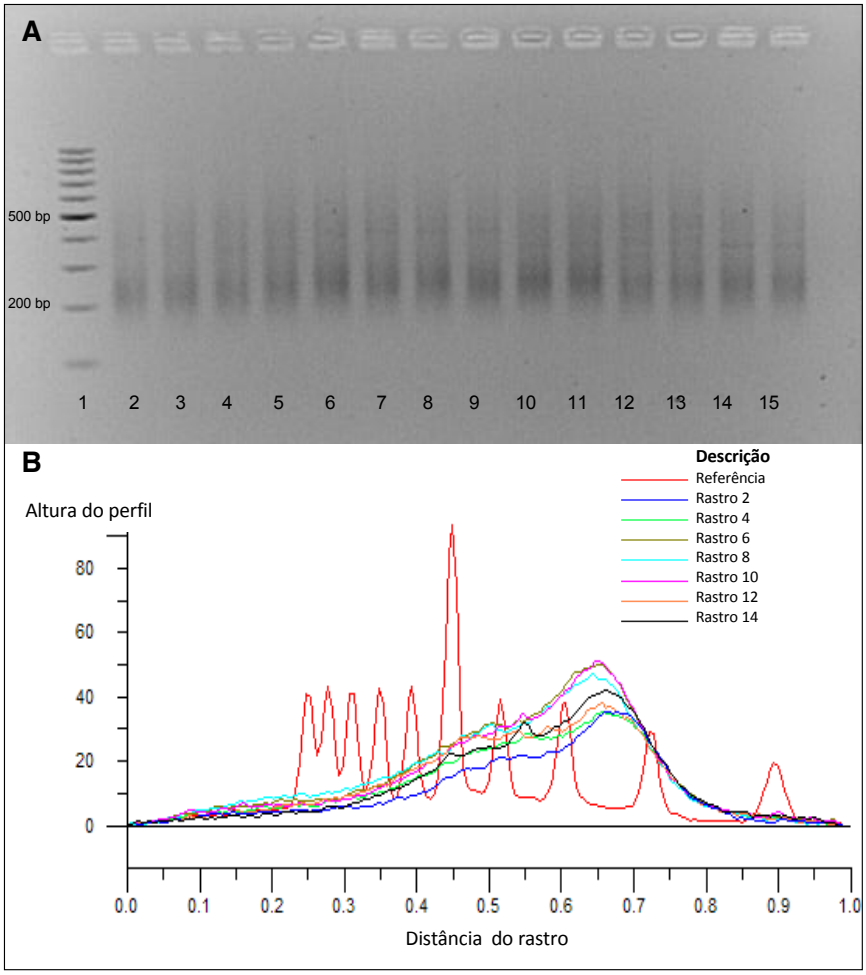


Figura 9. Resultado das sete combinações de enzimas de restrição testadas para a redução da complexidade genômica. Em **(A)** é observada a imagem do gel de agarose 1,2% mostrando a digestão com as diferentes enzimas de restrição em *E. grandis* e *E. urophylla* respectivamente: 2-3 *Pst*I(*Taq*I), 4-5 *Pst*I(*Bst*NI), 6-7 *Pst*I(*Msp*I), 8-9 *Pst*I(*Hpa*II), 10-11 *Pst*I(*Ban*II), 12-13

*Pst*I(*Mse*I), 14-15 *Pst*I(*Alu*I). Em (**B)** pode ser observado o perfil de intensidade da digestão do DNA de *E. grandis* obtido com os diferentes métodos de redução de complexidade. O marcador de fragmentos de tamanho conhecido (100bp ladder) é mostrado em vermelho. O rastro 2 (*PstI/TaqI* em azul) apresentou o perfil mais homogêneo e foi selecionado como a melhor combinação de enzimas de restrição para a redução de complexidade genômica.

3.3.2. Triagem das baterias piloto de clones DArT

A segunda etapa do desenvolvimento envolveu a construção de bibliotecas genômicas para cada um dos métodos de redução da complexidade selecionados e a triagem de um pequeno número de clones DArTs visando selecionar o método de redução que fornecesse maior proporção de clones DArT polimórficos. Para a construção destas bibliotecas dois conjuntos de amostras de DNA foram utilizados separadamente: o primeiro composto de 12 árvores de *E.grandis* e o segundo de 12 de *E.globulus*. Cada grupo de amostras foi digerido com combinações de ambas as enzimas: *PstI/TaqI* e *PstI/BstNI*. Portanto, quatro bibliotecas teste foram geradas, cada uma com 384 clones escolhidos aleatoriamente de cada uma das quatro bibliotecas, totalizando assim 1.536 clones DArT.

Os fragmentos de DNA clonados foram impressos em lâminas de vidro ou *slides* em duplicatas (aleatoriamente posicionados). Para cada uma das representações genômicas das 12 amostras de *E. grandis* e *E. globulus* foram gerados *targets* com duas réplicas que foram hibridizadas sobre o microarranjo. Cada *target* foi marcado com fluorescência verde (Cy3-dUTP) e cada réplica com fluorescência vermelha (Cy5-dUTP) e em seguida misturadas com fluorescência azul proveniente do polylinker do vetor utilizado para clonar os fragmentos o qual foi utilizado como controle de qualidade do arranjo. A mistura foi hibridizada sobre o microarranjo contendo 1.536 clones, os quais foram digitalizados para fluorescência azul, verde e vermelho.

Os dados foram extraídos utilizando o programa *DArTSoft* versão 7.44. O resultado da análise dos dois arranjos construídos com os diferentes métodos de redução da complexidade foram comparados em relação à frequência dos clones que revelaram marcadores DArT polimórficos, além do número de marcadores únicos revelados em cada método. Os critérios utilizados para declarar um clone como sendo um marcador polimórfico foram Reprodutibilidade > 97%, ou seja, aceitando um máximo de três discordâncias entre réplicas, e *Call Rate*> 80%. Dentre os dois métodos de redução da complexidade hibridizados, o gerado

pela combinação *PstI/TaqI* revelou uma maior proporção de marcadores polimórficos (21,7%) em comparação com o método gerado a partir da combinação *PstI/BstNI* (14,3%). Portanto o método de redução da complexidade escolhido para as etapas subsequentes de triagem de um maior número de clones DArT para a expansão do microarranjo foi *PstI/TaqI*.

3.3.3. Triagem da primeira bateria de clones DArT para seleção de clones polimórficos

Aos 1536 clones da triagem inicial foram adicionados 14.592 clones aleatoriamente amostrados, provenientes de 14 novas bibliotecas totalizando assim uma triagem de uma bateria de 16.128 clones (Tabela 1). Estas novas bibliotecas foram construídas com uma ampla variedade de genótipos representados por 254 amostras de sete espécies pertencentes aos dois gêneros mais importantes de eucaliptos (*Corymbia* e *Eucalyptus*) e os dois principais subgêneros de *Eucalyptus* (*Eucalyptus* e *Symphyomyrtus*). A ampla variedade de amostras é um ponto importante na estratégia de desenvolvimento do arranjo, já que permite aumentar a probabilidade de amostragem de segmentos genômicos que poderiam revelar marcadores polimórficos em uma ampla gama de origens genéticas (Wenzl, Carling *et al.*, 2004).

Tabela 1. Bibliotecas construídas e utilizadas para a triagem do primeiro microarranjo protótipo de marcadores DArT.

Nº de clones	Nº de amostras	Espécies*	Origem**
768	12	*Corymbia variegata*	Tasmania (AUS)
768	11	*E. camaldulensis*	Tasmania (AUS)
768	13	*E. globulus*	Portugal/Chile
1.920	12	*E. globulus*	Tasmania (AUS)
1.536	24	*E. globulus*	Tasmania (AUS)
768	12	*E. globulus*	West Australia
1.536	96	*E. grandis x E. urophylla*	Brasil
1.536	12	*E. grandis*	Tasmania (AUS)
2.688	9	*E. grandis*	South Africa
768	6	*E. nitens*	Chile
768	11	*E. nitens*	Tasmania (AUS)
768	12	*E. pilularis*	Tasmania (AUS)
768	12	*E. urophylla*	Tasmania (AUS)
768	12	*E. urophylla*	South Africa
16.128	254		

*_Corymbia variegata_ pertence a um gênero filogeneticamente relacionado com _Eucalyptus; E._camaldulensis, _E.globulus, E.grandis, E.urophylla e E.nitens_ pertencem ao subgenera _Symphyomyrtus; E.pilularis_ pertence ao subgênero _Eucalyptus._**Amostras obtidas de populações nativas ou de primeira geração de populações de melhoramento estabelecidas a partir de sementes provenientes de populações nativas.

Foram impressos slides incluindo todos os 16.128 clones e hibridizados com 284 indivíduos com réplicas completas representando oito espécies diferentes (Tabela 2). Com base na análise realizada pelo programa *DArTSoft*, foram identificados 7.677 marcadores polimórficos robustos, 47,6% do total de clones. A média de *Call Rate* observada entre os marcadores polimórficos selecionados foi de 95,3% e a média de reprodutibilidade neste mesmo grupo de marcadores foi de 99,7%.

A hibridização de amostras provenientes de *Corymbia* sobre o microarranjo composto majoritariamente de sondas de *Eucalyptus* (e vice versa) mostrou claramente que o sinal dos *targets* de *Corymbia* era baixo e não correlacionado com o sinal das espécies de *Eucalyptus* (e vice versa). Esta baixa transferibilidade observada entre os gêneros dificulta a descoberta de novos clones polimórficos para *Eucalyptus.* O desenvolvimento de DArT para o gênero *Corymbia* foi portanto abandonado focando exclusivamente em *Eucalyptus*.

Tabela 2. Espécies de *Eucalyptus* e número de indivíduos de cada espécie (total de 284) utilizados como *targets* para a triagem do primeiro microarranjo protótipo visando a seleção de marcadores polimórficos.

Nº de amostras	Espécies
135	*E. grandis x E. urophylla*
28	*E. pilularis*
27	*E. nitens*
35	*E. globulus*
12	*E. cladocalyx*
12	*E. grandis*
12	*E. urophylla*
12	*Corymbia variegata*
11	*E. camaldulensis*

Como a seleção de clones para construir as bibliotecas é um processo aleatório, certo nível de redundância de clones é incorporado ao microarranjo produzindo uma superestimação do polimorfismo. A análise de redundância calculada com o programa DArT

Tool Box foi realizada com a informação de genótipos das 284 amostras hibridizadas no microarranjo listadas na Tabela 2. Esta estimativa de redundância de clones obtida exclusivamente da distância Hamming entre clones calculada com base em genótipos, possibilitou a seleção de clones únicos ou com baixa redundância. Os 7.677 clones que revelaram polimorfismo pertenciam a estimados 2.652 *bins* não redundantes i.e. 34,5% (Tabela 3). Destes 2.652 *bins*, 1.330 *bins* (50,15% do total de *bins*) apresentaram um único clone por *bin*, ou seja foram sequências não redundantes, 1.199 *bins* (45,21%) tiveram de 2 a 9 clones por bin, 105 *bins* (3,96%) incluíam 10 a 19 clones por *bin*, 9 *bins* (0,34%) continham 20 á 29 clones, e 9 *bins* (0,34%) possuíam mais de 29 clones.

Tabela 3. Distribuição de clones DArT polimórficos dentro de cada classe de bin no primeiro microarranjo protótipo de triagem.

Nº de clones por bin	Nº de bins
1	1.330
2 - 9	1.199
10 - 19	105
20 - 29	9
30 - 39	4
40 - 49	2
>50	3
Total	**2.652**

A análise de redundância realizada com base em distância Hamming entre marcadores, envolve, entretanto, uma elevada probabilidade de superestimar a redundância ao agrupar em um mesmo *bin* sequências que na verdade são únicas. Para verificar a suspeita de superestimativa de redundância, foram sequenciados 134 clones amostrados ao acaso a partir de 9 *bins* que continham 30 ou mais clones (Tabela 4). O resultado deste sequenciamento revelou que dos 134 clones sequenciados a partir de nove bins e que, a princípio, representariam apenas 9 sequências não redundantes, na verdade foram recuperados 73 clones de sequência única, ou seja, 54%. Estes resultados de uma pequena amostra de clones sequenciados revelaram que a análise baseada em distância *hamming* é por demais estringente levando a uma superestimativa de redundância (Tabela 4). Embora esta elevada estringência seja positiva, pois tende a evitar redundância de sondas no microarranjo final e aumentar a cobertura genômica da genotipagem no final desta análise foram selecionados 4.608 clones que revelaram marcadores de alta qualidade, mantendo aproximadamente 30%

de potencial redundância, com uma frequência de seleção proporcional ao número de clones pertencentes ao bin.

Tabela 4. Resultado da análise de redundância de clones DArT com base no sequenciamento de clones amostrados em nove *bins* não redundantes declarados com base em distâncias *Hamming*.

# do bin	Nº de clones redundantes estimados com base em distância Hamming)	Nº de clones sequenciados	Nº de clones com sequencias de DNA únicas	% de clones de sequências únicas
1	116	33	19	57,6
2	75	20	12	60,0
3	59	17	8	47,1
4	43	13	6	46,2
5	41	6	4	66,7
6	39	10	5	50,0
7	37	16	11	68,8
8	31	10	6	60,0
9	30	9	2	22,2
Total		**134**	**73**	

3.3.4. Triagem da segunda bateria de clones DArT para seleção de clones polimórficos e montagem do microarranjo DArT operacional

Com o objetivo de aumentar a representatividade genômica do microarranjo DArT para análise genética de *Eucalyptus*, buscou-se aumentar a diversidade de fragmentos amostrados a partir de um número maior de espécies que representasse da melhor maneira possível a ampla variabilidade genética do gênero. Com este fim foram construídas quatro novas bibliotecas e isolados 7.680 novos clones de DNA para serem submetidos à triagem (Tabela 5). A primeira biblioteca foi construída a partir de um conjunto de amostras de seis pedigrees híbridos interespecíficos. A segunda biblioteca foi desenvolvida utilizando unicamente uma amostra da árvore BRASUZ1 (*E. grandis*), cujo genoma foi utilizado para a geração do genoma de referencia de *Eucalyptus.* As outras duas bibliotecas continham DNA de

62 espécies de *Eucalyptus* que foram construídas misturando quantidades equimolares de DNA de um individuo de cada espécie e cortando com dois métodos de redução de complexidade, *PstI* e *PstI/TaqI*. As representações digeridas unicamente com *PstI* ofereceram marcadores que estavam presentes em baixa frequência nas representações *PstI/TaqI* a ser clonadas e por tanto diminuíram o nível de redundância no conjunto final de clones. Das 62 espécies, 56 pertenciam ao subgênero *Symphyomyrtus* (representando 14 das 15 seções), enquanto que as outras espécies pertenciam a outros três subgêneros (*Alveolata, Eucalyptus* e *Minutifructus*).

Tabela 5. Origem dos clones DArT amostrados para a segunda bateria de triagem.

Biblioteca	**Espécies**	**Nº de amostras**	**Origem**	**Método de digestão**	**Nº de clones**
1	*E. grandis X E. urophylla (IP)*	16	Brazil	*PstI (TaqI)*	1.920*
	E. grandis X E. urophylla (VCP)	16	Brazil	*PstI (TaqI)*	
	E. camaldulensis X (E. urophylla x E. globulus)	16	Brazil	*PstI (TaqI)*	
	(E.grandis x E. urophylla) X (E. urophylla xE. globulus)	16	Brazil	*PstI (TaqI)*	
	(E. dunnii x E. grandis) X (E. urophylla xE. globulus)	16	Brazil	*PstI (TaqI)*	
	(E. dunnii x E. grandis) X E. urophylla	16	Brazil	*PstI (TaqI)*	
2	*Eucalytus grandis (BRASUZ1)*	1	Brazil	*PstI (TaqI)*	1.152**
3	*	62	Tasmania (AUS)	*PstI (TaqI)*	2.304***
4	*	62	Tasmania (AUS)	*PstI*	2.304***
	Total de clones				**7.680**

*Biblioteca construída com um conjunto de quantidades equimolares de DNA de 96 híbridos interespecíficos.**Biblioteca construída com DNA da árvore BRASUZ1 (E. grandis), cujo genoma foi sequenciado.***As seguintes espécies têm sido utilizadas para a construção da biblioteca: *E. albaleaf, E. albens, E. balladoniensis, E. bicostata, E. biterraneana, E. brassiana, E. brevistylis, E. camaldulensis, E. cladocalyx, E. coolabah, E. cordata, E. cornuta, E. cosmophylla, E. crebra, E. dalrympleana, E. deglupta, E. delicata, E. diversicolor, E. dundasii, E. dunnii, E. falcata, E. glaucescens, E. glaucina, E. globulus, E. gomphocephala, E. grandis, E. gunnii, E. hallii, E. houseana, E. howittiana, E. leucophloia, E. lockyeri, E. longifolia, E. lucasii, E. maidenii, E. michaeliana, E. microcorys, E. morrisbyi, E. nitens, E. obtusiflora, E. optima, E. ovata, E. pachycalyx, E. pachyphylla, E. paludicola, E. perriniana, E. polyanthemos, E. populnea, E. pseudoglobulus, E. pulverulenta, E. pumila, E. raveritiana, E. rubida, E.*

salmonophloia, E. scoparia, E. stoatei, E. tereticornis, E. torquata, E. urophylla, E. viminalis, E. wandoo, E. woodwardii

Para a triagem desta segunda bateria de 7.680 clones foram utilizados 190 indivíduos com réplicas completas de 7 espécies diferentes de *Eucalyptus* (*E. grandis, E. urophylla, E. camaldulensis, E. globulus, E. dunni, E. pilularis* e *E. nitens)* (Tabela 6).

Tabela 6. Populações de *Eucalyptus* e o correspondente número de indivíduos utilizados como *target* para a triagem da segunda bateria de 7.680 clones.

Nº de indivíduos	Espécies
71	*E. grandis X E. urophylla*
16	*E. camaldulensis X (E. urophylla x E. globulus)*
16	*(E.grandis x E. urophylla) X (E. urophylla x E. globulus)*
16	*(E. dunnii x E. grandis) X (E. urophylla x E. globulus)*
16	*(E. dunnii x E. grandis) X E. urophylla*
16	*E. pilularis*
16	*E. nitens*
23	*E. globulus*

Para a identificação de clones reveladores de marcadores polimórficos robustos e estimar o nível de redundância nesta segunda bateria, foram utilizados os programas *DArTSoft* e *DArT ToolBox* com os mesmos parâmetros de análise utilizados na primeira bateria de 16.128 clones. Dos 7.680 clones testados, *DArTSoft* detectou 5.653 revelando marcadores polimórficos (73,6%). A média de *Call Rate* e Reprodutibilidade foi similar à primeira bateria com 93,7% e 99,7% respectivamente. Entretanto, uma porcentagem mais alta de sondas revelando polimorfismo foi detectada nessa segunda bateria de clones, 73,6% em comparação a 47,6% da primeira bateria. Esta substancial diferença observada muito provavelmente foi derivada da maior diversidade genética capturada nas representações genômicas nas quatro novas bibliotecas que envolveram uma ampla gama de espécies do gênero.

Com o objetivo de minimizar a inclusão de marcadores redundantes entre a primeira e a segunda bateria de clones DArT avaliados, foi realizada uma análise consolidada de redundância envolvendo os 7.677 clones polimórficos revelados no primeiro arranjo e os 5.653 detectados na segunda bateria, totalizando assim 13.330 clones. Esta análise detectou 4.051

bins não redundantes, 30,4% (Tabela 7). Destes 4.051 *bins*, 2.143 apresentavam um único clone, ou seja sequências não redundantes (52,9% do total de *bins*), 1.737 *bins* tinham entre 2 a 9 clones (42,9%), 126 possuíam entre 10 a 19 clones por *bin* (3,1%), 17 *bins* apresentavam de 20 a 29 clones (0,4%) e finalmente 28 *bins* estavam formados por mais de 30 clones cada um (0,7%). A distribuição dos *bins* foi similar entre as duas baterias de clones submetidos à triagem.

Tabela 7. Distribuição de clones DArT polimórficos dentro de cada classe de bin da segunda bateria de clones submetidos à triagem.

Nº de clones por *bin*	**Nº de *bins***
1	2.143
2 - 9	1.737
10 - 19	126
20 - 29	17
30 - 39	8
40 - 49	3
>50	17
Total	**4.051**

Para minimizar a inclusão de clones polimórficos redundantes entre as duas baterias de clones analisados durante a montagem final do microarranjo operacional, foram escolhidos os bins que apresentavam de 1 a 7 clones. Estes *bins* foram classificados em três grupos, os que continham clones provenientes exclusivamente da primeira bateria, os *bins* cujos clones pertenciam exclusivamente à segunda bateria de clones e os *bins* que possuíam clones derivados de ambas as baterias (Tabela 9). Em cada classe de número de clones por *bin*, em media 52% dos *bins* eram oriundos da primeira bateria, 25% da segunda bateria e os restantes 23% continham clones derivados de ambas as baterias (Tabela 8). Nos *bins* com mais de quatro clones, o número de marcadores oriundos da segunda bateria foi reduzida de forma acentuada, e os *bins* maiores, com mais de sete clones, não contribuíam para a identificação de clones polimórficos adicionais uma vez que já tinham sido amostrados na primeira bateria. Portanto baseado nestas análises de polimorfismo e estimativa de redundância, foram selecionados 1.920 clones dos 7.680 clones da segunda bateria de clones DArT submetidos à triagem.

Tabela 8. Distribuição e origem de *bins* a partir dos quais clones DArT foram selecionados para integrar o microarranjo operacional DArT

Nº de clones por bin	Nº de *Bins*	Derivados da primeira bateria de clones	%	Derivados da segunda bateria de clones	%	Derivados de ambas as baterias de clones	%
1	2.143	1.201	56,0	942	44,0	0	0,0
2	774	395	51,0	265	34,2	114	14,7
3	395	188	47,6	128	32,4	79	20,0
4	196	91	46,4	48	24,5	57	29,1
5	122	69	56,6	19	15,6	34	27,9
6	84	44	52,4	12	14,3	28	33,3
7	72	39	54,2	8	11,1	25	34,7
Total	***3.786***	***2.027***		***1.422***		***337***	

Em resumo, o desenvolvimento do microarranjo envolveu a triagem de um total de 23.808 clones DArT derivados de 18 bibliotecas genômicas de complexidade reduzida. Deste total, foram identificados 13.300 clones que revelaram marcadores robustos e polimórficos. A análise de redundância dos 13.300 clones resultou em 4.051 grupos de clones não redundantes com base em distância Hamming. Para a montagem do microarranjo operacional foram selecionados 6.528 clones mantendo aproximadamente 30% da redundância estimada (Tabela 9). Para a montagem final do microarranjo operacional, aos 6.528 clones selecionados, foi adicionado um conjunto de 1.152 clones selecionados desenvolvidos a partir de uma biblioteca genômica da árvore BRASUZ1, a árvore de *E. grandis* cujo genoma foi sequenciado visando incluir no microarranjo uma quantidade substancial de clones que facilitassem a ancoragem futura de mapas genéticos com a sequência do genoma. Todos estes clones constituíram o microarranjo operacional DArT para *Eucalyptus* com 7.680 marcadores o qual vem sendo utilizado hoje com sucesso para diferentes aplicações na análise genética de Eucalyptus (Hudson, Kullan *et al.*, 2011; Kullan, Van Dyk *et al.*, 2011; Steane, Nicolle *et al.*, 2011; Resende, Resende *et al.*, 2012). As etapas de desenvolvimento deste microarranjo operacional foram detalhadas em um artigo científico (Sansaloni, Petroli *et al.*, 2010) (Anexo 1).

Tabela 9. Resultados do desenvolvimento do microarranjo DArT em *Eucalyptus* na fase protótipo e operacional incluindo a triagem de marcadores polimórficos e análise de redundância.

Etapa de desenvolvimento da tecnologia	Nº de clones DArT avaliados	Nº e (%) de clones polimórficos identificados	Nº de *bins* não redundantes	Nº e (%) de clones polimórficos selecionados
Primeira triagem de clones	16.128	7.677 (47,6%)	2,652 (16,4%)	4.608
Segunda triagem de clones	7.680	5,653 (73,6%)	n.d.*	1.920
Total de clones	23.808	13,300 (55,9%)	4.051 (17,0%)	6.528

*n.d. não definido.

3.3.5. Validação da plataforma de genotipagem DArT para *Eucalyptus* em estudos de diversidade e filogenia.

O presente trabalho de construção de um microarranjo DArT para *Eucalyptus* foi o primeiro a ser desenvolvido visando aplicações em mapeamento genético e estudos de diversidade genética e filogenia no gênero. Para avaliar o desempenho potencial da tecnologia DArT especificamente em estudos de diferenciação de espécies e reconstrução de populações silvestres, foram utilizados dois painéis de indivíduos. O primeiro painel, denominado de "painel de diversidade" era composto por um conjunto de 12 indivíduos geneticamente não relacionados de cada uma das sete principais espécies de *Eucalyptus* (*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis e E. cladocalyx)* e 12 indivíduos de *Corymbia variegata*. O segundo painel denominado "painel de filogenia" era formado por amostras individuais de 62 espécies diferentes de *Eucalyptus*, sendo 56 destas espécies pertencentes ao subgênero *Symphyomyrtus* e as outras espécies pertencentes a outros três subgêneros (*Alveolata, Eucalyptus* e *Minutifructus*). Os experimentos foram realizados em duplicata utilizando replicas de todas as amostras com o objetivo de aumentar a acurácia das análises. A avaliação foi realizada já ao longo do desenvolvimento do microarranjo, especificamente na etapa de triagem da primeira bateria de 16.128 clones DArT. Deste total de 16.128 clones foram identificados 7.960 marcadores polimórficos entre e dentro das oito espécies e com elevada reprodutibilidade no painel de diversidade (49,35%) (Figura 10). Uma vez determinada

a porcentagem de marcadores polimórficos detectados entre as oito espécies, foram eliminadas em diferentes etapas de análise as espécies mais distantes filogeneticamente visando avaliar o impacto na capacidade do microarranjo em posicionar corretamente as espécies em termos filogenéticos (Figuras 11, 12, 13 e 14). Com isso foi possível estimar a variação da porcentagem de marcadores informativos ao se reduzir a diversidade de espécies do painel (Tabela 10). Quando eliminada a espécie *Corymbia variegata* da análise observou-se uma diminuição de 5,69% no número de marcadores polimórficos. Na segunda etapa de eliminação foram mantidas unicamente as espécies pertencentes ao subgênero *Symphyomyrtus* onde se observou uma diminuição de 12,46% do polimorfismo. Quando foram utilizados para a análise somente as duas espécies mais próximas geneticamente (*E. grandis* e *E. urophylla*) o polimorfismo caiu somente 4,53%. Esta ultima diminuição da porcentagem foi similar (4,57%) ao analisar unicamente a espécie *E. grandis.* Estes resultados mostraram que mesmo com uma diminuição da diversidade de espécies que formaram o painel de análise, a proporção de clones DArT que revelam marcadores polimórficos continou elevada, demostrando não apenas o poder da técnica mas também uma substancial taxa de transferibilidade de marcadores entre espécies. Além disso, o posicionamento observado dos indivíduos das várias espécies nos dendogramas foi totalmente consistente com o esperado segundo o relacionamento filogenético conhecido. *E. grandis* e *E. urophylla* pertencem à mesma seção *Latoangulatae*; *E. camaldulensis* espécie predominantemente tropical pertence à seção *Exsertaria*, filogeneticamente mais próxima de *E. grandis* e *E. urophylla*. *E. globulus* e *E. nitens* são espécies subtropicais pertencentes à seção *Maidenaria*, filogeneticamente mais distante das tropicais. *E. cladocalyx* por sua vez pertence à seção monotípica, *Sejunctae*, ou seja com apenas esta espécie e filogeneticamente distinta das demais. Finalmente *E. pilularis* pertence a um outro subgênero, Monocalyptus, claramente distinto das espécies do subgênero Symphyomyrtus enquanto que *Corymbia variegata* representa outro gênero (Brooker, 2000).

Tabela 10. Análise de polimorfismo em painéis de diversidade decrescente de espécies.

Nº de espécies	**Espécies**	**Nº de clones DArTs polimórficos**
8	*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis e E. cladocalyx e*	7.960

	Corymbia variegata	
7	*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis* e *E. cladocalyx*	7.043
5	*E. nitens, E. globulus, E. urophylla, E. grandis* e *E. camaldulensis*	5.033
2	*E. grandis* e *E. urophylla*	4.302
1	*E. grandis*	3.565

No segundo painel de filogenia o número de marcadores polimórficos aumentou para 8.959 (61,39%), consistente com a elevada diversidade representada pelas 62 espécies de *Eucalyptus* analisadas. Estes marcadores polimórficos permitiram diferenciar todos os indivíduos e espécies com elevada precisão e consistência com a relação filogenética conhecida entre as espécies (Figura 15). Parte destes dados foi utilizada para gerar um artigo científico publicado na revista *Molecular Phylogenetics and Evolution* (Steane, Nicolle *et al.*, 2011)(Anexo2).

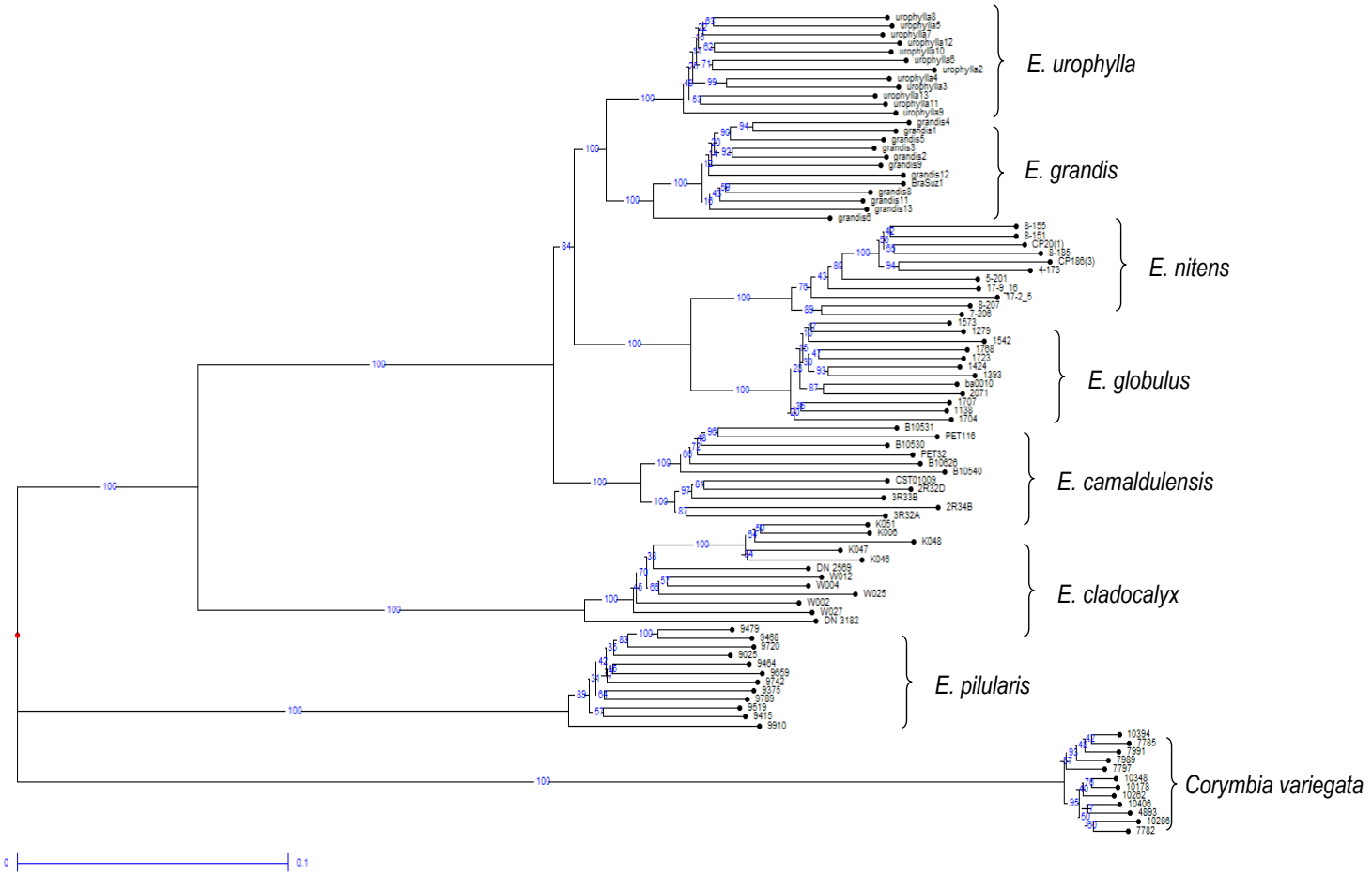


Figura 10. Dendrograma Neighbor Joining construído com 7.960 marcadores polimórficos, mostrando o posicionamento de. *E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis e E. cladocalyx* do gênero *Eucalyptus* e *Corymbia variegata* do género *Corymbia.*

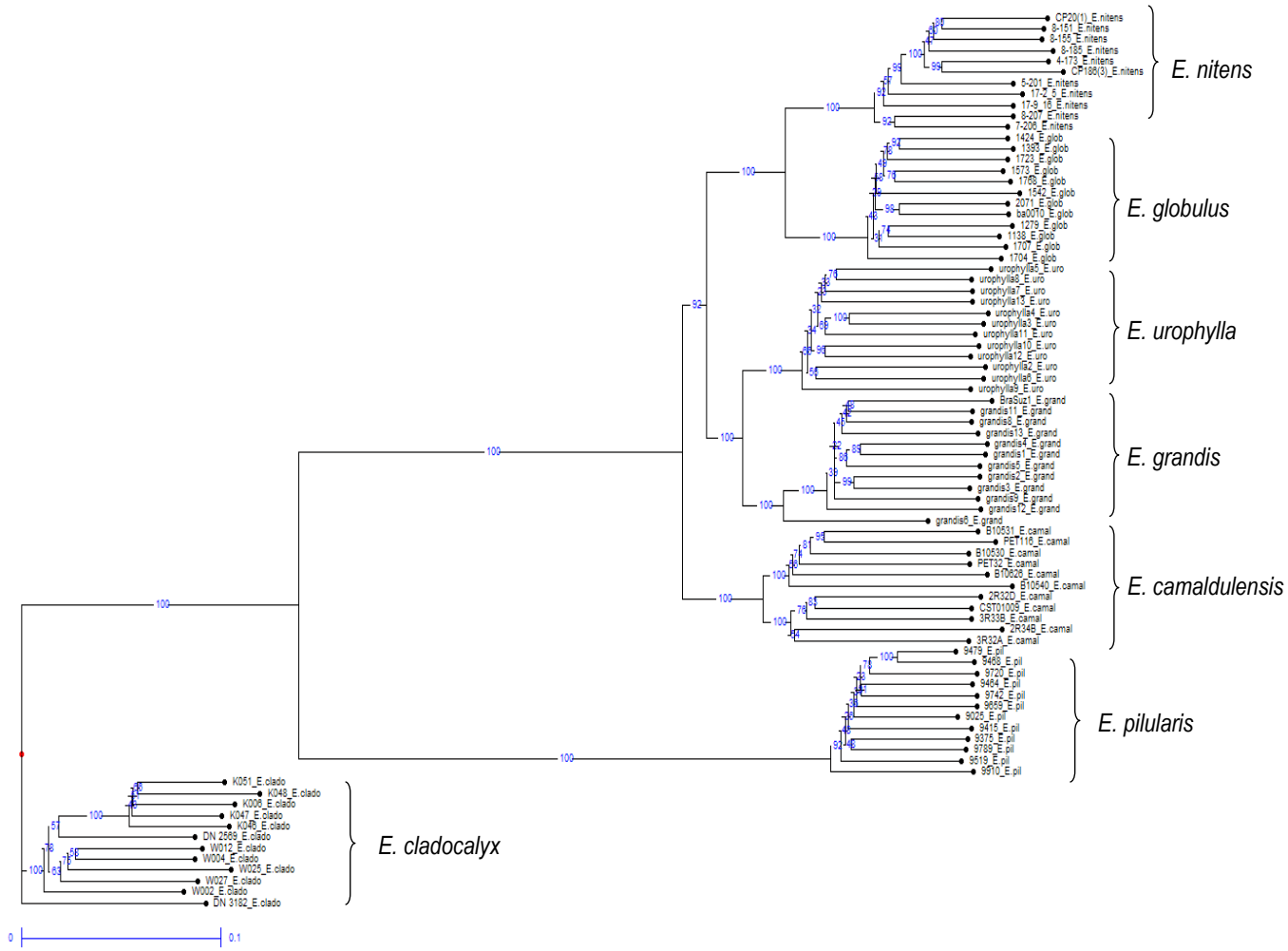


Figura 11. Dendrograma Neighbor Joining construído com 7.043 marcadores DArT polimórficos, mostrando o posicionamento de sete espécies do gênero *Eucalyptus* (*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis, E. pilularis e E. cladocalyx).*

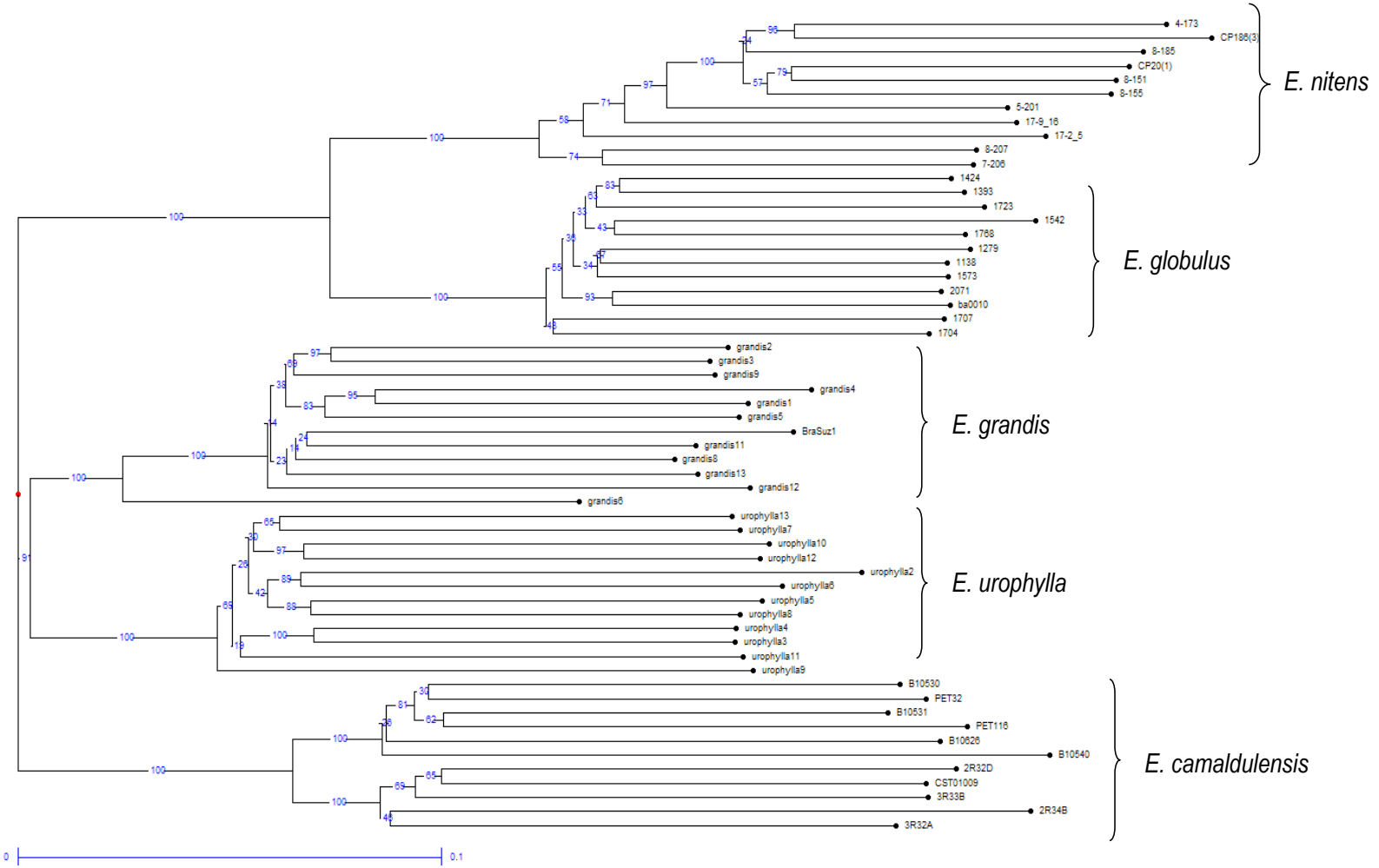


Figura 12. Dendrograma Neighbor Joining construído com 5.033 marcadores polimórficos, mostrando o posicionamento de 58 amostras de cinco espécies do subgênero *Symphyomyrtus* (*E. nitens, E. globulus, E. urophylla, E. grandis, E. camaldulensis).*

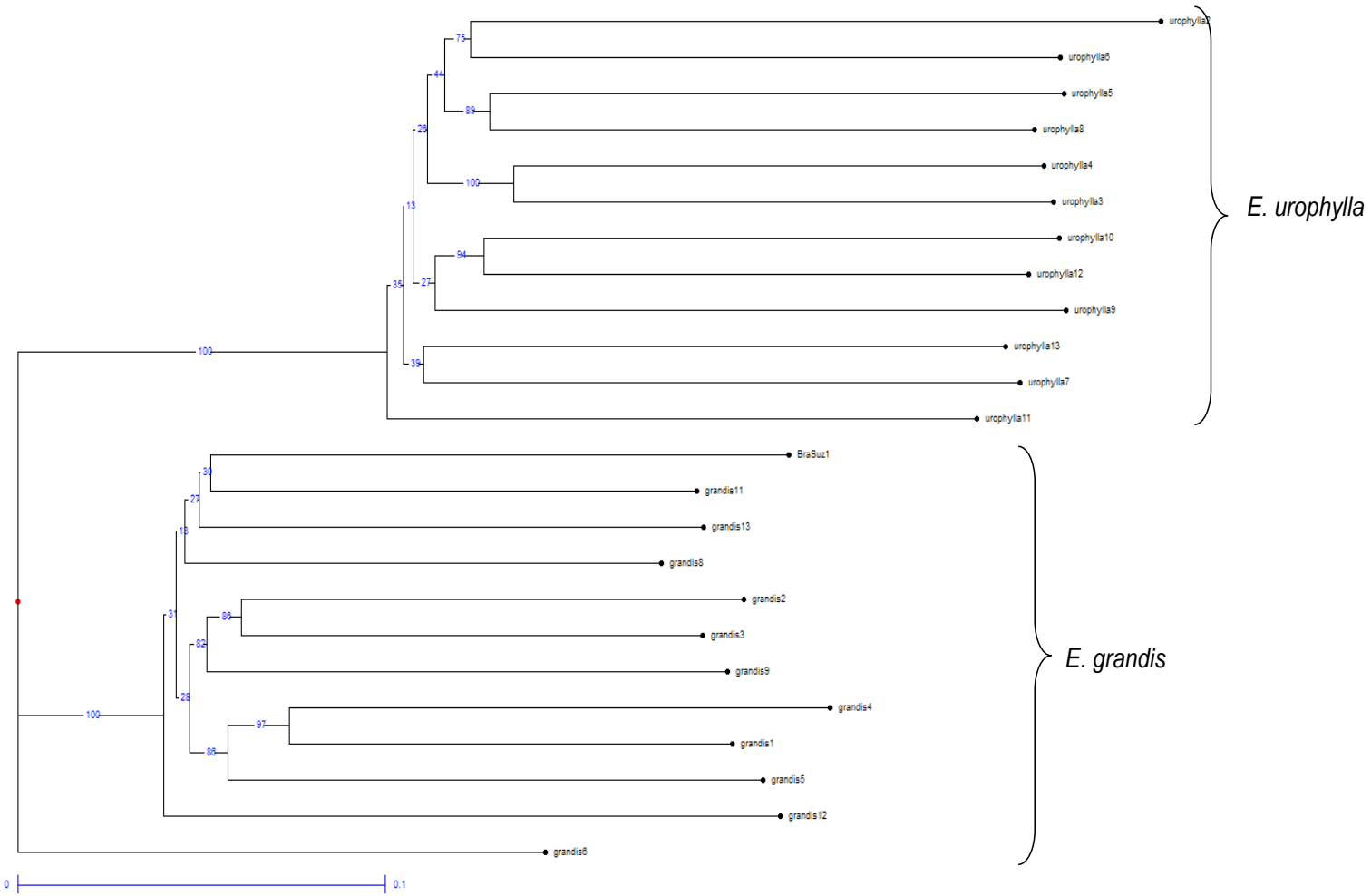


Figura 13. Dendrograma Neighbor Joining construído com 4.302 marcadores polimórficos, mostrando o posicionamento de 24 amostras das espécies *E. urophylla* e *E. grandis*, ambas pertencentes ao subgênero *Symphyomyrtus* e mesma seção Latoangulatae.

Figura 14. Dendrograma Neighbor Joining construído com 3.565 marcadores polimórficos, mostrando identificação individual de 12 amostras de *E. grandis*

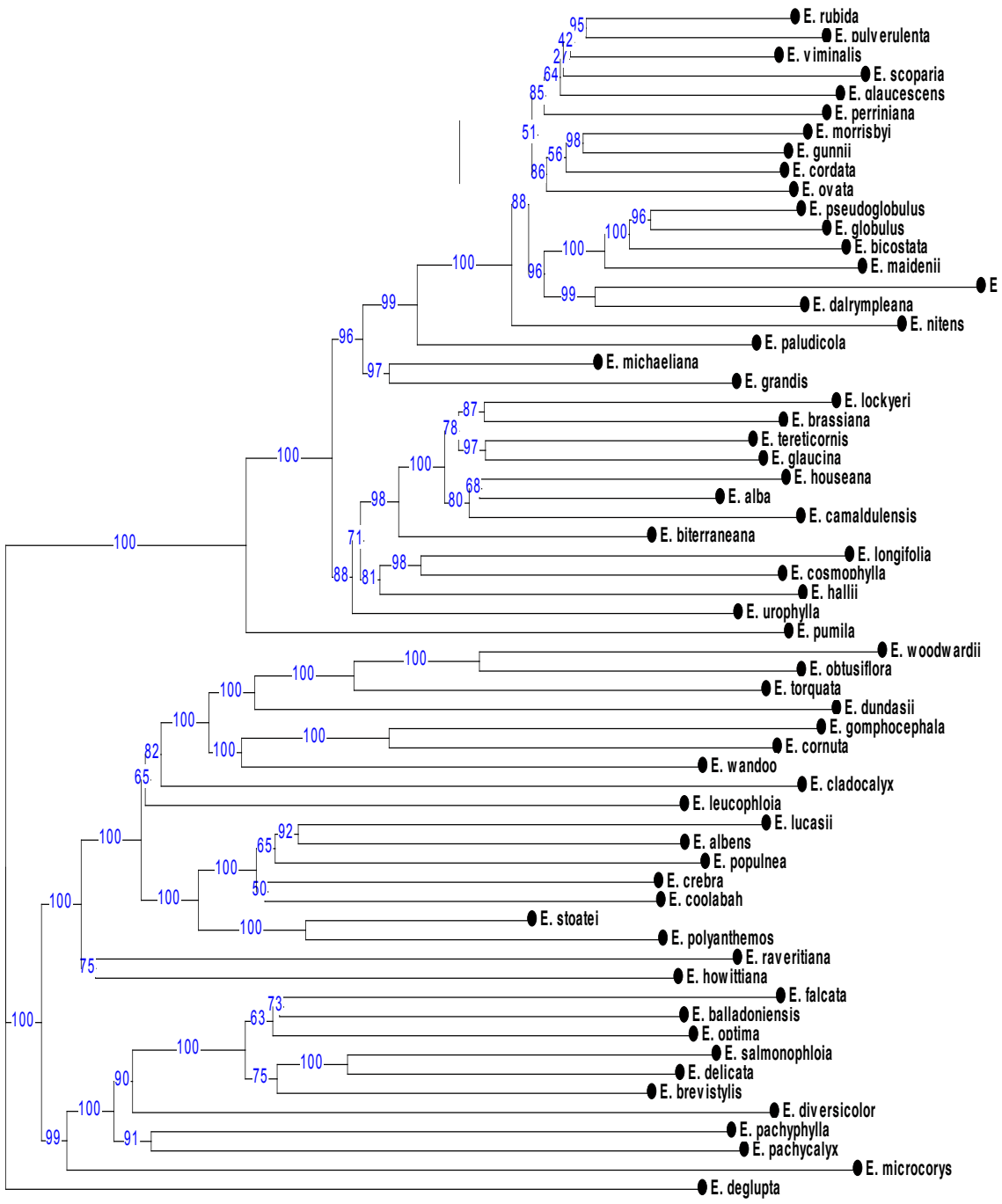


Figura 15. Dendograma Neighbor Joining representando as relações filogenéticas entre 62 espécies de *Eucalyptus*.

Uma etapa adicional de validação foi em seguida realizada já com o microarranjo operacional com 7680 clones selecionados. Foram analisados 174 indivíduos de seis espécies de *Eucalyptus*, as mesmas usadas para criar as bibliotecas genômicas das quais derivaram boa parte dos clones testados na segunda bateria de triagem (*E. grandis, E. urophylla, E. dunni, E. camaldulensis, E. globulus* e *E. nitens*). A análise de polimorfismo revelou 4.752 marcadores polimórficos entre os 7.680 clones (61,9%). Conforme esperado, nem todos os 7.680 clones se revelaram polimórficos uma vez que as 174 amostras estudadas não representavam o total da diversidade genética utilizada para a construção do arranjo.

3.3.6. Validação da plataforma de genotipagem DArT para mapeamento genético em *Eucalyptus*

Uma segunda validação do arranjo DArT operacional foi realizada por meio da observação da segregação e taxa de polimorfismo em diferentes populações de mapeamento. Nesta análise foram utilizadas 16 amostras de cada um de seis pedigrees de mapeamento totalizando 96 indivíduos. As seis populações eram compostas por híbridos interespecíficos envolvendo cinco espécies de *Eucalyptus* (*E. grandis, E. urophylla, E. globulus, E. calmadulensis* e *E. dunni*). Em média, foram detectados 2.211 marcadores polimórficos por pedigree (Tabela 11). O número de marcadores polimórficos compartilhados (polimórficos em dois pedigrees) entre as seis populações de mapeamento variou desde um mínimo de 859 marcadores a um máximo de 1.328 marcadores (Tabela 11). Observou-se que dos 7.680 clones presentes no microarranjo operacional de *Eucalyptus*, um total de 5.013 marcadores (65,3%) tinham segregado em pelo menos uma população de mapeamento, quando os dados dos seis pedigrees foram consolidados (Tabela 12). A análise também revelou que ao se analisar conjuntamente os seis pedgrees derivados de diferentes espécies um total de 150 marcadores segregaria e seria portanto informativo em todos eles permitindo uma comparação dos mapas de eventuais posições de QTLs entre eles caso este fosse o objetivo (Tabela 12).

Tabela 11. Número de marcadores polimórficos identificados em cada população de mapeamento de *Eucalyptus* (diagonal) e entre populações de mapeamento (acima da diagonal).

	C1xUGl	***DGxU2***	***DGxUGl***	***IP(GxU)***	***UGlxGU***	***VCP(GxU)***
C1xUGl	**2.394**	864	1.328	1.123	1.172	899
DGxU2		**1.818**	1.154	1.029	866	859
DGxUGl			**2.465**	1.251	1.284	953
IP(GxU)				**2.553**	1.175	1.144
UGlxGU					**2.176**	946
VCP(GxU)						**1.861**

C1 × UGl =*E. camaldulensis* × (*E. urophylla* × *E. globulus*); **DG × U =** (*E. dunnii* × *E. grandis*) × *E. urophylla*; **DG × UGl =** (*E. dunnii* × *E. grandis*) × (*E. urophylla* × *E. globulus*); **G × U(IP) =***E. grandis* × *E. urophylla* (pedigree IP); **UGlxGU =** (*E. urophylla* × *E. globulus*) × (*E.grandis* × *E. urophylla*); **G × U(VCP) =***E. grandis* × *E. urophylla* (VCP pedigree).

Tabela 12. Poder informativo dos marcadores DArT do microarranjo operacional para mapeamento genético, baseado na amostragem de seis populações de mapeamento (informação dos pedigrees na Tabela 10).

Nº de populações de mapeamento na qual o marcador foi polimórfico	**Nº de marcadores DArT polimórficos**	**% do número total de marcadores no microarranjo**
1	1.407	28,07%
2	1.154	23,02%
3	1.048	20,91%
4	761	15,18%
5	493	9,82%
6	150	3,0%
Total	**5.013**	**100%**

3.4. CONCLUSÕES

O microarranjo DArT desenvolvido para espécies de *Eucalyptus* é um dos microarranjos de melhor desempenho desenvolvidos até hoje do ponto de vista de recuperação de marcadores polimórficos de alta qualidade (DART Pty Ltd, resultados não publicados). A elevada proporção de sequências revealadoras de polimorfismo resulta em grande parte da domesticação recente e do modo de reprodução alógamo das espécies de

Eucalyptus (Myburg, Potts *et al.*, 2007). O alto nível de diversidade de sequência de espécies de *Eucalyptus* (Novaes, Drost *et al.*, 2008; Külheim, Yeoh *et al.*, 2009) representa um obstáculo para o desenvolvimento de plataformas altamente multiplexados de SNP como Golden Gate que normalmente requerem trechos razoavelmente longos de sequência sem SNPs secundários. Dificilmente seriam encontrados SNPs com estas características em uma ampla gama de espécies do gênero, fato este recentemente comprovado pela baixa transferibilidade de SNPs polimórficos entre espécies de *Eucalyptus* (Grattapaglia, Silva-Junior *et al.*, 2011). Marcadores DArT, por outro lado, pela sua natureza baseada em hibridização de longas sondas imobilizadas no arranjo, sofrem menos com isso e são portanto muito adequados para a genotipagem de milhares de marcadores genéticos em organismos com elevada heterozigosidade.

O microarranjo DArT produzido neste trabalho forneceu um número substancialmente maior de marcadores polimórficos e robustos para o *Eucalyptus* do que tecnologias anteriores (Grattapaglia e Kirst, 2008). Embora microssatélites co-dominantes sejam significativamente mais informativos em termos de loco individual, eles não permitem a genotipagem com alto desempenho e são caros do ponto de vista de custo por dado ponto (*data point*). A comparação de DArT com a análise de marcadores RAPD ou AFLP é mais apropriada uma vez que são todos marcadores dominantes. A questão que surge, entretanto, é o viés de amostragem e verificação (*ascertainment bias*) que ocorre ao selecionar primers RAPD, ou combinações primers/enzima para AFLP ou sondas DArT polimórficas. Esse viés é agravado pela população-alvo específica que é usada na seleção de polimorfismos e pelo rigor do experimentador em declarar marcadores polimórficos. Visando minimizar este viés, neste trabalho foram construídas 18 bibliotecas genômicas e submetidos à triagem mais de 20 mil clones de DNA gerados por redução de complexidade genômica que enriquece a fração para genes e sequências de alta complexidade.

O microarranjo DArT desenvolvido neste estudo proporciona pelo menos duas ordens de magnitude mais marcadores polimórficos do que um único ensaio RAPD ou AFLP. Em *Eucalyptus*, enquanto um primer RAPD selecionado pode fornecer até 10 bandas polimórficas robustas em uma corrida de gel (Grattapaglia e Sederoff, 1994) e uma combinação única de AFLP pode fornecer em média 30 a 40 marcadores polimórficos (Gaiotto, Bramucci *et al.*, 1997), um único ensaio DArT fornece entre 1.000 a 4.000 marcadores a partir de mais de 7.000 sondas polimórficas presentes no arranjo. Além disso, o conjunto de sondas DArT imobilizadas no microarranjo permite comparações de genotipagem entre uma ampla variedade de

espécies e populações, enquanto que genótipos de marcadores AFLP ou RAPD são muito menos passíveis de comparação e integração entre laboratórios e menos ainda entre espécies diferentes.

Neste trabalho, o elevado nível de multiplexagem dos marcadores DArT foi validado em uma extensa e representativa coleção de espécies e indivíduos de *Eucalyptus*. Os resultados indicam que o arranjo de genotipagem DArT fornece milhares de marcadores polimórficos para estudos de diversidade e representa uma plataforma muito eficaz para a geração de mapas de ligação de alta densidade com uma proporção substancial de marcadores compartilhados entre pedigrees. Em estudos de mapeamento de ligação, uma aplicação com um dos menores níveis de polimorfismo, pois a diversidade deriva de apenas dois genitores, os resultados da análise de um conjunto de pedigrees representativo indica que pode-se esperar entre 1500 e 2500 marcadores polimórficos e provavelmente grande parte mapeáveis (Tabela 11) a um custo de cerca de cinco centavos de dólar por marcador polimórfico. Este custo por amostra é pelo menos 50% menor do que o custo das plataformas atuais de genotipagem para um número equivalente de marcadores (ex. Illumina Golden Gate). Do ponto de vista de equipamento, a implementação da tecnologia DArT, da mesma forma como outras plataformas de genotipagem, envolveria, entretanto, um investimento substancial para a compra de equipamentos necessários para produzir os arranjos de alta densidade, câmaras de hibridização e um scanner de múltiplas fluorescências.

Outra vantagem importante dos marcadores DArT é a sua transferibilidade entre espécies, aspecto que é particularmente crítico quando se trata de um gênero como o *Eucalyptus* que conta com mais de 700 espécies passíveis de serem estudadas. Diversas espécies são chave em seus ecossistemas florestais e de grande interesse do ponto de vista de genética de populações, enquanto que outras são amplamente plantadas para o fornecimento de madeira para atividades industriais. Ao contrário de RAPD ou mesmo microssatélites, a transferibilidade de marcadores permitirá a comparação detalhada de parâmetros genéticos populacionais, mapas de ligação e posições de QTL entre diferentes estudos. Vale destacar, entretanto, que essa transferibilidade tem limites como o observado neste trabalho ao se verificar o baix desempenho do microarranjo ao se analisar uma espécie do gênero *Corymbia*. As consequências filogenéticas desta observação e do desempenho do arranjo DArT para uma ampla gama de espécies de *Eucalyptus* foram detalhadas em um estudo publicado recentemente que se baseou em dados gerados neste trabalho (Steane, Nicolle *et al.*, 2011).

O microarranjo DArT provou ser útil para a diferenciação genética seja dentro como entre espécies em *Eucalyptus globulus, Eucalyptus grandis, Eucalyptus nitens, Eucalyptus pilularis* e *Eucalyptus urophylla*. Ou seja, além de fornecer perfis genéticos individuais altamente discriminativos, diversos marcadores no microarranjo mostraram diferenças importantes de frequência entre espécies de forma a permitir fácil identificação da espécie à qual um determinado indivíduo pertence. O microarranjo DArT permitiu a identificação de híbridos de *E. nitens x E. globulus* e resolveu com sucesso situações sutís de estruturação geográfica entre procedências dentro *E. camaldulensis, E. cladocalyx, E. globulus* e *E. urophylla* que haviam sido descritas anteriormente com base em microssatélites. A análise filogenética envolvendo 94 espécies de Eucalyptus apresentou resultados que foram em grande parte congruentes com a taxonomia tradicional e filogenias baseadas em ITS (Inter transcribed spacer), mas forneceu maior resolução dentro dos clados principais do que análises anteriores. Análises filogenéticas realizadas com diferentes subconjuntos de marcadores derivados de diferentes bibliotecas genômicas construidas com DNA de diferentes espécies não apresentaram diferenças importantes, revelando que o microarranjo DArT não parece introduzir um viés derivado da amostragem das sondas nele contidas.

Uma limitação da tecnologia DArT em comparação com microssatélites multi-alélicos é o seu modo de herança dominante, o que impede o estudo de aspectos de variação intra-indivíduo. Marcadores dominantes são também menos informativos para a construção de mapas de ligação, a menos que um grande número deles seja utilizado quando então marcadores dominantes ligados em repulsão passam a ser quase tão informativos quanto marcadores co-dominantes (Plomion, Liu *et al.*, 1996). A ocorrência agrupada (*clustering*) de marcadores DArT no genoma poderia ser um problema derivado do método de geração das representações genômicas reduzidas com enzimas de restrição. No entanto, este problema não é exclusivo da tecnologia DArT. Uma avaliação rigorosa deste aspecto é possível pelo sequenciamento das sondas que hoje constituem o microarranjo e mapeamento físico das mesmas sobre o genoma de referência. A caracterização do conteúdo genômico deste microarranjo foi recentemente realizada iniciando com o seqüenciamento de boa parte dos 7.680 clones de DNA revelando uma distribuição relativamente homogênea dos marcadores cobrindo mais de 75% do genoma quando este foi dividido em intervalos de 500 kb (Petroli, Sansaloni *et al.*, 2011).

A disponibilidade das sequências dos marcadores DArT vai facilitar a integração de mapas de alta densidade e posições de QTL sobre o genoma de *Eucalyptus*. O microarranjo

DArT constitui assim uma ferramenta poderosa para realizar análises genéticas de alta resolução necessárias para aplicações tais como mapeamento genético fino, seleção genômica ampla e investigações filogenéticas e evolutivas. Além disso, a flexibilidade e a capacidade de expansão da tecnologia DArT abre a possibilidade de enriquecer ainda mais o microarranjo atual com marcadores adicionais, simplesmente realizando a triagem de mais clones e fazendo as análises de redundância necessárias para otimizar a representatividade das novas sondas incorporadas ao microarranjo. Os recentes trabalhos publicados utilizando este microarranjo DArT para a construção de mapas genéticos em diferentes espécies de *Eucalyptus* (Hudson, Kullan *et al.*, 2011; Kullan, Van Dyk *et al.*, 2011; Petroli, Sansaloni *et al.*, 2011), bem como estudos filogenéticos (Steane, Nicolle *et al.*, 2011) e a recente aplicação desta ferramenta de genotipagem para a execução de uma prova de conceito da abordagem de seleção genômica (Resende, Resende *et al.*, 2012), corroboram a importância e o excelente desempenho desta tecnologia de genotipagem.

4. CAPÍTULO 2

AVALIAÇÃO DA METODOLOGIA DE GENOTIPAGEM POR SEQUENCIAMENTO *DART-SEQ* PARA MAPEAMENTO GENÉTICO EM *Eucalyptus*

4.1. INTRODUÇÃO

Métodos de genotipagem de alto desempenho e com ampla cobertura genômica têm se tornado cada vez mais importantes para atender a resolução e velocidade necessárias em uma variedade de aplicações em genômica e melhoramento molecular de espécies florestais. A metodologia DArT (*Diversity Arrays Technology*) (Jaccoud, Peng *et al.*, 2001) tem recebido amplo interesse ao satisfazer eficientemente os requerimentos de desempenho, cobertura genômica e transferibilidade interespecífica de marcadores para mais de 60 espécies de plantas até o momento. Isto inclui espécies de *Eucalyptus*, que apresentaram o mais alto nível de polimorfismo de marcadores DArT até hoje (Sansaloni, Petroli *et al.*, 2010), juntamente com *Pinus taeda*, que também revelou elevada diversidade (Alves-Freitas, Kilian *et al.*, 2010) possivelmente em função da domesticação recente e hábito alógamo em comparação com a maioria das culturas anuais analisadas. A partir de 2.008 a empresa DArT Pty iniciou os trabalhos visando a migração da tecnologia de genotipagem DArT baseada em hibridização de sondas para a utilização de genotipagem por sequenciamento utilizando métodos de sequenciamento de próxima geração ou NGS (*Next Generation Sequencing*) (A. KIlian com. pessoal).

Uma série de tecnologias NGS tais como Roche/454, Illumina e AB SOLiD tem sido desenvolvidas nos últimos anos, todas elas capazes de gerar centenas de milhares ou dezenas de milhões de sequências de tamanho variável que vai hoje de 35 a quase 1000 pares de base a custos relativamente baixos (Mardis, 2008). Metodologias de NGS têm sido utilizadas de forma crescente para auxiliar projetos de sequenciamento de genoma completos e principalmente em projetos de re-sequenciamento de genomas de organismos para os quais existem genomas de referência (Metzker, 2010). Metodologias de NGS têm tido papel decisivo nos últimos anos para acelerar a descoberta de grande número de SNPs (*Single Nucleotide Polimorphisms*) para diversas espécies (Barbazuk, Emrich *et al.*, 2007; Brockman, Alvarez *et al.*, 2008; Hillier, Marth *et al.*, 2008; Van Tassell, Smith *et al.*, 2008; Wiedmann, Smith *et al.*, 2008) incluindo espécies de *Eucalyptus* (Novaes, Drost *et al.*, 2008; Külheim, Yeoh *et al.*, 2009; Grattapaglia, Silva-Junior *et al.*, 2011).

A abordagem utilizada para a descoberta de SNPs em geral tem se baseado no sequenciamento de representações genômicas reduzidas visando assim incrementar a cobertura de sequências em pontos específicos do genoma. Inicialmente a redução de complexidade genômica tem sido produzida pela utilização de RNA e síntese de cDNA de forma que SNPs foram descobertos essencialmente em regiões transcritas do genoma (Novaes, Drost *et al.*, 2008). Nos últimos anos, entretanto, metodologias de redução de complexidade genômica utilizando principalmente corte com enzimas de restrição tem sido utilizadas de forma crescente, com os trabalhos pioneiros em bovinos (Van Tassell, Smith *et al.*, 2008) e metodologias especificamente desenvolvidas para isso como aquela denominada RAD (*Restriction Associated DNA*) *sequencing* (Baird, Etter *et al.*, 2008) que tem sido utilizada de forma crescente para a descoberta de amplas baterias de SNPs em diferentes organismos (Maughan, Yourstone *et al.*, 2009; Hyten, Cannon *et al.*, 2010; Barchi, Lanteri *et al.*, 2011; Hohenlohe, Amish *et al.*, 2011; O'rourke, Yochem *et al.*, 2011; Rowe, Renaut *et al.*, 2011; Willing, Hoffmann *et al.*, 2011; Scaglione, Acquadro *et al.*, 2012).

A possibilidade de combinar metodologias otimizadas de redução de complexidade genômica com técnicas cada vez mais potentes de NGS e o uso de indexação de amostras com adaptadores específicos levou, por sua vez, ao uso desta abordagem visando diretamente a genotipagem das amostras e não apenas para a descoberta de SNPs. Isto foi inicialmente demonstrado utilizando RAD *sequencing* (Miller, Dunham *et al.*, 2007; Emerson, Merz *et al.*, 2010) e em seguida pela abordagem conhecida atualmente como genotipagem por sequenciamento ou GbS (*Genotyping-by-Sequencing*) (Maughan, Yourstone *et al.*, 2010; Myles, Chia *et al.*, 2010; Elshire, Glaubitz *et al.*, 2011; Poland, Brown *et al.*, 2012). Emboras estas duas metodologias sejam bastante parecidas no conceito central, a preparação de bibliotecas na abordagem GBS é muito mais simples do que na técnica RAD. GBS requer menor quantidade de DNA, evita corte aleatório e seleção do tamanho dos fragmentos, e é realizada em apenas dois passos, digestão/ligação e amplificação por PCR das representações genômicas. GbS tem sido portanto proposta como uma ferramenta simples, robusta, específica e altamente reproduzível (Elshire, Glaubitz *et al.*, 2011) que promete revolucionar a capacidade de genotipar milhares de amostras para milhares de marcadores a custos de poucas dezenas de dólares US, aumentando radicalmente a resolução de estudos populacionais, caracterização de germoplasma, melhoramento genético molecular de plantas e mapeamento de alta densidade.

A metodologia GbS se baseia resumidamente na redução de complexidade genômica da amostra de DNA total utilizando corte do DNA total com diferentes combinações de enzimas de restrição (ER) específicas otimizadas para cada espécie. Após a redução de complexidade via ER e antes do sequenciamento, cada amostra de DNA a ser genotipada recebe adaptadores com sequências indexadoras (*barcodes*) que permitem mais tarde rastrear as sequências geradas para cada amostra (Craig, Pearson *et al.*, 2008). Desta forma, 96 amostras podem ser sequenciadas conjuntamente em cada canaleta de sequenciamento, otimizando significativamente os custos. Para cada amostras são gerados alguns milhões de sequências curtas (40 a 100 bases) denominadas também de etiquetas (*tags*) para algumas dezenas de milhares de locos, de forma que cada loco é representado por uma dezena ou mais de sequências, permitindo assim a declaração de genótipos aos SNPs contidos nestas sequências curtas. Polimorfismos entre indivíduos podem resultar seja da presença ou ausência de sequências entre amostras derivadas da variabilidade na distribuição dos sítios de restrição ou de SNPs (polimorfismos de base individual) entre sequências em comum entre as amostras. Em cada corrida de sequenciamento em plataforma Illumina, por exemplo, podem ser genotipadas 768 amostras e gerados milhares ou dezenas de milhares de marcadores a depender da diversidade nucleotídica da espécie. Uma revisão recente dos métodos de redução de complexidade genômica acoplados a seqüenciamento NGS detalha as vantagens e limitações de cada abordagem (Davey, Hohenlohe *et al.*, 2011). A metodologia DArT que também se baseia na redução da complexidade genômica das amostras a serem genotipadas vem sendo adaptada de forma semelhante para a plataforma de NGS. Em outras palavras, trata-se de migrar de um procedimento no qual a detecção de polimorfismos é realizado via hibridização de sondas para um outro no qual a detecção é feita via sequenciamento. A DArT Pty Ltd, essencialmente adaptou o seu método de preparo de bibliotecas visando concretizar esta migração. O objetivo desta etapa do projeto de tese foi testar e avaliar este novo método, denominado DArT-Seq para *Eucalyptus*, com base no sucesso já alcançado pela DArT para espécies de genomas mais complexos como trigo, sorgo e cevada. A avaliação foi realizada comparando-se DArT-Seq com a metodologia DArT tradicional visando a construção de um mapa genético com uma progênie derivada do cruzamento da árvore *E. grandis* BRASUZ1 (árvore cujo genoma é hoje o genoma referência de *Eucalyptus)*.

4.2. MATERIAL E MÉTODOS

4.2.1. Material vegetal

As amostras utilizadas para avaliar a nova tecnologia de genotipagem por sequenciamento baseada em DArT foi uma população segregante de 89 indivíduos derivados de um cruzamento intraespecífico entre os indivíduos BRASUZ1 x M43D1, ambos *Eucalyptus grandis*, os quais foram fornecidos pela empresa Suzano. A árvore BRASUZ1 foi a utilizada para a obtenção do genoma de referência de *Eucalyptus* (http://www.phytozome.net/*Eucalyptus*.php).

Para a extração do DNA total foi utilizado tecido foliar de cada árvore. A extração foi feita pelo método utilizando o detergente CTAB 2% (Doyle e Doyle, 1987). Para a maceração das folhas foi utilizado o equipamento TissueLyser da Qiagen com esferas metálicas. Cada amostra foi macerada contendo o tampão de CTAB com Mercaptoetanol por 40 segundos. O precipitado foi ressuspendido em uma solução de Tris/EDTA (TE) pH 8,0 com Ribonuclease A (RNaseA) e incubado à temperatura de 37ºC por 20 minutos para ação da enzima. O DNA foi então quantificado em géis de agarose 1% corados com brometo de etídeo por comparação com DNA do fago lambda de concentração conhecida.

4.2.2. Análise de locos microssatélites e verificação de parentesco

Para certificar o correto parentesco dos 89 indivíduos da população segregante foram analisados cinco locos microssatélites, sendo estes EMBRA 12, EMBRA 38, EMBRA 28, EMBRA 210 e EMBRA 681. Posteriormente, o painel de microssatélites genotipados foi ampliado visando a inclusão destes marcadores no mapa de ligação totalizando 29 locos microssatélites. As amplificações dos locos microssatélites foram realizadas via PCR utilizando o kit Multiplex PCR da QIAGEN®. O volume final de cada reação da PCR foi de 5µl, contendo 1X do QIAGEN Multiplex PCR Master Mix 2X; 0,3µM de cada primer; 0,5X de Q-solution; 2ηg de DNA; e água RNase-free para completar o volume final. O programa de PCR utilizado seguia os padrões recomendados pelo fabricante do kit de amplificação, começando com uma desnaturação inicial a 95°C por 15 minutos, para ativar a enzima, seguidos de 35 ciclos envolvendo em cada um uma desnaturação a 95°C por 30 segundos, anelamento dos iniciadores a 57°C por 90 segundos, e extensão a 72°C por 1 minuto; após os 35 ciclos foi realizada uma extensão final a 72°C por 30 minutos. Posteriormente, o sucesso de amplificação foi detectado em eletroforese capilar em sequenciador automático ABI3100 (Applied Biosystems). De cada amostra foram reunidos 1 µl da PCR, 1 µl do padrão de fragmentos de tamanho conhecido marcado com

fluorescência ROX (Brondani e Grattapaglia, 2001), e 8 µl de formamida Hi-Di (Applied Biosystems).

4.2.3. Genotipagem com o microarranjo operacional DArT

Amostras de DNA dos parentais (BRASUZ1 e M43D1) e os 89 indivíduos segregantes foram processadas com o microarranjo de genotipagem operacional DArT descrito anteriormente (Sansaloni, Petroli *et al.*, 2010) com a finalidade de gerar dados de marcadores para análise comparativa com os dados de DArT baseados em NGS. O método de redução da complexidade escolhido foi *PstI/TaqI*, sendo o mesmo utilizado para o desenvolvimento do arranjo DArT.

4.2.4. Genotipagem por sequenciamento

As principais etapas da metodologia GBS são apresentadas na Figura 16. Nesta imagem é possível observar a redução da complexidade genômica e ligação de adaptadores como primeira medida, seguida da amplificação de *targets*, agrupamento ou *pooling* das amostras, amplificação de clusters, sequenciamento dos fragmentos e finalmente a análise de dados.

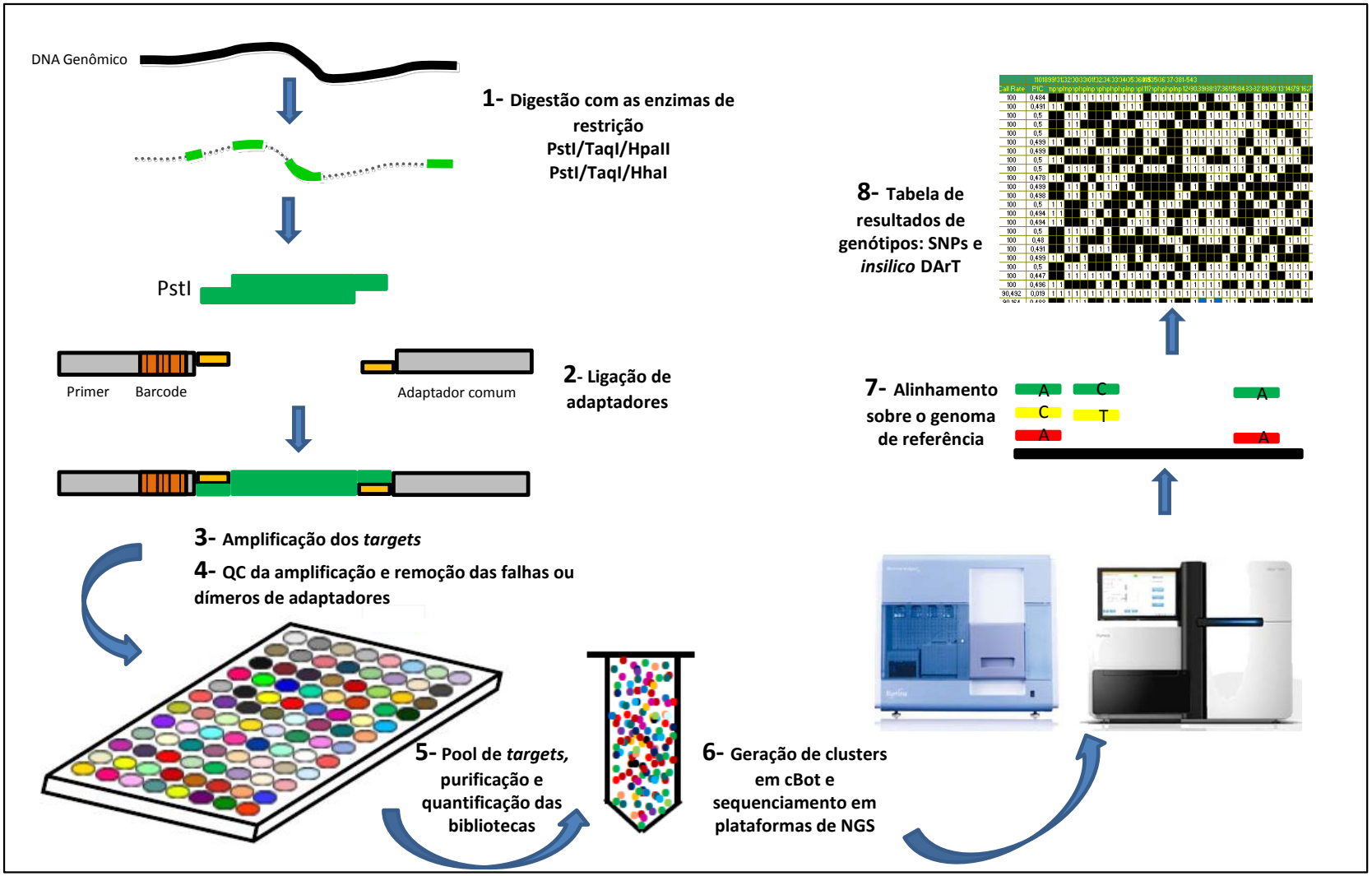


Figura 16. Fluxograma do procedimento DArT-Seq baseado em NGS (Next Generation Sequencing).

4.2.4.1. Redução da complexidade genômica e ligação de adaptadores

A primeira etapa na metodologia GbS é a redução da complexidade genômica. Esta foi realizada mediante digestão com a enzima de restrição *PstI* (CTGCAG) em combinação com duas enzimas adicionais. Para este experimento com *Eucalyptus* foram utilizados dois métodos de redução, em cada um dois quais se empregaram três enzimas diferentes *PstI/TaqI/ad-HpaII* e *PstI/TaqI/ad-HhaI*. A digestão com enzimas e a ligação de adaptadores foi realizada simultaneamente na mesma reação. Em contraste com a metodologia DArT baseada em hibridização (Sansaloni et al., 2010), nesta técnica foram incluídos dois adaptadores dentro da reação conforme detalhado a seguir.

- **Adaptador *forward* com sequência indexadora ou *barcode:*** este adaptador termina com uma sequência indexadora ou *barcode* de 4 a 8 pb, seguida de 4 pb em forma de cadeia simples expostas no extremo 3', as quais são complementares ao extremo gerado pelo corte com a enzima *PstI* (Figura 17). Estas sequências indexadoras permitiram mais tarde rastrear os fragmentos gerados para cada amostra. Um set de 96 adaptadores-*barcodes* chamados de "v1.3" tem sido preparados para este propósito e foram dispensados em formato de placa de 96 poços. A totalidade de adaptadores-*barcodes* foi rearranjada em diferentes posições da placa para criar uma segunda versão de adaptadores, e assim, equilibrar as pequenas variações que possam existir na contagem final de *tags* associados a cada adaptador. Esta nova versão foi chamada de "v1.3 balanced" e é utilizada para ser acoplada à réplica de cada amostra. Desta forma, foi possível analisar conjuntamente as 91 amostras (2 parentais + 89 descendentes F1) em uma canaleta de sequenciamento da *flow cell* de uma corrida Illumina GAIIx, otimizando significativamente os custos de sequenciamento.

a) PstI_BC114_Forward

5'- **ACACTCTTTCCCTACACGACGCTCTTCCGATCTtcaactgaTGCA** -3'

b) PstI_BC114_Reverse

5'- **tcagttgaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT** -3'

c) PstI_BC114

5' – **AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT** - 3'
…… 5' – **ACACTCTTTCCCTACACGACGCTCTTCCGATCTtcaactgaTGCA** - 3'
……3' – **TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAagttgact** - 5'

Figura 17. Sequências indexadoras ou barcodes. (a) sequência *forward* como adaptador (vermelho), o *barcode* de 4 a 8 pb (azul) e o sitio *PstI* (verde). (b) sequência *reverse* incluindo o adaptador (vermelho) e o *barcode* (azul). (c) sequência adaptador-*barcode* depois do anelamento, mostrando o iniciador de PCR no alinhamento (negro).

- **Adaptador comum:** este adaptador foi desenhado com a sequência de reconhecimento da enzima *HpaII* ou *HhaI* num extremo, dependendo do experimento (Figura 18). Este adaptador é o mesmo para todas as amostras e deve ser incluído na mistura de digestão/ligação.

a) EB4bp_CommonTop

5'- **CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT**-3'

b) EB_HpaII_CommonBot

5'- **CGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG**-3'

c) EB_HpaII

5'–**CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGG**–3'
.....–5'- **CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT**
..... –3'- **GAGCCGTAAGGACGACTTGGCGAGAAGGCTAGAGC**

Figura 18. Sequências do adaptador comum. (a) sequência *forward* do adaptador (vermelho); (b) sequência *reverse*do adaptador (vermelho) e duas bases complementares ao sitio de reconhecimento da enzima *HpaII* (azul). (c) sequência do adaptador comum depois do anelamento, mostrando o iniciador de PCR no alinhamento (preto).

Os oligonucleotídeos contendo a fita superior e inferior dos adaptadores *barcodes* e dos adaptadores comuns foram diluídos separadamente a 50µM e anelados em um termociclador (95°C por 2min; suave diminuição a 25°C a uma velocidade de 0.1°C/seg.; 25°C por 30min e manter a 4°C). As amostras (DNA + adaptadores) foram digeridas em uma reação de 16 µl contendo 2µl de Tampão RE (10x), 0,1µl de *PstI* (20u/µl), 0,1µl de *TaqI* (20u/µl), 0,2µl *HpaII* (10u/µl), 0,1µl T4 DNA ligase, 0,2µl de ATP (50mM), 0,2µl de BSA (100x),2µl de adaptador comum *HpaII* (0.1 µM) e 11µl de H_2O. Os adaptadores-*barcodes* foram adicionados separadamente na mesma reação de digestão/ligação (2 µl de adaptador-*barcode* (0,1 µM). As amostras foram incubadas por 2 horas a 37°C, seguido de outras 2 horas a 60°C.

4.2.4.2. Amplificação dos *targets*

A amplificação dos fragmentos digeridos utiliza dois iniciadores com sequências complementares aos adaptadores ligados. A combinação de iniciadores usados para a PCR depende do método de redução da complexidade utilizado. O primeiro iniciador reconhece o extremo do adaptador-*barcode* (EB_PCRI primer: 5'–AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT–3') e o segundo iniciador reconhece o adaptador comum (EB_HpaII_pcr: (CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGG) ou

(EB_HhaI_pcr: CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGC). A reação total de PCR de 50 µl continha 2µl da amostra digerida, 5µl de tampão *Red Taq* (10x), 1µl de dNTPs (10nM), 0,1µl do iniciador EBPCRI (100µM), 0,1µl do iniciador EB_*HpaII*_PCR (100µM) ou EB_*HhaI*_pcr (dependendo do experimento), 2µl de polymerase *Red Taq* e 39,8 de H_2O. Esta reação foi amplificada a 94°C por 1 min., 94°C por 20 seg., diminuição de 2,4°C/seg. até 58°C, 58°C por 30seg., aumento de 2,4°C/seg. até 72°C, 72°C por 45 seg., repetição de 29 ciclos mais a partir do segundo passo, finalmente 72°C por 7 min. e manter a 10°C.

Através de uma pequena alíquota (5µl), os alvos amplificados foram verificados mediante eletroforese em um gel de agarose 1,2% para confirmar o rastro homogêneo dos fragmentos e para visualizar a distribuição dos tamanhos. A presença de bandas de tamanho pequeno significa a amplificação de dímeros de adaptadores, que devem ser removidos antes de proceder com a etapa seguinte.

4.2.4.3. Agrupamento, purificação e quantificação dos *targets*

Todos os *targets* amplificados com sucesso foram agrupados em um tubo de 1,5ml, enquanto que as falhas foram removidas antes do agrupamento (*pooling*). Posteriormente, a mistura de *targets* foi purificada utilizando o kit comercial "QIAquick PCR Purification" (Qiagen, Valencia, CA) seguindo as instruções sugeridas pelo fabricante. Uma vez purificados os *targets*, estes foram quantificados em gel de agarose 1,2% com o objetivo de confirmar a ausência de dímero de adaptadores. Finalmente, os mesmos *targets* foram quantificados em PicoGreen para calcular o volume requerido para a corrida de sequenciamento.

4.2.4.4. Geração de clusters em sistema cBot™ (Illumina) e sequenciamento no Genome Analizer GAII Illumina

O Sistema cBot™ (Illumina) proporciona automação completa de um processo complexo como é a geração de clusters. Com muito pouca mão de obra e sem preparação de reagentes, o cBot utiliza amplificação ponte para criar simultaneamente centenas de milhões de moldes de moléculas de DNA individuais em quatro horas. O cBot™ dispensa reagentes a partir de uma placa de 96 poços pre-aliquotada, controlando também o tempo de reação e temperatura. A corrida é configurada usando a interface do *software* cBot™, o que simplifica a operação e fornece um relatório visual do estado de execução. Um leitor de código de barras no instrumento registra os reagentes e *flow cell* utilizados em cada experimento. O cBot gera dentro da *flow cell* entre 700 a 800 clusters/mm^2. Para garantir esta quantidade de clusters, o pool de amostras purificadas deve estar corretamente quantificado para assim calcular de maneira precisa a diluição necessária para a corrida de sequenciamento.

Cada *flow cell* possui 8 canaletas de sequenciamento, das quais neste experimento foram utilizadas 4. A preparação das diluições do pool de amostras purificado foi realizado em um strip de 8 tubos. Em um segundo strip foi realizada a desnaturação dos fragmentos colocando 16µl de tampão Tris PH 7,5, 3µl da diluição das amostras (do primeiro tubo) e 1µl de NaOH (2M). Esta solução foi misturada, exposta a um spin e incubada em gelo por 5 minutos. Posteriormente, foram colocados 994µl de tampão Hyb (proporcionado pelo kit de reagentes do cBot™ (Illumina) em 8 tubos de 1,5ml e dentro de cada um deles foram acrescentados 6 µl de produto da desnaturação. Esta solução é misturada por inversão, aproximadamente umas 10 vezes. Na etapa final, são colocados 120µl da mistura anterior em um strip de 8 tubos prontos para serem colocados no equipamento cBot. Além da solução preparada anteriormente, foram colocados no equipamento cBot, a placa de reagentes e a *flowcell*.

Cada *flow cell* possui 8 canaletas de sequenciamento, cada uma delas constituída por duas colunas, e 60 campos ou "*tiles"* a partir das quais são emitidas quatro imagens por ciclo, uma imagem para cada base (Figura 19). Foram então utilizadas quatro canaletas de sequenciamento com as 91 amostras cada uma, duas com a biblioteca gerada a partir do método de redução da complexidade *PstI/TaqI/ad-HpaII* e as outras duas com o método *PstI/TaqI/ad-HhaI*.

Uma vez terminada a amplificação dos clusters, a *flowcell* foi colocada no "*Genome Analyzer II* ou *GAII*" (Illumina, Inc., San Diego, CA) para sequenciar 77 bases "*single read*", o qual levou um tempo de aproximadamente 77 horas (uma hora por base).

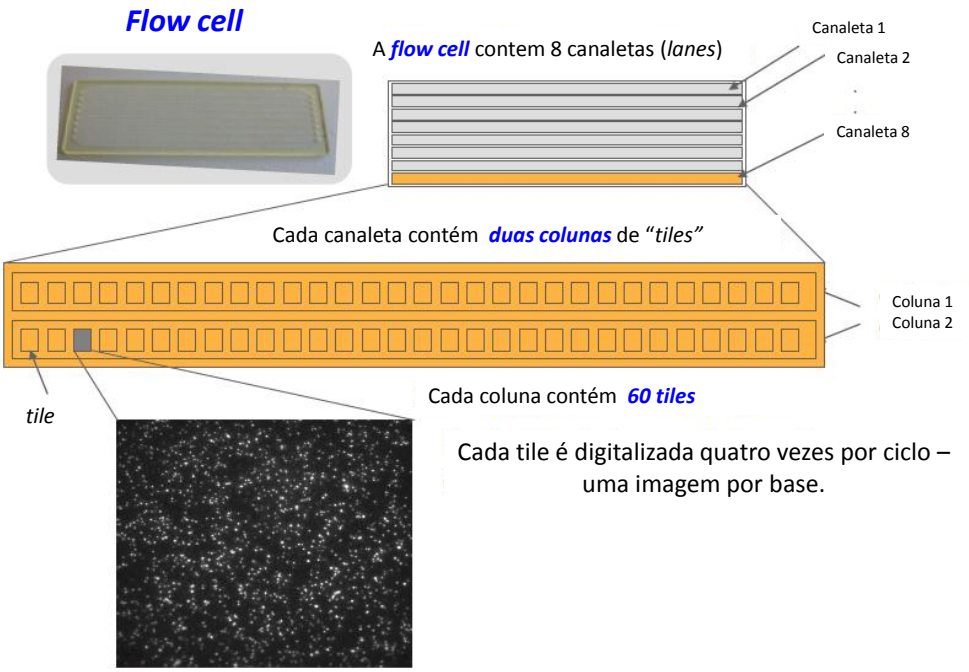


Figura 19. Representação esquemática da *flowcell* de uma corrida de sequenciamento de GAIIX.

4.2.4.5. Processamento dos dados brutos e alinhamento das sequências de DNA

Após a corrida de sequenciamento, os dados brutos foram exportados em arquivos QSEQ, posteriormente transformados em arquivos FASTQ para começar a filtragem. Estes arquivos foram produzidos com os dados das sequências com 77 pb. O primeiro filtro foi realizado utilizando um script de divisão de *barcodes* (DArT Lt.), considerando as primeiras 9 bases de cada *read*, incluindo as 4 a 8 bases do *barcode* e as primeiras bases do sitio *PstI*. Neste sentido, foram aceitos os *reads* com um mínimo de 75% da sequência superando um valor Q Phred igual a 30, o que implica uma probabilidade de 1:1000 da base ser chamada incorretamente (acurácia de 99,9%). Posteriormente, um segundo filtro foi aplicado sobre a sequência completa do *read,* com o objetivo de eliminar sequências de baixa qualidade, aceitando os *reads* com um mínimo de 50% da sequência superando o valor de qualidade Q Phred igual a 10, representando uma probabilidade de 1:10 de a base ser chamada

incorretamente (acurácia de 90%). As sequências que não superaram este segundo filtro, também foram eliminadas da análise.

Após produzir várias etapas estatísticas de controle de qualidade e limpeza de sequências, estas foram alinhadas sobre o genoma de referência publicado em 15/1/2011 no site Phytozome (http://www.phytozome.net/*Eucalyptus*.php) utilizando o programa BWA (Burrows-Wheeler Aligner) (Li e Durbin, 2009) aceitando um máximo de 4 *mismatches*. Os arquivos de saída do alinhamento foram processados utilizando um pipeline de análise desenvolvido pela empresa DArT Pty Ltd, o qual gerou dois arquivos diferentes a partir destes dados. No primeiro arquivo, foram produzidos clusters baseados em similaridade de sequências, aceitando um máximo de 3 bases divergentes por *read*. A partir destes clusters, o polimorfismo foi detectado pela presença ou ausência de sequências entre amostras, sendo registrado através de um sistema binário (1 ou 0) e estes chamados de *InSilico DArT*. Este polimorfismo é derivado da variabilidade na distribuição dos sítios de restrição entre sequências em comum entre amostras. O segundo arquivo foi produzido utilizando a variação de bases simples (SNPs) dentro das sequências. Para isto foi selecionada a variante alélica mais abundante dentro do cluster, a qual é considerada referência, e a variante menos abundante é considerada um SNP. Assim, estas variantes podem ser classificadas pela presença/ausência das mesmas dentro de cada sequência. Genótipos de ambos os tipos de marcadores foram exportados para planilhas Excel para uma melhor manipulação dos dados.

4.2.4.6. Seleção de marcadores para análise de mapeamento genético

Os marcadores polimórficos (*Insilico* DArT e SNPs) passaram por um processo de seleção para serem posteriormente mapeados geneticamente. Para escolher os melhores marcadores, uma análise de qualidade foi realizada com base em uma série de parâmetros rigorosos. Os marcadores *Insilico DArT* foram selecionados com base em três parâmetros principais: (1) *Row average* > 10, o que significa que a média de número per *reads* detectados deve ser maior a 10; (2) *GTZ (Grater than zero) Ratio* (0,3-0,7), onde serão escolhidos os marcadores com uma frequência de presença "1" entre 30% e 60%; e (3) Q>2,5, expressa a qualidade do marcador. As exigências de seleção dos marcadores SNPs foram determinadas também através de três parâmetros de qualidade: (1) Call rate> 0,8, o que significa que somente é aceito um máximo de 20% de dados perdidos para cada marcador; (2) *One Ratio Ref/SNP* (0,3-0,7), envolve marcadores com uma frequência de presença "1" entre 30% e 70%

por marcador; e finalmente (3) *Num of Heterozigous* < 0,7, inclui marcadores com 70% de genótipos heterozigotos.

4.2.4.7. Construção do mapa de ligação de alta densidade

A construção do mapa de ligação foi realizada com o programa desenvolvido na empresa DArT Pty. Ltd. (não publicado) usando scripts em R especialmente desenvolvidos para o agrupamento, ordenação e mapeamento dos marcadores. Este mapa envolveu marcadores DArT baseados em microarranjo e marcadores *Insilico DArTs*. O programa compreende basicamente três passos principais. O primeiro passo ou "*binning"* é a construção de grupos de marcadores baseada na similaridade de genótipos (scores), portanto, todos os marcadores que possuem uma distância menor que 0,01 são colocados no mesmo *bin*. Após esta etapa, o programa calcula uma matriz de distância (Hamming) e divide os marcadores em grupos de ligação. Se a distância entre os marcadores é menor do que 0,05, estes são colocados no mesmo grupo de ligação utilizando o algoritmo TSP (*travelling salesman problem*). Simultaneamente, o *software* calcula um "*fit value*" para determinar a melhor posição do marcador dentro de cada grupo de ligação (GL). Locos com valores baixos de "*fit value*" são removidos do mapa. No segundo passo ou "*joining groups*" é calculada a distância entre os extremos de todos os grupos, de maneira a poder unir grupos de ligação mediante seus extremos. Finalmente, o terceiro passo ou "*attach remaining markers*", adiciona os marcadores restantes de menor qualidade que não entraram no mapa *framework*.

4.3. RESULTADOS E DISCUSSÃO

4.3.1. Genotipagem DArT baseada em microarranjo

A genotipagem da população segregante BRASUZ1 utilizando a plataforma de microarranjo DArT (Sansaloni, Petroli *et al.*, 2010) detectou 1.088 marcadores polimórficos robustos e de alta qualidade, com valores de reprodutibilidade >97% e Call Rate >84%, representando 14,16% do total de sondas DArT presentes no microarranjo. Estimou-se que da totalidade de marcadores DArT polimórficos, 505 (46,4%) tiveram uma segregação tipo pseudo-cruzamento teste 1:1, enquanto que os restantes 583 (53,6%) segregaram em uma proporção 3:1 ou F2. Observando a primeira categoria de marcadores (1:1), foi possível determinar que 59% (298 marcadores) tinha origem no parental M43D e o restante 41% (207 marcadores) era derivada do parental BRASUZ1. A expectativa de obter um baixo número

relativo de marcadores polimórficos nesta população segregante foi confirmada, quando comparada com outras populações de mapeamento em *Eucalyptus* utilizando a mesma tecnologia (Kullan, Van Dyk *et al.*, 2011; Petroli, Sansaloni *et al.*, 2011). A razão pode estar no fato desta família ter sido produzida por um cruzamento intraespecífico (*E. grandis* x *E. grandis),* enquanto que em famílias produzidas a partir de cruzamentos interespecíficos foi observado um nível mínimo de polimorfismo de 23,67% (*E. dunni/E. grandis* x *E. urophylla*) e um máximo de 33,24% (*E. grandis* x *E. urophylla*) (Sansaloni, Petroli *et al.*, 2010). Uma explicação adicional é o baixo nível de heterozigosidade de sequência do parental autofecundado BRASUZ1 o que reduz o número de marcadores segregantes na sua progênie.

4.3.2. Genotipagem DArT-Seq baseado em NGS

Neste trabalho foram utilizados 91 indivíduos da família BRASUZ1, sendo dois parentais (BRASUZ1 e M43D) e 89 descendentes. Todas estas amostras foram submetidas à etapa de digestão/Ligação utilizando dois métodos de redução de complexidade genômica (*PstI_TaqI_ad_HpaII* e *PstI_TaqI_ad_HhaI*), com réplicas completas de cada amostra. Posteriormente, os *targets* das representações genômicas criadas foram amplificados em termociclador e testados em gel de agarose para confirmar o sucesso da amplificação e a ausência de dímeros de adaptadores (Figura 20). Nas quatro placas produzidas foi identificada a presença de dímeros de adaptadores em duas amostras, as quais foram removidas na etapa seguinte de *pooling* de amostras.

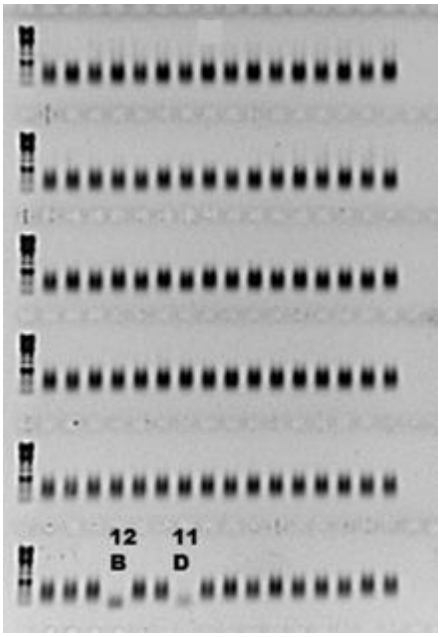


Figura 20. Controle de qualidade de *targets*. Gel de eletroforese 1,2% mostrando um rastro homogêneo de fragmentos amplificados. Nas posições 12B e 11D da placa observam-se bandas de pequeno tamanho representando dímeros de adaptadores. O marcador de peso molecular utilizado foi 1Kb DNA ladder.

Após a purificação do pool de *targets*, levou-se a cabo uma nova etapa de controle de qualidade. Desta vez, cada pool a ser introduzido em uma canaleta da *flow cell*, foi analisado em gel de agarose 1,2% com a finalidade de observar a presença de algum dímero de adaptadores que possa ter permanecido na amostra e comprovar a ausência total de bandas marcantes no rastro (Figura 21).

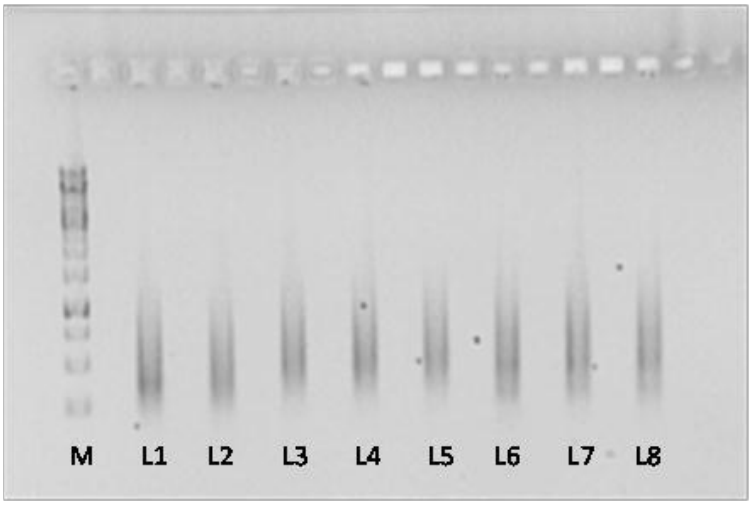


Figura 21. Controle de qualidade das representações genômicas após a purificação. Gel de eletroforese 1,2% após a purificação. Os dois métodos utilizados para *Eucalyptus* encontram-se nas colunas L3 - L7 (*PstI_TaqI_HpaII*) - L4 e L8 (*PstI_TaqI_HhaI*). M é o marcador de peso molecular 100kb ladder. A ausência de dímeros de adaptadores indica uma condição ótima para dar inicio à incorporação do pool de amostras na *flow cell*.

A corrida de sequenciamento no sequenciador Illumina GAIIx em quatro canaletas de uma única *flow cell* gerou 166,5 milhões de *reads* com um tamanho de 77 bases para cada sequência. Foi calculada uma média de 41,6 milhões de *reads* por canaleta, sendo que 86,7 milhões eram provenientes do método *PstI_TaqI_HpaII* e 79,8 milhões foram originados pelo método *PstI_TaqI_HhaI* (Tabela 13). Após a primeira filtragem de qualidade baseada na região do *barcode*, das 166 milhões de sequências totais restaram 133 milhões (80%) e um remanescente de 128 milhões *reads* (77,10%) foi aceito uma vez aplicado o segundo filtro de qualidade. Portanto, um total de 22,9% das sequências originais foram eliminadas. Uma

percentagem similar de sequências de alta qualidade (70%), que superara os filtros recomendados por Illumina, foi descrito por em um estudo utilizando a metodologia GBS em milho (Elshire, Glaubitz *et al.*, 2011). Este trabalho também demonstrou que os valores recomendados por Illumina para os parâmetros de filtragem parecem subestimar a qualidade dos *reads*, concluindo que 83% do total de *reads* eram de ótima qualidade (Elshire, Glaubitz *et al.*, 2011). A partir do total de sequências que passaram com sucesso nos dois filtros aplicados na nossa análise, foram identificados 30,6 milhões de *reads* únicos, representando 23,9% do total de *reads* aceitos, com uma média de 6,1 milhões de *reads* únicos por canaleta da *flow cell*.

Tabela 13. Resumo dos filtros utilizados para a seleção de reads para análise. Para a região do *barcode* o parâmetro de filtro utilizado foi: 75% da sequência com um valor Q Phred ≥30. Para a qualidade da sequência completa: 50% da sequência com um valor Q Phred ≥10.

Canaleta	Filtro	Nº de reads	Reads eliminados	Reads aceitos
3	Região barcode	42665527	8516433 (20%)	34149094
3	Seq. completa	34149094	1506568 (4%)	32642526
7	Região barcode	44000185	9966708 (23%)	34033477
7	Seq. completa	34033477	1410996 (4%)	32622481
4	Região barcode	38774949	7266582 (19%)	31508367
4	Seq. completa	31508367	799621 (3%)	30708746
8	Região barcode	41031005	7652085 (19%)	33378920
8	Seq. completa	33378920	1087557 (3%)	32291363

Várias etapas de controle de qualidade de sequências foram realizadas sobre os *tags* únicos através do *pipeline* de análise desenvolvido pela empresa DArT Pty Ltd. Esta avaliação detectou um total de 38.057 e 21.823 *tags* referência para os métodos *PstI_TaqI_ad_HpaII* e *PstI_TaqI_ad_HhaI*, respectivamente. Posteriormente alinharam-se todos estes *tags* únicos sobre o genoma de *Eucalyptus.* Constatou-se que 32.172 (84,54%) das *tags* geradas pelo primeiro método alinharam-se corretamente, dentre as quais 29.891 (92,90%) o fizeram sobre os 11 *scaffolds* principais correspondendo aos 11 cromossomos de *Eucalyptus*, e um número muito menor foi localizado nos *scaffolds* adicionais não montados. Para as *tags* produzidas pelo segundo método de redução de complexidade testado, identificaram-se 17.024 (78%) *tags* alinhadas sobre o genoma de referência, das quais 15.929 (93,56%) posicionadas nos 11

scaffolds maiores (Tabela 14). Observou-se que 87,81% dos *tags* alinhados gerados pelo método *PstI_TaqI_ad_HpaII*, tiveram uma única posição no genoma de referência, enquanto que para o método *PstI_TaqI_ad_HhaI*, isto ocorreu para 88,98% dos *tags* (Tabela 15). Uma análise similar realizada em milho mostrou que 78% das *tags* alinharam perfeitamente em uma única posição no genoma de referência dessa espécie, enquanto que os 22% restantes alinharam-se em múltiples posições (Elshire, Glaubitz *et al.*, 2011). Isto demonstra que a metodologia GBS amostra regiões genômicas de baixa cópia aumentando a possibilidade de identificar sequências ricas em genes.

Uma vez finalizada a análise de qualidade dos *reads*, revelaram-se dois tipos de polimorfismos. Marcadores *insilico* DArT, baseados na presença/ausência de *tags* entre amostras, derivados da variabilidade na distribuição dos sítios de restrição, e marcadores SNPs detectados a partir de polimorfismo de base individual entre as sequências sobrepostas para um determinado *tag* em comum entre amostras. Com a finalidade de produzir um mapa genético de alta densidade e qualidade, uma análise rigorosa de qualidade destes marcadores foi realizada para cada um dos dois métodos de redução de complexidade.

Tabela 14. Distribuição do número de *tags* alinhados com ambos os métodos de redução de complexidade sobre o genoma de referência de *Eucalyptus*.

	Nº de *tags*	
Scaffold	***PstI_TaqI_ad_HpaII***	***PstI_TaqI_ad_HhaI***
scaffold_1	2.300	1.284
scaffold_2	3.252	1.744
scaffold_3	3.066	1.655
scaffold_4	2.074	1.169
scaffold_5	2.929	1.505
scaffold_6	3.132	1.663
scaffold_7	2.616	1.334
scaffold_8	3.607	1.902
scaffold_9	2.016	1.076
scaffold_10	2.370	1.241
scaffold_11	2.529	1.356
> scaffold_11	2.281	1.095

Não alinharam	5.885	4.799
Total	**38.057**	**21.823**

Tabela 15. Distribuição do número de posições onde um *tag* único alinha-se no genoma de referência de *Eucalyptus* com os dois métodos de redução de complexidade.

	Nº de *tags*	
Nº de posições de alinhamento	***PstI_TaqI_ad_HpaII***	***PstI_TaqI_ad_HhaI***
1	28.252	15.148
2	2.699	1.373
3	727	295
4	251	119
5	81	33
6	60	23
7	31	8
8	24	7
>9	47	18
Total de reads alinhados	***32.172***	***17.024***

4.3.3. Detecção de polimorfismos de marcadores

4.3.3.1. Método de redução da complexidade *PstI_Taq_ad_HpaII*

- *Insilico DArT*: Foi exportado um total de 17.700 marcadores *insilico* DArT, os quais foram submetidos a um processo seletivo através de diversos parâmetros de qualidade: *Call rate*>80; *One ratio* entre 0,3-0,6; Q>2,5, e *Row Avg* ~14. Após as filtragens realizadas, foram selecionados 4.870 marcadores polimórficos de alta qualidade, representando 27,51% do total de marcadores *insilico DArT* detectados. Destes, 2.751 segregaram para o parental M43D (56,49%) e 1.845 a partir do parental BRASUZ1 (37,88%) e 274 tiveram uma segregação 3:1 (5,62%).

- SNPs: Um total de 141.906 marcadores SNPs foi detectado na análise, sobre os quais também foram aplicados os seguintes parâmetros de qualidade: *Call rate*>80; *One ratio* entre

0,3-0,7 e *Num of heterozygous* <0,7. Um total de 5.902 marcadores foi selecionado, dos quais 2.951 pertenciam ao alelo referência (Ref) e 2.951 ao alelo variante (SNP).

4.3.3.2. Método de redução da complexidade *PstI_Taq_ad_HhaI*

- *Insilico DArT*: Após as filtragens de qualidade mencionadas no método anterior para marcadores *insilico DArT*, 3.112 sequências foram selecionados a partir de um total de 21.449. Destes, 1.869 segregavam a partir do parental M43D (60%) e 1.243 (40%) a partir do parental BRASUZ1.

- SNP: para este método foram identificados 83.158 marcadores SNPs, dos quais 4.374 (5,25%) foram selecionados após as filtragens dos parâmetros de qualidade acima mencionados para esta classe de marcadores, sendo 2.187 do alelo referência (Ref) e 2.187 do alelo variante (SNP).

4.3.4. Mapeamento genético

Para a construção do mapa de ligação da família M43DxBRASUZ1 foram utilizados 5.958 marcadores polimórficos, sendo 1.088 marcadores DArT detectados em microarranjo e 4.870 *insilico* DArT, estes últimos obtidos com o método de redução da complexidade *PstI_TaqI_ad_HpaII*, selecionado para a construção do mapa por apresentar maior número de marcadores polimórficos em relação ao outro método utilizado. Os marcadores SNPs ainda não entraram na análise de mapeamento. Utilizando somente os marcadores DArT convencionais e *insilico* DArT que segregavam 1:1 para ambos os parentais foi realizada uma análise de distribuição sobre o genoma de referência (Tabela 16). Nesta tabela pode se observar que no parental M43D 64,97% dos marcadores alinharam nos 11 *scaffolds* do genoma de referência, 4,86% alinharam no genoma, mas fora dos 11 scaffolds e 30,16% não alinharam, enquanto que no parental BRASUZ1 80,35% alinharam sobre os 11 scaffolds, 12,35% alinharam no genoma, mas fora dos 11 scaffolds e 7,30% não alinharam. Vale ressaltar que esta pequena proporção de marcadores não alinhados é esperada pois os marcadores não foram selecionados quanto ao alinhamento ou não no genoma para entrarem na análise de mapeamento. Além disso, sabe-se que o genoma de referência não necessariamente é completo e ainda podem existir sequências no genoma dos parentais que são ausentes no genoma de referência. No que se refere aos marcadores DArT via microarranjo não alinhados,

vale lembrar que somente 6.918 dos 7.680 clones imobilizados no microarranjo foram sequenciados (Petroli, Sansaloni *et al.*, 2011).

Tabela 16. Distribuição no genoma de referência de *Eucalyptus* dos marcadores polimórficos DArTs e *insilico* DArTs de ambos os parentais selecionados para mapeamento.

	M43D			**BRASUZ1**		
Scaffolds	**Nº de Insilico DArT via DArT-Seq**	**Nº de DArTs via microarranjo**	**Total**	**Nº de Insilico DArT via DArT-Seq**	**Nº de DArTs via microarranjo**	**Total**
1	141	40	181	49	10	59
2	212	47	259	179	36	215
3	217	45	262	267	30	297
4	107	21	128	89	14	103
5	233	37	270	197	32	229
6	159	20	179	177	39	216
7	165	28	193	117	18	135
8	211	34	245	136	25	161
9	76	20	96	92	17	109
10	86	41	127	52	9	61
11	106	24	130	111	22	133
> 11	143	12	155	251	13	264
não alinharam	895	66	961	128	28	156
Total	***2751***	***435***	***3186***	***1845***	***293***	***2138***

Para a construção do mapa de ligação foi utilizado o programa desenvolvido pela DArT Pty. Ltd. o qual posicionou 5.281 marcadores polimórficos sobre os 11 cromossomos de *Eucalyptus* no mapa *framework*, com uma média de 480,1 marcadores por grupo de ligação. Esta quantidade de marcadores representou 92,9% do total de marcadores com potencial de serem mapeados (Figura 23). Da totalidade de marcadores mapeados, 4.683 (88,67%) eram *Insilico* DArTs e 598 (11,32%) DArT baseados em microarranjo (Tabela 17). O grupo de ligação que apresentou maior densidade de marcadores foi o GL3 com 851 marcadores, sendo 759 *insilico* DArTs e 92 DArTs, enquanto que o grupo de menor densidade foi o GL10 com 240 marcadores, 200 *insilico* e 40 DArTs.

Tabela 17. Estatística do mapa de ligação da família BRASUZ1 construido com o software da DArT com base em distância Hamming.

Grupo de Ligação	Nº total de marcadores mapeados	Nº de *insilico* DArT via DArT-Seq	Nº de DArTs via microarranjo	Comprimento do GL (Distância *Hamming)*
1	262	222	40	2,17
2	629	552	77	5,40
3	851	759	92	7,96
4	274	240	34	3,33
5	736	663	73	5,76
6	529	465	64	6,41
7	524	467	57	4,48
8	565	511	54	6,15
9	334	296	38	3,25
10	240	200	40	2,92
11	337	308	29	4,51
Total	***5.281***	***4.683***	***598***	***52,38***
Media	***480,09***	***425,72***	***54,36***	***4,76***

Utilizando todos os 5.281 marcadores mapeados, foi identificada a distribuição dos mesmos segundo a segregação para cada um dos parentais nos 11 grupos de ligação (Figura 22). Neste gráfico pode se observar uma distribuição homogênea de marcadores entre ambos os parentais com exceção do grupo de ligação 1. A ausência de marcadores provenientes do parental BRASUZ1 no grupo de ligação 1 era previsível, em função do menor número de marcadores segregantes observada dentre os selecionados para construir o mapa (Tabela 16). Este número menor provavelmente deriva do fato deste cromossomo ter entrado em homozigose em maior proporção do que os demais cromossomos no indivíduo BRASUZ1 sem causar depressão por endocruzamento que afetasse a sobrevivência da árvore. A maior homozigose de BRASUZ1 neste cromossomo resultou, evidentemente, em um menor número de marcadores segregando e consequentemente mapeados na progênie.

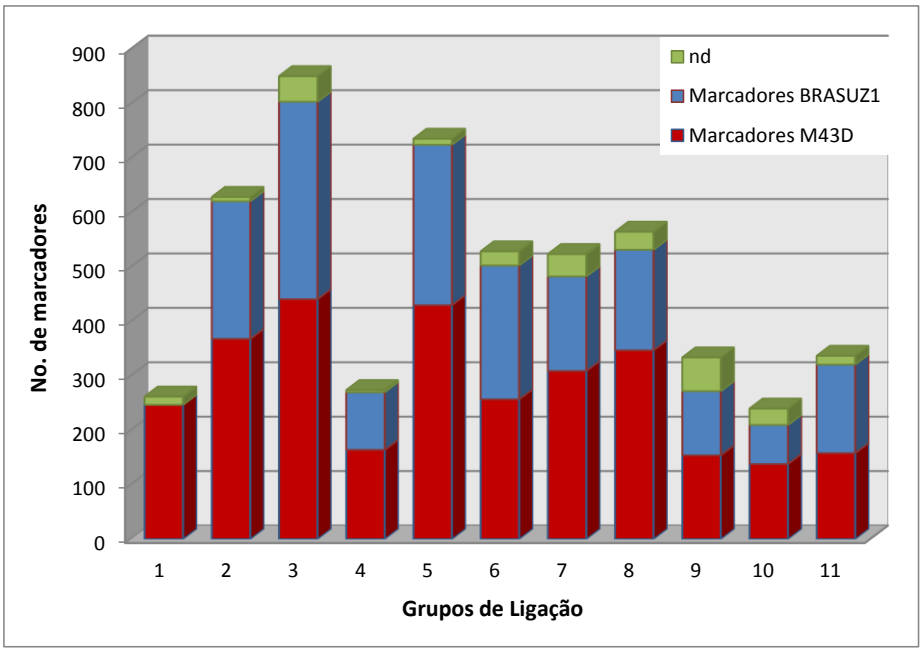
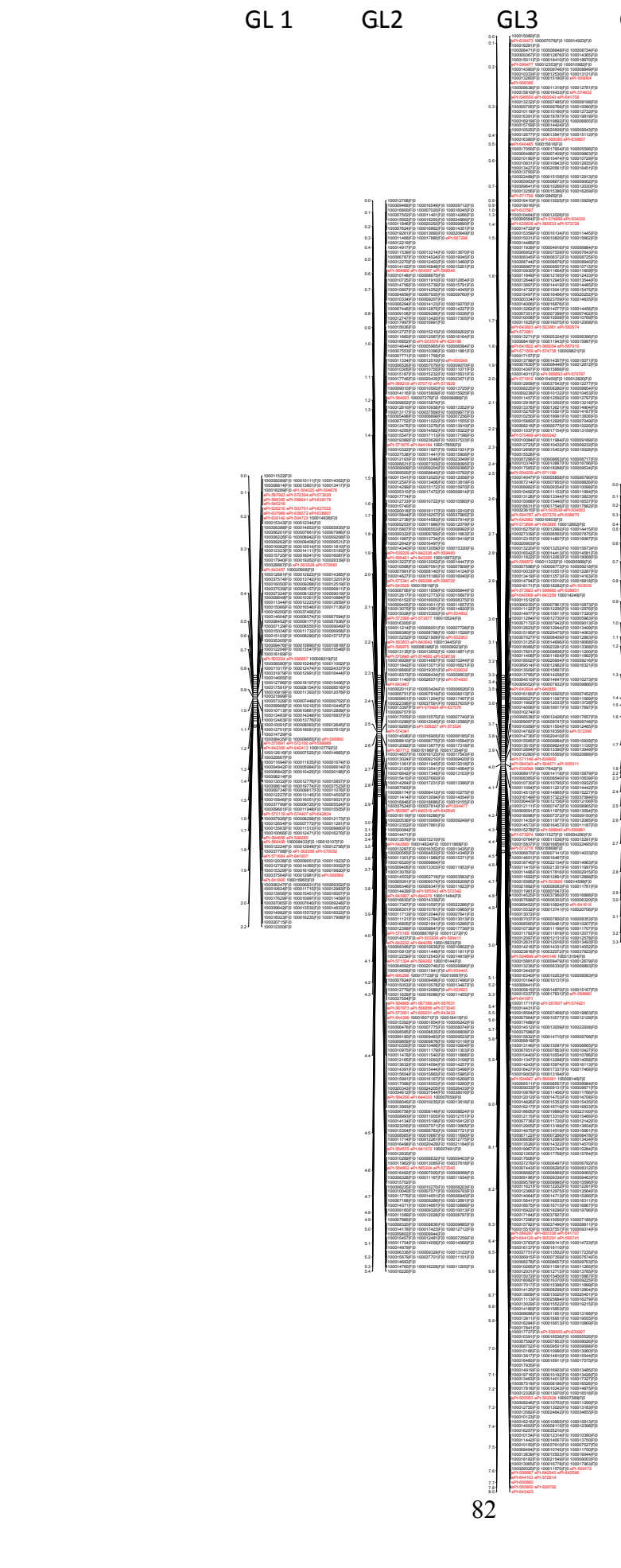


Figura 22. Distribuição do número de marcadores mapeados segundo a segregação dos parentais em cada um dos 11 grupos de ligação. Nas colunas se observam os marcadores polimórficos que segregaram a partir do parental M43D (vermelho), a partir do parental BRASUZ1 (azul); e os marcadores sem origem definida (verde).

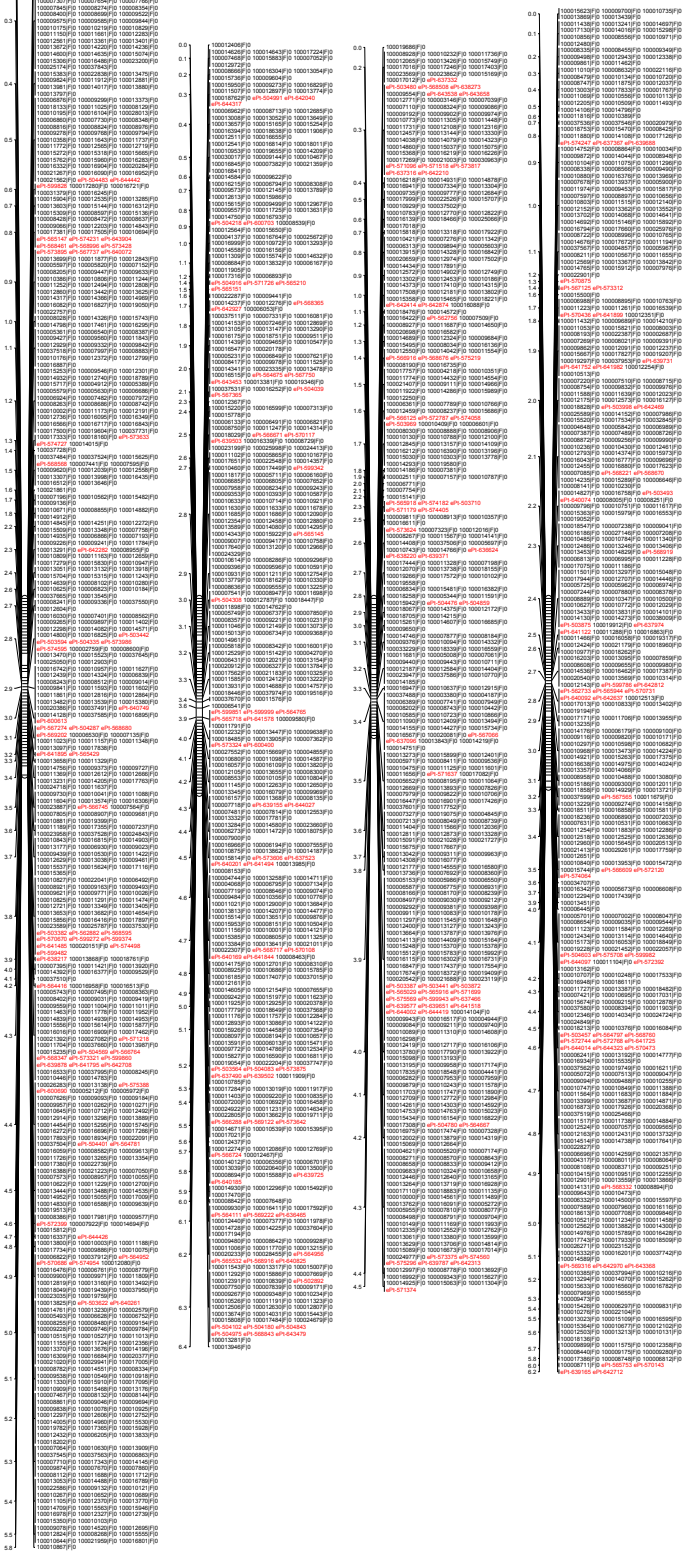
Mapas genéticos de alta densidade são úteis para a identificação de marcadores "genome-wide" ligados a QTLs, isolamento de genes via clonagem posicional, mapeamento comparativo, e estudos de evolução de genoma (Varshney 2007). Em *Eucalyptus* mapas genéticos utilizando marcadores RAPD, RFLP e microssatélites têm sido produzidos com uma densidade de 100 a 300 marcadores (Grattapaglia e Sederoff, 1994; Byrne, Murrell *et al.*, 1995; Verhaegen e Plomion, 1996; Brondani, Brondani *et al.*, 2002; Thamarus, Groom *et al.*, 2002; Brondani, Williams *et al.*, 2006). Recentemente, aplicando a tecnologia DArT baseada em microarranjo em combinação com microssatélites, foram construídos mapas genéticos para cruzamentos intraespecíficos de *E. globulus* com 1060 e 564 marcadores mapeados (Hudson, Kullan *et al.*, 2011), ou seja número equivamente àquele obtido para o pedigree intraespecífico de *E. grandis* deste estudo, 598 marcadores. Por outro lado mapas de mais alta densidade com cerca de 2300 e 2.400 marcadores DArT baseados em microarranjo foram gerados respectivamente para pedigrees interespecíficos (*E. grandis x E. urophylla*) (Kullan, Van Dyk *et al.*, 2011; Petroli, Sansaloni *et al.*, 2011) possivelmente resultantes do nível de polimorfismo substancialmente maior entre os parentais. Neste estudo, utilizando a tecnologia DArT-Seq baseada em NGS, foi possível mapear 4.683 marcadores Insilico DArT, ou seja, do tipo presença/ausência o que corresponde a cerca de 8 vezes mais do que o número de

marcadores mapeados com o microarranjo DArT. Caso um pedigree interspecífico tivesse sido utilizado neste experimento de validação da tecnologia DArT-Seq, e mantendo as mesmas proporções comparativas observadas, a expectativa seria de um total de 8 vezes ~2300 marcadores o que somaria cerca de 16.000 a 17.000 marcadores potencialmente mapeados. Finalmente, vale lembrar que além dos 4.683 marcadores *InSilico* DArT mapeados ainda restam cerca de 4.000 a 5.000 marcadores SNPs detectados a serem integrados a este mapa.

O problema que surge com este número muito elevado de marcadores é o devido ordenamento ao longo dos cromossomos. O pedigree utilizado neste estudo evidentemente é muito limitado com apenas 89 indivíduos. Pedigrees com várias centenas de indivíduos seriam necessários para se ter alguma possibilidade de posicionar números da ordem de vários milhares de marcadores com elevada confiança estatística, isto posto que um software que conseguir lidar com esta quantidade de dados seja disponível o que ainda não é exatamente o caso, apesar de esforços recentes neste sentido (Cheema e Dicks, 2009). Neste estudo foi possível propor um mapa genético com mais de 5.000 marcadores com base em uma abordagem via distâncias Hamming cuja robustez não pode ser avaliada vis a vis com outros softwares comumente utilizados. Do ponto de vista de representação gráfica do mapa no momento o que resta é a construção de mapas genéticos do tipo framework, utilizando o softare JoinMap nos quais um número muito limitado de marcadores é posicionado com algum nível de confiança. Para a progênie do BRASUZ1 foi possível, portanto, posicionar somente 1390 marcadores dos mais de 5.000 tentativamente mapeados com o software da DArT (Figura 23).


GL 1
GL2
GL3
GL4

GL 5 GL6 GL7 GL8



GL 9 GL10 GL11

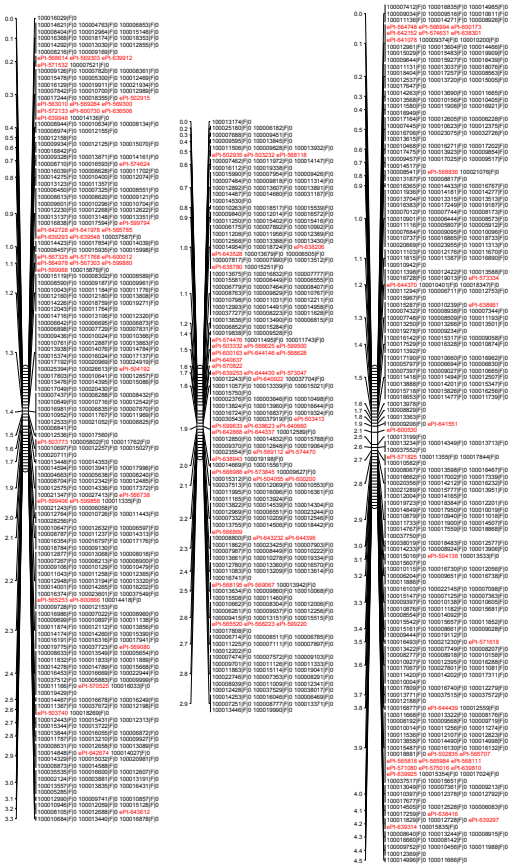


Figura 23. Mapa de ligação da família BRASUZ1. Um total de 5.281 marcadores foram posicionados com elevada confiança (mapa framework), sendo 4.683 marcadores DArT baseados em NGS (em preto) e 598 marcadores DArT baseados em microarranjo (em vermelho). Distâncias cumulativas são mostradas à esquerda de cada grupo de ligação.

4.4. CONCLUSÕES E PERSPECTIVAS

Os resultados preliminares deste estudo foram recentemente publicados de forma sintética (Sansaloni, Petroli *et al.*, 2011) (Anexo 3) e serão objeto de uma publicação mais detalhada em breve. Estes resultados demonstram que a utilização combinada da técnica DArT como método robusto de redução da complexidade junto ao protocolo otimizado de sequências indexadoras e sequenciamento NGS, pode fornecer cerca de 7 a 8 vezes mais marcadores dominantes mapeados do que o método DArT convencional baseado em

microarranjos. Resultados recentes da genotipagem de uma população de 1000 indivíduos híbridos de cerca de 40 famílias geneticamente não relacionadas de *E. grandis x E. urophylla* com a mesma metodologia utilizada neste trabalho forneceu cerca de 22.000 marcadores dominantes InSilico DArT e aproximadamente 8.000 SNPs de alta qualidade (Lima e Grattapaglia, dados não publicados).

Metodologias de genotipagem-por-sequenciamento ainda constituem um avanço muito recente e até o momento somente três trabalhos foram efetivamente publicados com pequenas diferenças quanto ao método de redução de complexidade e parâmetros de filtragem das sequências para a declaração de genótipos. Em *Drosophila* o método foi denominado MSG (*Multiplexed Shotgun Sequencing*) e foi utilizado com sucesso para mapear geneticamente mais de 400 contigs previamente não integrados no genoma de *D. simulans* (Andolfatto, Davison *et al.*, 2011). A metodologia batizada de GbS foi demonstrada em genomas mais complexos do que *Drosophila* ao ser aplicada para o mapeamento genético em milho, trigo e cevada. Em cevada foi possível mapear 24.186 *tags* do tipo presença/ausência, enquanto que em milho foram mapeados 25.185 marcadores SNPs (Elshire, Glaubitz *et al.*, 2011). Em um trabalho posterior de GbS com uma metodologia otimizada que se baseia no uso de duas enzimas de restrição, mais de 34.000 SNPs e 240.000 *tags* foram mapeados em um mapa de referência de cevada e 20.000 SNPs e 367.000 *tags* em um mapa de referência de trigo (Poland, Brown *et al.*, 2012). Embora a aplicação de GbS em princípio demandasse a disponibilidade de um genoma de referência, trabalhos recentes mostraram que esta não é uma condição estritamente necessária e algoritmos de análise que dispensam este recuso já são disponíveis (Willing, Hoffmann *et al.*, 2011; Poland, Brown *et al.*, 2012).

O desenvolvimento de metodologias GbS otimizadas para espécies de plantas com genomas complexos representa um avanço formidável na história do desenvolvimento e aplicações da análise genética com marcadores moleculares. Representa ainda uma variação importante do uso das novas tecnologias de sequenciamento que podem se provar mais úteis justamente para a genotipagem do que propriamente para o sequenciamento de genomas, onde novas tecnologias que geram sequências mais longas deverão surgir e se tornar preferenciais. Além dos números de marcadores várias ordens de magnitude superiores aos números tipicamente usados até hoje, as técnicas GbS tem um custo accessível da ordem de algumas dezenas de dólares por amostra e utilizam reagentes universais, vantagem equivalente aos marcadores RAPD que em seu tempo causaram uma verdadeira revolução na análise genética de plantas e animais. A combinação do elevado número de marcadores, baixo

custo da genotipagem e metodologia relativamente accessível, aponta para um uso crescente de GbS nos próximos anos. GbS deverá ser a ferramenta de escolha nas mais diversas aplicações em estudos de genética de populações, investigações evolutivas e principalmente em apoio ao melhoramento acelerando e aumentando a precisão da seleção direcional de características multifatoriais complexas.

Paralelamente à adoção crescente de GbS, otimizações das plataformas de sequenciamento permitirão a análise multiplexada de um número cada vez maior de amostras, reduzindo os custos progressivamente. Especificamente no caso de espécies de *Eucalyptus*, organismo alvo deste trabalho, GbS poderá ser o "*workhorse*" (cavalo de batalha) em aplicações operacionais da seleção genômica. Resultados experimentais recentes utilizando cerca de 3.500 marcadores DArT convencionais dominantes já permitiram construir modelos preditivos que alcançaram acurácias seletivas equivalentes àquelas obtidas com seleção fenotípica tradicional (Resende, Resende *et al.*, 2012). A possibilidade de genotipar entre 10 e 20 mil marcadores de alta qualidade não apenas deverá permitir a adoção de seleção genômica em populações com tamanho efetivo da ordem de Ne = 100 (Grattapaglia e Resende, 2010), mas poderá ainda fornecer acurácias seletivas melhores às obtidas com os processos convencionais de seleção, concretizando, assim, as perspectivas de seleção assistida por marcadores apontadas aos melhoristas florestais mais de 20 anos atrás.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Abraf. Associação Brasileira de Produtores de Florestas Plantadas 2011.

Akbari, M., P. Wenzl, *et al*. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet, v.113, n.8, Nov, p.1409-20. 2006.

Alves-Freitas, D. M. T., A. Kilian, *et al*. Towards a high-density DArT (Diversity Arrays Technology) microarray for high-throughput genotyping of Pinus taeda and closely related species. Resumos do 56º Congresso Brasileiro de Genética. Santos, 2010. 182 p.

Andolfatto, P., D. Davison, *et al*. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res, v.21, n.4, Apr, p.610-7. 2011.

Baird, N. A., P. D. Etter, *et al*. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One, v.3, n.10, p.e3376. 2008.

Barbazuk, W. B., S. J. Emrich, *et al*. SNP discovery via 454 transcriptome sequencing. The Plant Journal, v.51, n.5, p.910-918. 2007.

Barchi, L., S. Lanteri, *et al*. Identification of SNP and SSR markers in eggplant using RAD tag sequencing. Bmc Genomics, v.12, Jun 10. 2011.

Baril, C. P., D. Verhaegen, *et al*. Structure of the specific combining ability between two species of Eucalyptus. I. RAPD data. Theoretical and Applied Genetics, v.94, p.796-803. 1997.

Bedo, J., P. Wenzl, *et al*. Precision-mapping and statistical validation of quantitative trait loci by machine learning. BMC Genetics, v.9, n.1, p.35. 2008.

Bernatzky, R. e S. D. Tanksley. Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. Genetics, v.112, n.4, Apr, p.887-98. 1986.

Bhattramakki, D., M. Dolan, *et al*. Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol Biol, v.48, n.5-6, Mar-Apr, p.539-47. 2002.

Botstein, D., R. L. White, *et al*. Construction of a genetic map in man using restriction legnth polymorphism. Am J Hum Genet, v.32, p.314-331. 1980.

Brockman, W., P. Alvarez, *et al*. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Research, January 1, 2008. 2008.

Brondani, R., C. Brondani, *et al*. Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. Theor Appl Genet, v.97, p.816 - 827. 1998.

Brondani, R., E. Williams, *et al*. A microsatellite-based consensus linkage map for species of Eucalyptus and a novel set of 230 microsatellite markers for the genus. Bmc Plant Biology, v.6, p.20. 2006.

Brondani, R. P. V., C. Brondani, *et al*. Towards the construction of a genus wide reference linkage map for *Eucalytpus* based on microsatellite markers. Molecular and General Genomics, v.267, p.338-347. 2002.

Brondani, R. P. V. e D. Grattapaglia. Cost-effetive method to synthetise a fluorescent internal DNA standard for automated fragment sizing. . Biotechniques, v.31, p.793-795. 2001.

Brooker, M. I. H. A new classification of genus Eucalyptus L´Her. (Myrtaceae). 2000 (Australian Systematic Botany)

Brown, A. H. D., A. C. Matheson, *et al*. Estimation of mating system of Eucalyptus obliqua L'Herit. by using allozyme polymorphisms. Australian Journal of Botany, v.23, p.931–949. 1975.

Brown, R. G., G. P. Gill, *et al*. Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U S A, v.101, n.42, p.15255-15260. 2004.

Bundock, P. C., M. Hayden, *et al*. Linkage maps of Eucalyptus globulus using RAPD and microsatellite markers. Silvae Genetica, v.49, n.4-5, p.223-232. 2000.

Butcher, P., M. Mcdonald, *et al*. Congruence between environmental parameters, morphology and genetic structure in Australia's most widely distributed eucalypt, *Eucalyptus camaldulensis*. Tree Genetics & Genomes, v.5, n.1, p.189-210. 2009.

Butcher, P. A., A. Otero, *et al*. Nuclear RFLP variation in Eucalyptus camaldulensis Dehnh. from northern Australia. Heredity, v.88, n.5, May, p.402-12. 2002.

Byrne, M., M. I. Marquezgarcia, *et al*. Conservation and Genetic Diversity of Microsatellite loci in the Genus Eucalyptus. Australian Journal of Botany, v.44, 01/01/1996, p.331-341. 1996.

Byrne, M., J. C. Murrell, *et al*. An integrated genetic linkage map for eucalypts using RFLP, RAPD and isozyme markers. Theoretical and Applied Genetics, v.91, p.869 - 875. 1995.

Byrne, M., T. L. Parrish, *et al*. Nuclear RFLP diversity in Eucalyptus nitens. Heredity v.81, p.225–233. 1998.

Carlson, J. E., L. K. Tulsieram, *et al*. Segregation of random amplified DNA markers in F1 progeny of conifers. Theoretical and Applied Genetics, v.83, p.194-200. 1991.

Chambers, G. K. e E. S. Macavoy. Microsatellites: consensus and controversy. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, v.126, n.4, p.455-476. 2000.

Cheema, J. e J. Dicks. Computational approaches and software tools for genetic linkage map estimation in plants. Briefings in Bioinformatics, v.10, n.6, November 1, 2009, p.595-608. 2009.

Costa, P., D. Pot, *et al*. A genetic map of Maritime pine based on AFLP, RAPD and protein markers. Theoretical and Applied Genetics, v.100, p.39-48. 2000.

Craig, D. W., J. V. Pearson, *et al*. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Meth, v.5, n.10, p.887-893. 2008.

Crossa, J., G. D. L. Campos, *et al*. Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics, v.186, n.2, October 2010, p.713-724. 2010.

Cupertino, F., J. Leal, *et al*. Genetic diversity of Eucalyptus hybrids estimated by genomic and EST microsatellite markers. Biologia Plantarum, v.55, n.2, p.379-382. 2011.

Davey, J. W., P. A. Hohenlohe, *et al*. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, v.12, n.7, Jul, p.499-510. 2011.

Devey, M. E., K. D. Jermstad, *et al*. Inheritance of RFLP loci in a loblolly pine three-generation pedigree. TAG Theoretical and Applied Genetics, v.83, n.2, p.238-242. 1991.

Doughty, R. W. The eucalyptus: a natural and commercial history of the gum tree: Johns Hopkins University Press. 2000

Doyle, J. J. e J. L. Doyle. Isolation of plant DNA from fresh tissue. Focus, v.12, p.13-15. 1987.

Eckert, A. J., B. Pande, *et al*. High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (Pinus taeda L.). Tree Genetics & Genomes, v.5, n.1, Jan, p.225-234. 2009.

Edwards, A., A. Civitello, *et al*. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet, v.49, n.4, Oct, p.746-56. 1991.

Eldridge, K., J. Davidson, *et al*. Eucalypt Domestication and Breeding: Oxford University Press, USA. 1994

Elliott, C. P. e M. Byrne. Phylogenetics and the conservation of rare taxa in the *Eucalyptus angustissima* complex in Western Australia. Conservation Genetics, v.5, n.1, p.39-47. 2004.

Elshire, R. J., J. C. Glaubitz, *et al*. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS One, v.6, n.5, p.e19379. 2011.

Emerson, K. J., C. R. Merz, *et al*. Resolving postglacial phylogeography using high-throughput sequencing. Proceedings of the National Academy of Sciences of the United States of America, v.107, n.37, Sep 14, p.16196-16200. 2010.

Ferreira, M. Melhoramento e a Sivilcutura intensiva clonal. IPEF, v.45, p.8. 1992.

Ferreira, M. E. e D. Grattapaglia. Introdução ao uso de marcadores moleuculares em análise genética. Brasília: Embrapa. 1998. 220 p. (Introdução ao uso de marcadores moleuculares em análise genética)

Freeman, J. S., J. M. O'reilly-Wapstra, *et al*. Quantitative trait loci for key defensive compounds affecting herbivory of eucalypts in Australia. New Phytologist, v.178, n.4, p.846-851. 2008.

Gaiotto, F. A., M. Bramucci, *et al*. Estimation of outcrossing rate in a breeding population of Eucalyptus urophylla with dominant RAPD and AFLP markers. TAG Theoretical and Applied Genetics, v.95, n.5, p.842-849. 1997.

Gale, M. D. e K. M. Devos. Plant Comparative Genetics after 10 Years. Science, v.282, n.5389, October 23, 1998, p.656-659. 1998.

Gion, J.-M., A. Carouche, *et al*. Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: Eucalyptus. BMC Genomics, v.12, n.1, p.301. 2011.

Gion, J. M., P. Rech, *et al*. Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. Molecular Breeding, v.6, p.441-449. 2000.

Glaubitz, J. C., L. C. Emebiri, *et al*. Dinucleotide microsatellites from Eucalyptus sieberi: inheritance, diversity, and improved scoring of single-base differences. Genome, v.44, n.6, Dec, p.1041-5. 2001.

Gonzalez-Martinez, S. C., E. Ersoz, *et al*. DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in Pinus taeda L. Genetics, v.172, n.3, Mar, p.1915-26. 2006.

Grattapaglia, D. Integrating genomics into Eucalyptus breeding. Genet Mol Res, v.3, n.3, p.369-79. 2004.

Grattapaglia, D. e M. Kirst. Eucalyptus applied genomics: from gene sequences to breeding tools. New Phytologist, v.179, n.4, p.911-929. 2008.

Grattapaglia, D., C. Plomion, *et al*. Genomics of growth traits in forest trees. Current Opinion in Plant Biology, v.12, n.2, p.148-156. 2009.

Grattapaglia, D. e M. D. V. Resende. Genomic Selection in forest tree breeding. Tree Genetics & Genomes, v.7, n.2, p.241-255. 2010.

Grattapaglia, D., V. J. Ribeiro, *et al*. Retrospective selection of elite parent trees using paternity testing with microsatellite markers: an alternative short term breeding tactic for Eucalyptus. Theor Appl Genet, v.109, n.1, Jun, p.192-9. 2004.

Grattapaglia, D., C. P. Sansaloni, *et al*. Genomic Selection In Eucalyptus: Marker Assisted Selection Coming To Reality In Forest Trees. Plant and Animal Genome XVIII Conference. San Diego**:** Abstract W 237 p. 2010.

Grattapaglia, D. e R. Sederoff. Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. Genetics, v.137, n.4, Aug, p.1121-37. 1994.

Grattapaglia, D., O. B. Silva-Junior, *et al*. High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. BMC plant biology, v.11, n.1, p.65. 2011.

Grosse, W. M., S. M. Kappes, *et al*. Single nucleotide polymorphism (SNP) discovery and linkage mapping of bovine cytokine genes. Mammalian Genome, v.10, n.11, p.1062-1069. 1999.

Hamming, R. W. Error detecting and error correcting codes. Bell System Technical Journal v.29, n.2, p.147-160. 1950.

He, X. e Å. Bjørnstad. Diversity of North European oat analyzed by SSR, AFLP and DArT markers. TAG Theoretical and Applied Genetics, p.1-14. 2012.

Heaton, M. P., G. P. Harhay, *et al*. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. Mammalian Genome, v.13, n.5, p.272-281. 2002.

Helentjaris, T., D. F. Weber, *et al*. Use of monosomics to map cloned DNA fragments in maize. Proc Natl Acad Sci U S A, v.83, n.16, Aug, p.6035-9. 1986.

Hillier, L. W., G. T. Marth, *et al*. Whole-genome sequencing and variant discovery in C. elegans. Nat Meth, v.5, n.2, p.183-188. 2008.

Hippolyte, I., F. Bakry, *et al*. A saturated SSR/DArT linkage map of Musa acuminata addressing genome rearrangements among bananas. BMC Plant Biol, v.10, p.65. 2010.

Hoelzel, A. e A. Green. PCR protocols and population analysis by direct DNA sequencing and PCR - based DNA fingerprinting. 1998

Hohenlohe, P. A., S. J. Amish, *et al*. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Molecular Ecology Resources, v.11, Mar, p.117-122. 2011.

Horton, M. W., A. M. Hancock, *et al*. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet, v.44, n.2, p.212-216. 2012.

Hudson, C., A. Kullan, *et al*. High synteny and colinearity among Eucalyptus genomes revealed by high-density comparative genetic mapping. Tree Genetics & Genomes, p.1-14. 2011.

Huttner, E., P. Wenzl, *et al*. Diversity Arrays Technology: A novel tool for harnessing the genetic potential of orphan crops. Conference of The World Biological Forum Discovery to Delivery: BioVisionAlexandria: CABI Publishing, 2004. p.

Huynh, B.-L., H. Wallwork, *et al*. Quantitative trait loci for grain fructan concentration in wheat (<i>Triticum aestivum</i> L.). TAG Theoretical and Applied Genetics, v.117, n.5, p.701-709. 2008.

Hyten, D. L., S. B. Cannon, *et al*. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. Bmc Genomics, v.11, Jan 15, p.-. 2010.

Jaccoud, D., K. Peng, *et al*. Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res, v.29, n.4, Feb 15, p.E25. 2001.

Jones, T. H., R. E. Vaillancourt, *et al*. Detection and visualization of spatial genetic structure in continuous Eucalyptus globulus forest. Mol Ecol, v.16, n.4, Feb, p.697-707. 2007.

Junghans, D. T., A. C. Alfenas, *et al*. Resistense to rust (*Puccinia psidii* Winter) in Eucalyptus: mode of inheritance and mapping of a major gene with RAPD markers. Theoretical and Applied Genetics, v.108, n.175-180. 2003.

Keane, P. J., G. A. Kile, *et al*. Diseases and pathogens of eucalypts: CSIRO. 2000

Keil, M. e A. R. Griffin. Use of random amplified polymorphic DNA (RAPD) markers in the discrimination and verification of genotypes in Eucalyptus. . Theoretical and Applied Genetics v.89, p.442–450. 1994.

Kijas, J. W., J. A. Lenstra, *et al*. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. PLoS Biol, v.10, n.2, p.e1001258. 2012.

Kilian, A., E. Huttner, *et al*. The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. International Congress In the Wake of the Double Helix: From the Green Revolution to the Gene Revolution: May 27-31 2003, v.2003, p.443 - 461. 2005.

Kirst, M., C. M. Cordeiro, *et al*. Power of microsatellite markers for fingerprinting and parentage analysis in Eucalyptus grandis breeding populations. J Hered, v.96, n.2, Mar-Apr, p.161-6. 2005.

Kruglyak, L. e D. A. Nickerson. Variation is the spice of life. Nat Genet, v.27, n.3, Mar, p.234-6. 2001.

Külheim, C., S. H. Yeoh, *et al*. Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways. BMC Genomics, v.10, n.452. 2009.

Kullan, A., M. Van Dyk, *et al*. High-density genetic linkage maps with over 2,400 sequence-anchored DArT markers for genetic dissection in an F2 pseudo-backcross of Eucalyptus grandis x E. urophylla. Tree Genetics & Genomes, p.1-13. 2011.

Levinson, G. e G. A. Gutman. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Molecular Biology and Evolution, v.4, n.3, May 1, 1987, p.203-221. 1987.

Li, H. e R. Durbin. Fast and accurate short read alignment with Burrowsâ€“Wheeler transform. Bioinformatics, v.25, n.14, July 15, 2009, p.1754-1760. 2009.

Lima, B., O. Silva-Junior, *et al*. Assessment of SNPs for linkage mapping in Eucalyptus: construction of a consensus SNP/microsatellite map from two unrelated pedigrees. BMC Proceedings, v.5, n.Suppl 7, p.P31. 2011.

Litt, M. e J. A. Luty. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet, v.44, n.3, Mar, p.397-401. 1989.

Mamani, E. M. C., N. W. Bueno, *et al*. Positioning of the major locus for *Puccinia psidii* rust resistance (*Ppr1*) on the *Eucalyptus* reference map and its validation across unrelated pedigrees. Tree Genetics & Genomes, v.6, n.6, Dec, p.953-962. 2010.

Marcucci Poltri, S. N., N. Zelener, *et al*. Selection of a seed orchard of Eucalyptus dunnii based on genetic diversity criteria calculated using molecular markers. Tree Physiol, v.23, n.9, Jun, p.625-32. 2003.

Mardis, E. R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, v.9, p.387-402. 2008.

Marques, C. M., J. A. Araújo, *et al*. AFLP genetic maps of *Eucalyptus globulus* and *E. tereticornis*. Theoretical and Applied Genetics, v.96, p.727-737. 1998.

Maughan, P. J., S. M. Yourstone, *et al*. Single-Nucleotide Polymorphism genotyping in mapping populations via genomic reduction and next generation sequencing: proof of concept. The Plant Genome, v.3, n.3, p.166-178. 2010.

_____. SNP Discovery via Genomic Reduction, Barcoding, and 454-Pyrosequencing in Amaranth. Plant Gen., v.2, n.3, p.260-270. 2009.

Mccartney, C., R. Stonehouse, *et al*. Mapping of the oat crown rust resistance gene <i>Pc91</i&gt. TAG Theoretical and Applied Genetics, v.122, n.2, p.317-325. 2011.

Menendez, D., O. Krysiak, *et al*. A SNP in the flt-1 promoter integrates the VEGF system into the p53 transcriptional network. Proceedings of the National Academy of Sciences of the United States of America, v.103, n.5, January 31, 2006, p.1406-1411. 2006.

Messier, W., S.-H. Li, *et al*. The birth of microsatellites. Nature, v.381, n.6582, p.483-483. 1996.

Metzker, M. L. Sequencing technologies - the next generation. Nat Rev Genet, v.11, n.1, Jan, p.31-46. 2010.

Meuwissen, T. H., B. J. Hayes, *et al*. Prediction of total genetic value using genome-wide dense marker maps. Genetics, v.157, n.4, Apr, p.1819-29. 2001.

Milczarski, P., H. Bolibok-Bragoszewska, *et al*. A High Density Consensus Map of Rye (Secale cereale L.) Based on DArT Markers. Plos One, v.6, n.12, Dec 6. 2011.

Miller, M. R., J. P. Dunham, *et al*. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res, v.17, n.2, Feb, p.240-8. 2007.

Mora, A. L. e C. H. Garcia. A cultura do eucalipto no Brasil. São Paulo - SBS, p.1. 2000.

Moran, G. F. Patterns of genetic diversity in Australian tree species. New Forests v.6, p.49–66. 1992.

Moran, G. F. e J. C. Bell. Eucalyptus. In: S. D. Tanksley e T. J. Orton (Ed.). Isozymes in plant genetics and breeding. Amsterdam: Elsevier, 1983. Eucalyptus, p.423-441

Mueller, U. G. e L. L. Wolfenbarger. AFLP genotyping and fingerprinting. Trends in Ecology & Evolution, v.14, n.10, p.389-394. 1999.

Myburg, A. A., A. R. Griffin, *et al*. Comparative genetic linkage maps of Eucalyptus grandis, Eucalyptus globulus and their F1 hybrid based on a double pseudo-backcross mapping approach. Theor Appl Genet, v.107, n.6, Oct, p.1028-42. 2003.

Myburg, A. A., B. M. Potts, *et al*. Eucalyptus. In: C. Kole (Ed.). Genome mapping and molecular breeding in plants. New York, NY, USA: Springer, v.Vol. 7: Forest trees., 2007. Eucalyptus, p.115 - 160

Myles, S., J. M. Chia, *et al*. Rapid Genomic Characterization of the Genus Vitis. PLoS One, v.5, n.1, Jan 13, p.-. 2010.

Neale, D. B. Genomics to tree breeding and forest health. Current Opinion in Genetics & Development, v.17, n.6, p.539-544. 2007.

Nesbitt, K. A., Potts, B.M., Vaillancourt, R.E., West, A.K., Reid, J.B.,. Partitioning and distribution of RAPD variation in a forest tree species, Eucalyptus globulus (Myrtaceae). Heredity, v.74, p.628-637. 1995.

Neves, L., E. M. C. Mamani, *et al*. A high-density transcript linkage map with 1,845 expressed genes positioned by microarray-based Single Feature Polymorphisms (SFP) in Eucalyptus. BMC Genomics, v.12, n.1, p.189. 2011.

Niedziela, A., P. Bednarek, *et al*. Aluminum tolerance association mapping in triticale. BMC Genomics, v.13, n.1, p.67. 2012.

Novaes, E., D. R. Drost, *et al*. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics, v.9, p.312. 2008.

O' Malley, D. M., D. Grattapaglia, *et al*. Molecular markers, forest genetics and tree breeding. In: J. P. Gustafson e R. B. Flavell (Ed.). Genomes of plants and animals. New York: Plenum Press, 1996. Molecular markers, forest genetics and tree breeding., p.87-102

O'rourke, S. M., J. Yochem, *et al*. Rapid Mapping and Identification of Mutations in Caenorhabditis elegans by Restriction Site-Associated DNA Mapping and Genomic Interval Pull-Down Sequencing. Genetics, v.189, n.3, Nov, p.767-U108. 2011.

Oliver, R., E. Jellen, *et al*. New Diversity Arrays Technology (DArT) markers for tetraploid oat (Avena magna Murphy et Terrell) provide the first complete oat linkage map and markers linked to domestication genes from hexaploid A. sativa L. TAG Theoretical and Applied Genetics, v.123, n.7, p.1159-1171. 2011.

Ottewell, K. M., S. C. Donnellan, *et al*. Multiplexed microsatellite markers for the genetic analysis of Eucalyptus leucoxylon (Myrtaceae) and their utility for ecological and breeding studies in other eucalyptus species. J Hered, v.96, n.4, Jul-Aug, p.445-51. 2005.

Paterson, A. H., J. E. Bowers, *et al*. Comparative Genomics of Plant Chromosomes. The Plant Cell Online, v.12, n.9, September 1, 2000, p.1523-1540. 2000.

Paux, E., P. Sourdille, *et al*. A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B. Science, v.322, n.5898, October 3, 2008, p.101-104. 2008.

Payn, K. G., W. S. Dvorak, *et al*. Microsatellite diversity and genetic structure of the commercially important tropical tree species Eucalyptus urophylla, endemic to seven islands in eastern Indonesia. Tree Genetics & Genomes, v.4, p.519 - 530. 2008.

Perrier, X. e J. P. Jacquemoud-Collet. DARwin 5.0.158. Http://Darwin.Cirad.Fr/Darwin 2006.

Petroli, C., C. Sansaloni, *et al*. Genomic characterization, high-density mapping and anchoring of DArT markers to the reference genome of Eucalyptus. BMC Proceedings, v.5, n.Suppl 7, p.P35. 2011.

Plomion, C., B. H. Liu, *et al*. Genetic analysis using trans-dominant linked markers in an F-2 family. Theoretical and Applied Genetics, v.93, n.7, Nov, p.1083-1089. 1996.

Poke, F. S., R. E. Vaillancourt, *et al*. Genomic research in Eucalyptus. Genetica, v.125, n.1, Sep, p.79-101. 2005.

Poland, J. A., P. J. Brown, *et al*. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. Plos One, v.7, n.2, p.e32253. 2012.

Potts, B. M. Genetic improvement of eucalypts. Encyclopedia of forest sciences. Oxford UK: Elsevier**:** 1480-1490 p. 2004.

Potts, B. M. e L. A. Pederick. Morphology, phylogeny, origin, distribution and genetic diversity of the eucalypts. In: B. N. Brown (Ed.). Diseases and Pathogens of Eucalypts. Collingwood: CSIRO Publishing, 2000. Morphology, phylogeny, origin, distribution and genetic diversity of the eucalypts, p.11-34

Resende, M. D. V., M. F. R. Resende, *et al*. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytologist, v.194, n.1, p.116-128. 2012.

Rodríguez-Suárez, C., M. Giménez, *et al*. Development of wild barley (<i>Hordeum chilense</i>)-derived DArT markers and their use into genetic and physical mapping. TAG Theoretical and Applied Genetics, v.124, n.4, p.713-722. 2012.

Rowe, H. C., S. Renaut, *et al*. RAD in the realm of next-generation sequencing technologies. Molecular Ecology, v.20, n.17, Sep, p.3499-3502. 2011.

Sale, M. M., B. M. Potts, *et al*. Relationships within *Eucalyptus* using chloroplast DNA. . Aust. Syst. Bot., v.6, p.p.127-138. 1993.

Sansaloni, C., C. Petroli, *et al*. Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proceedings, v.5, n.Suppl 7, p.P54. 2011.

Sansaloni, C. P. Desenvolvimento, caracterização e mapeamento de microssatélites de tetra e pentanucleotídeos em Eucalyptus spp. Departamento de Biologia Celular - Curso de Pós-Graduação em Biologia molecular, Universidade de Brasília, Brasília, 2008. 114 p.

Sansaloni, C. P., C. D. Petroli, *et al*. A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in *Eucalyptus*. Plant Methods, v.6, p.16. 2010.

Scaglione, D., A. Acquadro, *et al*. RAD tag sequencing as a source of SNP markers in Cynara cardunculus L. Bmc Genomics, v.13, Jan 3. 2012.

Schlotterer, C. Microsatellites. In: (Ed.). Molecular genetic analysis of populations - A practical approach.: Oxford University Press, 1998. Microsatellites

Schmid, K. J., T. R. Sorensen, *et al*. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. Genome Res, v.13, n.6A, Jun, p.1250-7. 2003.

Sessitsch, A., E. Hackl, *et al*. Diagnostic microbial microarrays in soil ecology. New Phytologist, v.171, n.4, p.719-736. 2006.

Simko, I., I. Eujayl, *et al*. Empirical evaluation of DArT, SNP, and SSR marker-systems for genotyping, clustering, and assigning sugar beet hybrid varieties into populations. Plant Sci, v.184, Mar, p.54-62. 2012.

Steane, D., N. Conod, *et al*. A comparative analysis of population structure of a forest tree, Eucalyptus globulus (Myrtaceae), using microsatellite markers and quantitative traits. Tree Genetics & Genomes, v.2, p.30 - 38. 2006.

Steane, D. A., D. Nicolle, *et al*. Population genetic analysis and phylogeny reconstruction in Eucalyptus (Myrtaceae) using high-throughput, genome-wide genotyping. Mol Phylogenet Evol, v.59, n.1, Feb 7, p.206-224. 2011.

Steane, D. A., R. E. Vaillancourt, *et al*. Development and characterisation of microsatellite loci in Eucalyptus globulus (Myrtaceae). Silvae Genetica, v.50, n.2, p.89-91. 2001.

Steane, D. A., A. K. West, *et al*. Restriction fragment length polymorphisms in chloroplast DNA from six species of *Eucalyptus*. Aust. J. Bot., v.39, p.p.399-414. 1992.

Stephens, M., N. J. Smith, *et al*. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet, v.68, n.4, Apr, p.978-89. 2001.

Tanksley, S. D., N. D. Young, *et al*. RFLP Mapping in Plant Breeding: New Tools for an Old Science. Nat Biotech, v.7, n.3, p.257-264. 1989.

Tautz, D. Hypervariabflity of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research, v.17, n.16, August 25, 1989, p.6463-6471. 1989.

Thamarus, K., K. Groom, *et al*. A genetic linkage map for Eucalyptus globulus with candidate loci for wood, fibre and floral traits. Theor Appl Genet, v.104, p.379 - 387. 2002.

Thornsberry, J. M., M. M. Goodman, *et al*. Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet, v.28, n.3, Jul, p.286-9. 2001.

Thudi, M., A. Bohra, *et al*. Novel SSR Markers from BAC-End Sequences, DArT Arrays and a Comprehensive Genetic Map with 1,291 Marker Loci for Chickpea (<italic>Cicer arietinum</italic> L.). PLoS One, v.6, n.11, p.e27275. 2011.

Tinker, N. A., A. Kilian, *et al*. New DArT markers for oat provide enhanced map coverage and global germplasm characterization. BMC Genomics, v.10, p.39. 2009.

Van Schalkwyk, A., P. Wenzl, *et al*. Bin mapping of tomato diversity array (DArT) markers to genomic regions of <i>Solanum lycopersicum</i> × <i>Solanum pennellii</i> introgression lines. TAG Theoretical and Applied Genetics, v.124, n.5, p.947-956. 2012.

Van Tassell, C. P., T. P. L. Smith, *et al*. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth, v.5, n.3, p.247-252. 2008.

Van Uum, C. M. J., S. J. C. Stevens, *et al*. SNP array-based copy number and genotype analyses for preimplantation genetic diagnosis of human unbalanced translocations. Eur J Hum Genet. 2012.

Verhaegen, D. e C. Plomion. Genetic mapping in *Eucalyptus urophylla* and *Eucalyptus grandis* using RAPD markers. Genome, v.39, p.1051-1061. 1996.

Vos, P., R. Hogers, *et al*. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research, v.23, n.21, January 1, 1995, p.4407-4414. 1995.

Weber, J. L. e P. E. May. Abundant class of human DNA polymorhisms which can be typed using the polymerare chain reaction. Am. J. Hum. Genet., v.44, p.388-396. 1989.

Welsh, J. e M. Mcclelland. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Research, v.18, n.24, January 1, 1990, p.7213-7218. 1990.

Wenzl, P., J. Carling, *et al*. Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci U S A, v.101, n.26, Jun 29, p.9915-20. 2004.

Wenzl, P., H. Li, *et al*. A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC Genomics, v.7, p.206. 2006.

Wheeler, M. A., M. Byrne, *et al*. Little genetic differentiation within the dominant forest tree, Eucalyptus marginata Myrtaceae) of South Western Australia. Silvae Genetica, v. 52, p. 254–259. 2003.

White, J., J. R. Law, *et al*. The genetic diversity of UK, US and Australian cultivars of Triticum aestivum measured by DArT markers and considered by genome. Theoretical and Applied Genetics, v.116, n.3, Feb, p.439-453. 2008.

Wiedmann, R. T., T. P. Smith, *et al*. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet, v.9, p.81. 2008.

Williams, J. G. K., A. R. Kubelik, *et al*. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research, v.18, n.22, November 25, 1990, p.6531-6535. 1990.

Willing, E. M., M. Hoffmann, *et al*. Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics, v.27, n.16, Aug 15, p.2187-2193. 2011.

Winter, A., W. Krämer, *et al*. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proceedings of the National Academy of Sciences, v.99, n.14, July 9, 2002, p.9300-9305. 2002.

Wittenberg, A., T. Lee, *et al*. Validation of the high-throughput marker technology DArT using the model plant Arabidopsis thaliana. Molecular Genetics and Genomics, v.274, p.30 - 39. 2005.

Wittenberg, A. H. Genetic mapping using the Diversity Arrays Technology (DArT) - Application and validation using the whole-genome sequences of Arabidopsis thaliana and the fungal wheat pathogen Mycosphaerella graminicola. Wageningen University, Netherlands, 2007.

Xia, L., K. Peng, *et al*. DArT for high-throughput genotyping of cassava (Manihot esculenta) and its wild relatives. TAG Theoretical and Applied Genetics, v.110, p.1092 - 1098. 2005.

Xie, Y., K. Mcnally, *et al*. A High-throughput Genomic Tool: Diversity Array Technology Complementary for Rice Genotyping. Journal of Integrative Plant Biology, v.48, n.9, p.1069-1076. 2006.

Yang, S. Y., R. K. Saxena, *et al*. The first genetic map of pigeon pea based on diversity arrays technology (DArT) markers. J Genet, v.90, n.1, Apr, p.103-9. 2011.

Zhang, L., D. Liu, *et al*. Investigation of genetic diversity and population structure of common wheat cultivars in northern China using DArT markers. BMC Genetics, v.12, n.1, p.42. 2011.

Zhivotovsky, L. A. Estimating population structure in diploids with multilocus dominant DNA markers. Molecular Ecology, v.8, n.6, p.907-913. 1999.

Zhu, Y. L., Q. J. Song, *et al*. Single-nucleotide polymorphisms in soybean. Genetics, v.163, n.3, Mar, p.1123-34. 2003.

Ziegle, J. S., Y. Su, *et al*. Application of automated DNA sizing technology for genotyping microsatellite loci. Genomics, v.14, n.4, Dec, p.1026-31. 1992.

6. ANEXOS

ANEXO I. Cópia do artigo publicado:

Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D, Vaillancourt RE, Kilian A (2010) A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. Plant Methods 6:16

ANEXO II. Cópia do artigo publicado:

Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J, Kilian A, Myburg AA, Grattapaglia D, Vaillancourt RE (2011) Population genetic analysis and phylogeny reconstruction in Eucalyptus (Myrtaceae) using high-throughput, genome-wide genotyping. Mol Phylogenet Evol 59:206-224

ANEXO III. Cópia do artigo publicado:

Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A (2011) Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proceedings 5:P54

PLANT METHODS

METHODOLOGY Open Access

A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in *Eucalyptus*

Carolina P Sansaloni[1,2], César D Petroli[1,2], Jason Carling[3], Corey J Hudson[4], Dorothy A Steane[4], Alexander A Myburg[5], Dario Grattapaglia[1,2,6], René E Vaillancourt*[4] and Andrzej Kilian[3]

Abstract

Background: A number of molecular marker technologies have allowed important advances in the understanding of the genetics and evolution of *Eucalyptus*, a genus that includes over 700 species, some of which are used worldwide in plantation forestry. Nevertheless, the average marker density achieved with current technologies remains at the level of a few hundred markers per population. Furthermore, the transferability of markers produced with most existing technology across species and pedigrees is usually very limited. High throughput, combined with wide genome coverage and high transferability are necessary to increase the resolution, speed and utility of molecular marker technology in eucalypts. We report the development of a high-density DArT genome profiling resource and demonstrate its potential for genome-wide diversity analysis and linkage mapping in several species of *Eucalyptus*.

Findings: After testing several genome complexity reduction methods we identified the *Pst*I/*Taq*I method as the most effective for *Eucalyptus* and developed 18 genomic libraries from *Pst*I/*Taq*I representations of 64 different *Eucalyptus* species. A total of 23,808 cloned DNA fragments were screened and 13,300 (56%) were found to be polymorphic among 284 individuals. After a redundancy analysis, 6,528 markers were selected for the operational array and these were supplemented with 1,152 additional clones taken from a library made from the *E. grandis* tree whose genome has been sequenced. Performance validation for diversity studies revealed 4,752 polymorphic markers among 174 individuals. Additionally, 5,013 markers showed segregation when screened using six inter-specific mapping pedigrees, with an average of 2,211 polymorphic markers per pedigree and a minimum of 859 polymorphic markers that were shared between any two pedigrees.

Conclusions: This operational DArT array will deliver 1,000-2,000 polymorphic markers for linkage mapping in most eucalypt pedigrees and thus provide high genome coverage. This array will also provide a high-throughput platform for population genetics and phylogenetics in *Eucalyptus*. The transferability of DArT across species and pedigrees is particularly valuable for a large genus such as *Eucalyptus* and will facilitate the transfer of information between different studies. Furthermore, the DArT marker array will provide a high-resolution link between phenotypes in populations and the *Eucalyptus* reference genome, which will soon be completed.

Background

A number of molecular marker technologies have been developed and used for species of *Eucalyptus* in the last 20 years [1]. Each of these technologies allowed important advances in the understanding of the multifaceted genetics, evolution and breeding of this vast genus that includes over 700 species, some of which are globally important plantation forestry species [2]. Molecular markers have been used to resolve phylogenetic issues [3], describe the genetic structure of natural populations [4,5], solve questions related to the management of genetic variation in breeding populations [6] and build linkage maps [7-9] that in turn have led to the identification of QTLs for important traits [10-13]. Nevertheless, the genotyping density achieved even with technologies such as AFLP [14] remains at a few hundred markers per

* Correspondence: R.Vaillancourt@utas.edu.au
[4] School of Plant Science and Cooperative Research Centre for Forestry, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia
Full list of author information is available at the end of the article

sample and because AFLP is gel-based it is relatively labour-intensive. Multiplexing has allowed moderate-level throughput in microsatellite studies. However, the transferability of microsatellites across species is notoriously poor and needs to be investigated and optimized before microsatellites can be used in a new species [1]. Wider genome coverage and higher throughput genotyping methods are necessary to increase resolution and speed for a variety of applications. Diversity Arrays Technology (DArT) [15] provides a promising alternative to satisfy the requirements of throughput, genome coverage and transferability. DArT is a complexity reduction, DNA hybridization-based method that simultaneously assays hundreds to thousands of markers across a genome. DArT preferentially targets low-copy genomic regions, allows automation of data acquisition and is cost competitive. Although developed some years ago, this marker technology has recently gained increasing attention [16-20]. We report the development of the first version of a high density operational DArT genotyping microarray with over 7,000 markers and demonstrate its potential for diversity and linkage mapping studies in species of *Eucalyptus* across the two most important subgenera.

Results and Discussion

This paper describes the various steps that were taken in developing the eucalypt DArT array (Figure 1). The first step was to find a successful method for reducing genome complexity. Once this was done, a prototype microarray was developed and tested. The DArT array was subsequently expanded and again tested for redundancy. The final step was to validate the operational microarray for genome-wide genotyping in *Eucalyptus*.

Genome complexity reduction

The first necessary step in the development of DArT markers (Figure 1) is choosing a genome complexity reduction method (see http://www.diversityarrays.com/molecularprincip.html). The DArT genome complexity reduction method is based on restriction enzyme (RE) digestion of total genomic DNA, adapter ligation and amplification of adapter-ligated fragments. DNA extraction was done with a CTAB protocol [21]. Seven methods of genome complexity reduction were tested for their performance in *Eucalyptus* (Additional File 1). DNA samples were prepared by digestion with the rare cutting *Pst*I RE as a primary cutter in combination with a frequently cutting enzyme (*Taq*I, *Bst*NI, *Msp*I, *Hpa*II, *Ban*II, *Mse*I or *Alu*I) as a secondary cutter. *Pst*I is sensitive to CpG methylation, thereby excluding heavily methylated repetitive DNA from the representation. Adapters, complementary to the "sticky-ends" of the fragments generated by *Pst*I digests were ligated (protocol modified slightly from the original [15,16]), to allow PCR amplification of only the *Pst*I fragments that had not been cut with the secondary enzyme. A desirable genome complexity reduction method will produce a smear of products with few or no distinct bands when representations are visualised on agarose gels following electrophoresis. Strong banding indicates the amplification of repetitive sequences and such representations are unsuitable for DArT development [22]. The genomic representations produced by the digestion with *Pst*I in combination with either *Taq*I (*Pst*I/

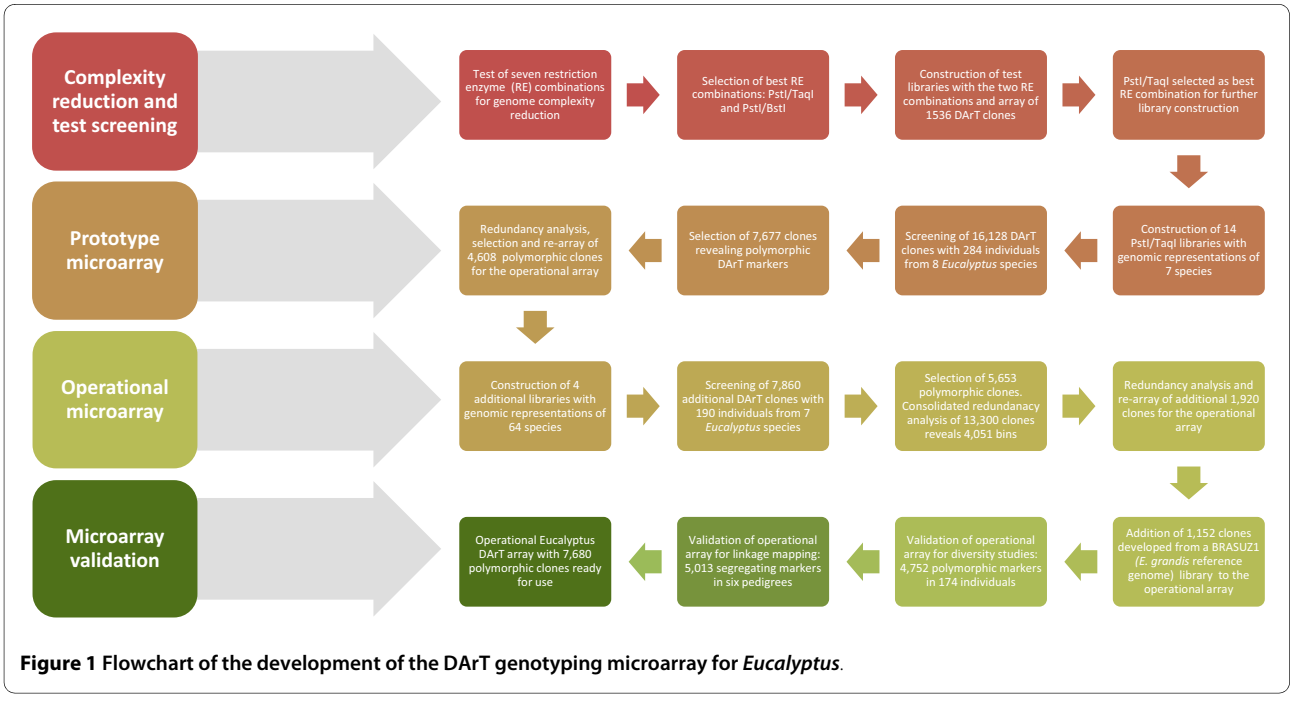


Figure 1 Flowchart of the development of the DArT genotyping microarray for *Eucalyptus*.

*Taq*I) or *Bst*NI (*Pst*I/*Bst*NI) were considered the most suitable for *Eucalyptus* to advance to the subsequent development steps (Figure 1, Additional File 1).

Test screening of clones for polymorphic DArT markers

The second step (Figure 1) entailed the construction of small genomic libraries for each of the selected complexity reduction methods and the screening of the resulting DNA clones (probes) to reveal polymorphic markers. For library construction, two sets of pooled DNA samples were utilized separately: the first from 12 *E. grandis* and the second from 12 *E. globulus* trees. Each pooled sample was digested with both enzyme combinations: *Pst*I/*Taq*I and *Pst*I/*Bst*NI. Four testing libraries were generated, each with 384 randomly picked clones, with a total of 1,536 DArT clones to be screened for polymorphism. The cloned DNA fragments were printed onto glass slides for the first test array in duplicates (randomly positioned within the array) as is normally done for DArT. Genomic representations of each of the 12 *E. grandis* and 12 *E. globulus* individuals were prepared to generate 'targets' that were hybridized to the arrays. For each species, the 12 genotypes were assayed with two technical replicates per genotype. Each target was labeled with a green fluorescent dye (Cy3-dUTP) and red fluorescent dye (Cy5-dUTP), and then mixed with a blue fluorescently-labeled polylinker from the vector used for cloning the DNA fragments in the libraries that provided a reference value for the quantity of amplified DNA fragment present in each 'spot' of the microarray, as well as an in-built quality control for spots on the microarrays. This mixture was hybridized to a 1,536-clone microarray, that was scanned for blue, green & red fluorescence and data were extracted using *DArTSoft* version 7.44. *DArTSoft* localizes the individual spot features of the microarrays and then compares the relative intensity (blue versus green) and (blue versus red) values obtained for each clone across all slides/targets to detect the presence of clusters of higher and lower values corresponding to marker scores of '1' (high) and '0' (low) respectively. The quality parameters used in this study were: Call Rate (percentage of targets that could be scored as '0' or '1') and Reproducibility value (reproducibility of scoring between replicated target assays) [16]. The results of the *DArTSoft* analysis for the two arrays prepared using DNA clones derived from either *Pst*I/*Taq*I or *Pst*I/*Bst*NI digestions were compared with regard to the frequency of clones revealing polymorphic DArT markers. The criteria used for declaring a clone as revealing a polymorphic marker were Reproducibility > 97% and Call Rate > 80%. From the analysis of the two species hybridized in duplicate to the two arrays, the complexity reduction method using *Pst*I/*Taq*I was found to yield a higher proportion (21.7%) of candidate polymorphic markers according to the above criteria compared to the *Pst*I/*Bst*NI method (14.3%).

Prototype Eucalyptus *DArT microarray*

The *Pst*I/*Taq*I genome complexity reduction method was used in the development of the prototype *Eucalyptus* DArT microarray (Figure 1). The initial test array, with 1,536 clones, was expanded by picking an additional random set of 14,592 discovery DArT clones, this time derived from a total of 14 libraries (Table 1). A broader sample of genotypes (254 samples from seven eucalypt species representing the two most important genera of eucalypts [*Corymbia* and *Eucalyptus*] and the two most important subgenera of *Eucalyptus* [*Eucalyptus* and *Symphyomyrtus*]) were used for library construction resulting in a broader sample of DNA sequences, therefore increasing the probability of sampling genomic segments that could reveal polymorphic markers across a wider range of genetic backgrounds [16]. A total of 16,128 clones were printed twice on each slide and were hybridized with DNA from each of 284 individuals ("targets"; Table 2) representing eight different species with replication, following the methods described above. The results were analyzed with *DArTSoft* and assessed using the threshold criteria of Reproducibility > 97% and Call Rate > 80%. This analysis revealed 7,677 clones (47.6%) as robust polymorphic markers (Table 3). The Call Rate average was 95.3% and the Reproducibility average was 99.7% (this value was calculated on the basis of duplicate genotyping assays for all test samples).

Testing *Corymbia* targets on the array composed primarily of *Eucalyptus* probes (and *vice versa*) showed very clearly that the overall array signal of *Corymbia* targets was low and uncorrelated to signal from *Eucalyptus* species (and *vice versa*). Because of this poor transferability across genera, we abandoned the development of DArT for *Corymbia*. As clones used to build an array are picked at random from the libraries, clone redundancy (i.e. DNA fragments with the same or very similar/overlapping sequence) is an issue. Redundancy of the polymorphic DArT clones was evaluated with the software package *DArT ToolBox* http://www.diversityarrays.com/ by constructing a Hamming distance matrix between clones, followed by distance binning, in which all clones with zero distance were placed into the same bin. This was done using the 284 samples used as targets listed in Table 2. This estimation of clone redundancy based on similar score pattern enabled the selection of unique or low redundancy clones prior to the availability of sequence information for the clones. The redundancy estimation based on distance binning of the 7,677 polymorphic markers resulted in 2,652 unique bins, i.e. 34.5% non-redundant marker scoring patterns (Table 4). With a limited number of effective scores for calculating the dis-

Table 1: Libraries and corresponding numbers of clones screened for the prototype *Eucalyptus* DArT microarray

No. of clones	No. of individuals	Species*	Source**
768	12	*Corymbia variegata*	Australia
768	11	*E. camaldulensis*	Australia
768	13	*E. globulus*	Portugal/Chile
1920	12	*E. globulus*	Australia
1536	24	*E. globulus*	Australia
768	12	*E. globulus*	Australia
1536	96	*E. grandis* × *E. urophylla*	Brazil
1536	12	*E. grandis*	South Africa
2688	9	*E. grandis*	South Africa
768	6	*E. nitens*	Chile
768	11	*E. nitens*	Australia
768	12	*E. pilularis*	Australia
768	12	*E. urophylla*	South Africa
768	12	*E. urophylla*	South Africa

* *Corymbia variegata* belongs to a genus phylogenetically closely related to *Eucalyptus*; *E. camaldulensis, E. globulus, E. grandis, E. urophylla* and *E. nitens* belong to *Eucalyptus* subgenus *Symphyomyrtus*; *E. pilularis* belongs to *Eucalyptus* subgenus *Eucalyptus*.

** Sourced from native stands or first-generation breeding populations established from seed stocks derived from native stands.

tance matrix for markers and a clear genetic structure in the materials used for initial marker discovery, there was a high likelihood of unique sequences being grouped to a single bin, especially in large bins. Therefore, a total of 4,608 clones were selected for re-arraying, keeping approximately 30% of the potentially redundant markers, with frequency of retention proportional to the bin size. In order to verify the redundancy estimation, we sequenced re-arrayed clones that belonged to nine bins that had at least 30 clones. Sequencing results revealed that on average 53% of the DArT clones in these large bins represented unique DNA sequences. Binning results were therefore, as anticipated, conservative and yielded an overestimation of redundancy (Table 5).

Table 2: *Eucalyptus* species and number of individuals of each species (total of 284) used as targets to screen the prototype DArT microarray for polymorphic markers

No. of individuals	Species
135	*E. grandis* × *E. urophylla*
28	*E. pilularis*
27	*E. nitens*
35	*E. globulus*
12	*E. cladocalyx*
12	*E. grandis*
12	*E. urophylla*
12	*Corymbia variegata*
11	*E. camaldulensis*

Interim and operational Eucalyptus DArT microarrays

In order to enrich the *Eucalyptus* DArT array for polymorphic markers, four additional genomic libraries were constructed that provided a total of 7,680 new clones that were screened for polymorphism (Table 6). Two of these libraries contained DNA from 62 eucalypt species and were built by pooling equimolar DNA quantities from one individual of each species and cutting either with *Pst*I or *Pst*I/*Taq*I. The *Pst*I representation allowed markers that were present at low frequency in the *Pst*I/*Taq*I representation to be cloned and therefore minimized redundancy in the final clone set. Most species (56) were from subgenus *Symphyomyrtus* (representing 14 of the 15 sections and missing only a minor one); the other species were from three other subgenera (*Alveolata, Eucalyptus*, and *Minutifructus*). Screening these new libraries for polymorphism (Figure 1) was performed using a set of 190 individuals from seven different *Eucalyptus* species (*E. grandis, E. urophylla, E. camaldulensis, E. globulus, E. dunni, E. pilularis and E. nitens*) with targets in full replication (Table 7). *DArTSoft* and *DArT ToolBox* were used to identify robust markers and to estimate redundancy as described for the first array (with the same parameters and thresholds). *DArTSoft* detected 5,653 polymorphic markers among the 7,680 clones (73.6%). The average Call Rate and Reproducibility were similar to the first array with 93.7% and 99.7% respectively. However, a sig-

Table 3: Summary of results of the *Eucalyptus* DArT microarray development involving screening for polymorphism and score signature-based redundancy analysis in the prototype and operational arrays (see text for details; n.d. not determined)

Technology development phase	No. of DArT clones screened	No. and (%) of polymorphic DArT clones	No. of bins with unique scoring pattern
Initial libraries	16,128	7,677 (47.6%)	2,652 (16.4%)
Array expansion libraries	7,680	5,653 (73.6%)	n.d.
All libraries	23,808	13,300 (55.9%)	4,051 (17.0%)

nificantly higher percentage of polymorphic markers (73.6% versus 47.6%) was found in the array expansion stage (Table 3), most likely due to the greater genetic diversity that was captured in the genomic representations from the four new libraries. A consolidated analysis of redundancy based on binning was carried out to minimize redundancy between the 7,677 polymorphic clones selected initially for the prototype microarray and the additional 5,653 clones. From a total of 13,300 clones, 4,051 bins were found in the interim array (Tables 3 and 4). On the basis of polymorphism analysis and the additional redundancy assessment, 1,920 new clones were selected from the 7,680, to create a second re-arrayed library. The two re-arrayed libraries (the first one with 4,608 clones and the second with 1,920 clones), were supplemented with 1,152 clones developed primarily from a genomic library of BRASUZ1, the *Eucalyptus grandis* tree whose genome is being sequenced (Table 6), to constitute an operational DArT genotyping array for *Eucalyptus* with 7,680 markers.

Validation of DArT array for diversity and linkage mapping

The performance of the operational DArT array for diversity studies was first validated by genotyping 174 individuals from six of the *Eucalyptus* species used to create the libraries (*E. grandis, E. urophylla, E. dunni, E. camaldulensis, E. globulus* and *E. nitens*). These individuals were a subset of those used to create the libraries. This analysis revealed 4,752 polymorphic markers out of the 7,680 clones (61.9%) among the 174 individuals. As expected, not all the 7,680 clones were found to yield polymorphic markers since the 174 samples assayed did not represent the total genetic diversity used to construct the array.

As a second validation, an assessment of DArT marker segregation and rate of polymorphism was carried-out with 94 samples in full replication, including 15-16 samples from each of six mapping pedigrees. Most of these individuals were not used in library construction and represented a test of the level of polymorphism that could be expected in diverse linkage mapping experiments. There were 2,211 polymorphic markers per pedigree on average (Table 8). The number of shared polymorphic markers (polymorphic in two pedigrees) amongst the six mapping pedigrees varied from a minimum of 859 to a maximum of 1,328 (Table 8). A total of 5,013 markers (65.3%) out of the 7,680 clones showed segregation within at least one mapping population, when data from the six pedigrees were consolidated (Table 9). Table 9 shows the number of DArT markers that were exclusively polymorphic in one pedigree only (1,154 markers or 23%) through to those that were polymorphic in an increasing number of pedigrees up to all six pedigrees (150 markers: 3%).

Table 4: Distribution of the number of polymorphic DArT clones within each binning class in the prototype and interim phases of the DArT microarray development

No. of clones in bin	No. of bins in prototype array *(7,677 polymorphic clones)*	No. of bins in interim array *(13,300 polymorphic clones)*
1	1,330	2,143
2-9	1,199	1,737
10-19	105	126
20-29	9	17
30-39	4	8
40-49	2	3
≥ 50	3	17
Total	2,652	4,051

Table 5: Results of DArT clone redundancy analysis based on DNA sequencing of clones selected from the nine bins that had at least 30 clones per bin, based on Hamming distance of zero (no difference in scoring pattern between markers in the bin)

Bin #	No. of clones per bin based on Hamming distance	No. of clones selected for re-arraying and sequencing	No. of re-arrayed clones with unique DNA sequences	% of re-arrayed clones with unique DNA sequences
1	116	33	19	57.6
2	75	20	12	60
3	59	17	8	47.1
4	43	13	6	46.2
5	41	6	4	66.7
6	39	10	5	50
7	37	16	11	68.8
8	31	10	6	60
9	30	9	2	22.2
Average	52.3	14.8	8.1	53.2

Conclusions

This eucalypt DArT array is one of the best performing DArT arrays yet developed (DArT Pty Ltd, unpublished results). The high frequency of polymorphic markers is likely to be a function of the high level of sequence variation in the *Eucalyptus* genome [23] and, to a much lesser extent, a function of its relatively small genome size and low proportion of repetitive DNA [1]. Interestingly, the high level of sequence diversity in *Eucalyptus* species [23] could be a serious impediment to the development of highly multiplexed SNP platforms that usually require reasonably long stretches of sequence without secondary SNPs. It may prove challenging to find good targets for SNP assay design which would be invariable across a range of *Eucalyptus* species. In this context, DArT analysis is not constrained by high sequence polymorphism and is therefore very suitable for genotyping thousands of

Table 6: Four additional genomic representation libraries and corresponding numbers of clones used for the development of the interim and operational DArT microarray

No. of clones in library	No. of individuals	Species	Source	RE digestion
1920 *	16	*E. grandis* × *E. urophylla (IP pedigree)*	Brazil	*Pst*I/*Taq*I
	16	*E. grandis* × *E. urophylla (VCP pedigree)*	Brazil	*Pst*I/*Taq*I
	16	*E. camaldulensis* × *(E. urophylla* × *E. globulus)*	Brazil	*Pst*I/*Taq*I
	16	*(E. grandis* × *E. urophylla)* × *(E. urophylla* × *E. globulus)*	Brazil	*Pst*I/*Taq*I
	16	*(E. dunnii* × *E. grandis)* × *(E. urophylla* × *E. globulus)*	Brazil	*Pst*I/*Taq*I
	16	*(E. dunnii* × *E. grandis)* × *E. urophylla*	Brazil	*Pst*I/*Taq*I
1152 **	1	*E. grandis (BRASUZ1)*	Brazil	*Pst*I/*Taq*I
2304 ***	62	Several species	Australia	*Pst*I/*Taq*I
2304 ***	62	Several species	Australia	*Pst*I

* Library built by pooling equimolar quantities of DNA of 96 inter-specific hybrids.

** Library built with DNA of *E. grandis* tree BRASUZ1, whose full genome is being sequenced.

*** The following species were used: *E. albens, E. balladoniensis, E. bicostata, E. biterranea, E. brassiana, E. brevistylis, E. camaldulensis, E. cladocalyx, E. coolabah, E. cordata, E. cornuta, E. cosmophylla, E. crebra, E. dalrympleana, E. deglupta, E. delicata, E. diversicolor, E. dundasii, E. dunnii, E. falcata, E. glaucescens, E. glaucina, E. globulus, E. gomphocephala, E. grandis, E. gunnii, E. hallii, E. houseana, E. howittiana, E. leucophloia, E. lockyeri, E. longifolia, E. lucasii, E. maidenii, E. michaeliana, E. microcorys, E. morrisbyi, E. nitens, E. obtusiflora, E. optima, E. ovata, E. pachycalyx, E. pachyphylla, E. paludicola, E. perriniana, E. platyphylla, E. polyanthemos, E. populnea, E. pseudoglobulus, E. pulverulenta, E. pumila, E. ravereana, E. rubida, E. salmonophloia, E. scoparia, E. stoatei, E. tereticornis, E. torquata, E. urophylla, E. viminalis, E. wandoo, E. woodwardii*.

Table 7: *Eucalyptus* pedigrees and corresponding numbers of individuals used as targets to screen the 7,680 clones for degree of polymorphism

No. of individuals	Species
71	*E. grandis* × *E. urophylla*
16	*E. camaldulensis* × (*E. urophylla* × *E. globulus*)
16	(*E. grandis* × *E. urophylla*) × (*E. urophylla* × *E. globulus*)
16	(*E. dunnii* × *E. grandis*) × (*E. urophylla* × *E. globulus*)
16	(*E. dunnii* × *E. grandis*) × *E. urophylla*
16	*E. pilularis*
16	*E. nitens*
23	*E. globulus*

genetic markers in highly outbred organisms such as *Eucalyptus*.

DArT generated a substantially larger number of robust polymorphic markers for *Eucalyptus* species than previous technologies. Although co-dominant microsatellites are significantly more informative at the single locus level they are low-throughput and expensive per data-point. Comparing DArT with RAPD or AFLP analysis would be more appropriate as they are all dominant markers. The complicating issue, however, is the ascertainment bias that takes place when selecting RAPD primers, AFLP primer/enzyme combinations or DArT polymorphic probes. This bias is exacerbated by the specific target population that is used when selecting polymorphisms and by the rigor of the experimenter in declaring these polymorphisms. It is important to note that the DArT array developed in this study provides at least two orders of magnitude more polymorphic markers in a single assay than RAPD or AFLP analysis. In *Eucalyptus*, while a selected RAPD primer can provide up to 10 robust polymorphic bands in a single gel run and a selected AFLP combination can provide on average 30 to 40 polymorphic markers, a single DArT assay provides 1,000 to 4,000 polymorphic markers from the7,680 probes present on the current array. In addition, the standard probe set selected for routine DArT genotyping allows comparisons of markers across a range of species and populations while both AFLP and RAPD markers are much less amenable to integration across laboratories and even less so across different species.

The high level of DArT marker multiplexing was validated in a large collection of eucalypt species and individuals. The results indicated that the DArT genotyping array will deliver thousands of polymorphic markers for population diversity studies and provide a very efficient platform with which to generate high-density linkage maps with a substantial proportion of markers shared across pedigrees. This array will be especially useful for applications that benefit from access to a large number of markers. The cost per data point (per sample per marker) will of course depend on the application and the facility generating the data. Using the fully costed service provided by the technology development partner, DArT Pty Ltd, the cost per data point for polymorphic markers is expected to vary between one and five cents US (assuming an assay cost per sample of 50 USD, not counting shipping and DNA extraction costs). In linkage mapping studies, an application where one of the lowest degrees of polymorphism is expected because diversity comes essentially from only two parents, we expect that a minimum of 1,000 polymorphic markers could be mapped at a cost of approximately five cents US per polymorphic marker. The per sample cost is much cheaper than current SNP genotyping platforms assaying an equivalent number of markers (e.g. Illumina GoldenGate). The in-house use of DArT arrays would involve purchasing the equipment necessary to spot high density arrays, hybrid-

Table 8: Number of polymorphic DArT markers in each *Eucalyptus* mapping pedigree (diagonal) and shared among mapping pedigrees (above the diagonal)

	C1 × UGI	DG × U	DG × UGI	G × U(IP)	UGI × GU	G × U(VCP)
C1 × UGI	2394	864	1328	1123	1172	899
DG × U		1818	1154	1029	866	859
DG × UGI			2465	1251	1284	953
G × U(IP)				2553	1175	1144
UGI × GU					2176	946
G × U(VCP)						1861

C1 × UGI = *E. camaldulensis* × (*E. urophylla* × *E. globulus*); DG × U = (*E. dunnii* × *E. grandis*) × *E. urophylla*; DG × UGI = (*E. dunnii* × *E. grandis*) × (*E. urophylla* × *E. globulus*); G × U(IP) = *E. grandis* × *E. urophylla* (pedigree IP); UGIxGU = (*E. urophylla* × *E. globulus*) × (*E.grandis* × *E. urophylla*); G × U(VCP) = *E. grandis* × *E. urophylla* (VCP pedigree).

Table 9: Informativeness of DArT markers from the operational array for genetic mapping based on sampling six different pedigrees (see Table 8 for list of pedigrees)

No. of mapping pedigrees in which a marker was polymorphic	No. of DArT markers in the class	% of total number of polymorphic markers
1	1,407	28.1
2	1,154	23.0
3	1,048	21.0
4	761	15.2
5	493	9.8
6	150	3.0
Total	5,013	

ization chambers and a multi color scanner and therefore would require a very high throughput operation to make such investment worthwhile.

Another significant advantage of the DArT markers is their transferability across species, which is particularly valuable when dealing with a genus like *Eucalyptus* with over 700 species, of which many are foundation species in their forest ecosystems, and several are commercially useful in either temperate or sub-tropical regions of the world. This transferability will allow the detailed comparison of linkage maps and QTL positions across studies. However, this transferability appears to have limits as we obtained poor transferability across eucalypt genera (*Corymbia* to *Eucalyptus*). We will address the phylogenetic consequences of this finding and the performance of the DArT array across the full range of *Eucalyptus* species in a related study (Steane *et al.* submitted).

A limitation of the DArT technology compared to multi allelic microsatellites is their dominant inheritance, which precludes studying aspects of within-individual variation, although methodologies are being developed that can mitigate this [24]. Dominant markers are also less informative for constructing linkage maps, unless a large number of them are available and population sizes are large, in which case they can be as useful as co-dominant markers. Finally, clustering of DArT markers across the genome could potentially be an issue due to the reduced representation method by which that DArT probes are developed. However, this is not exclusive to the DArT technology and an assessment of this will only be possible by linkage mapping DArT markers in multiple pedigrees and/or physically mapping them on the upcoming *Eucalyptus* reference genome.

To better characterize the genomic content of this array, all 7,680 DNA clones on the operational DArT array are being sequenced. The availability of DNA sequences for the DArT markers will facilitate the integration of high-density maps and QTL locations with the *Eucalyptus* genome assembly. The operational DArT array constitutes a powerful tool with which to undertake high resolution genetic analyses required for applications such as fine QTL mapping, genome-wide selection and complex phylogenetic and evolutionary investigations. Moreover, the flexibility and expandability of the DArT technology opens the possibility of further enriching the current array with additional polymorphic markers by simply screening additional sets of clones. A number of mapping (Grattapaglia *et al.* in prep; Kullan *et al.* in prep), population and phylogenetic (Steane *et al.* submitted) studies currently underway with DArT in several *Eucalyptus* species are corroborating the excellent performance of this technology and will be the subject of upcoming reports.

Methods

For the development of the *Eucalyptus* DArT microarray, DNA samples from many different species and provenances were used both in the prototype and technology development steps (Tables 1, 2, 6 and 7). DNA was extracted from either fresh leaf tissue or bark cambium in three different laboratories (Australia, South Africa, Brazil) all using a CTAB protocol [21]. DNA quality was checked on agarose gels with DNA digested with the restriction enzyme *Hind*III together with undigested DNA to check that (1) undigested DNA formed a tight band of high molecular weight without RNA contamination; (2) fully-digested DNA formed a smear of mid- to low molecular weight. DNA concentration was adjusted to 50-100 ng/μL, targeting a concentration of 75 ng/μL.

Methods of genome complexity reduction to generate genomic representations

Digestion and adapter ligation were performed simultaneously on 75 ng of genomic DNA in a 10 μL aqueous solution containing 2 Units of each restriction enzyme, 80 Units of T4 DNA Ligase and 0.05 μM adapter (5'-CAC GAT GGA TCC AGT GCA-3' annealed with 5'-CTG GAT CCA TCG TGC A-3'). Reactions were incubated at

37°C for 2 hours, followed by 2 hours at 60°C as required by the enzyme combinations. 1 μL of digestion/ligation reaction product was used as a template for PCR amplification in a 50 μL reaction using DArT *Pst*I primer (5'-GAT GGA TCC AGT GCA G-3') with the following cycling parameters: 94°C for 1 min, followed by 30 cycles of 94°C for 20 sec, 58°C for 40 sec, 72°C for 1 min, and finished with an extension at 72°C for 7 min. Initial assessment of the tested methods was performed by resolving 5 μL of amplification product in a 1.2% agarose gel stained with ethidium bromide.

Construction of small clone DArT libraries

The genomic representations of each species/complexity reduction method combination were pooled and cloned using the *TOPO TA Cloning Kit* (Invitrogen) as specified by the manufacturer's instructions. Individual bacterial colonies were picked into 384-well plates containing LB medium with 4.4% glycerol, 100 μg/mL ampicillin and a mixture of salts to facilitate PCR from the LB cultures (unpublished observation) and grown at 37° for 18 hours. A PCR amplification was performed using 0.5 μL of bacterial culture template, 0.2 μM "M13 Forward" and "M13 Reverse" primers (Invitrogen), and the following PCR program: 95° for 4 min, 57° for 35 sec, 72° for 1 min, followed by 35 cycles of 94° for 35 sec, 52° for 35 sec and 72° for 1 min and a final step of 72° for 7 min. The PCR products were dried at 37°C and washed with 70% ethanol before being dissolved in "DArTspotter" spotting buffer, designed for use with poly-L-lysine coated micro-array slides (Wenzl *et. al.* in preparation, available from DArT Pty Ltd). Arrays were spotted using a *MicrogridII* arrayer (Biorobotics) on poly-L-lysine coated glass microarray slides (Erie Scientific). Slides were aged on the bench for 24 hours before being immersed in Milli-Q water at 95°C for 2 min, to denature the DNA spotted onto the slides, then in Milli-Q water with 0.1 mM DTT and 0.1 mM EDTA at 20°C, and finally being dried by centrifugation at 500 × g for 7 min and vacuum desiccation for 30 min.

Fluorescent labeling of genomic representations

Genomic representations of the 12 samples of *E. grandis* and *E. globulus* were prepared as described above for library construction, to generate 'targets' for hybridizing to the arrays. The products of amplification were precipitated individually with isopropanol, washed with 70% ethanol and air dried at room temperature for 12 hours. For each species the 12 genotypes were assayed with two replicates per genotype. Targets were labeled in a 10 μL reaction volume with 2.5 nM of Cy3-dUTP or Cy5-dUTP (Amersham Bioscience), 2.5 units of Klenow exo$^-$ fragment of *E. coli* Polymerase I (New England Biolabs) and 25 μM random decamers in 1 × NEB Buffer 2 (New England Biolabs). The labelling reactions were incubated at 37°C for 3 hours.

Test hybridization to microarrays

The labeled targets were mixed with a hybridisation buffer containing a 50:5:1 mixture of Express Hyb (Clonetech), herring sperm DNA (Promega) and FAM-labeled polylinker region of the pCR 2.1 TOPO vector (Invitrogen) used for cloning the libraries, plus 2 mM EDTA at pH 8.0. The target mixtures were denatured at 95°C for 2 min before hybridization to the microarrays, which was carried out at 62.5°C for 18 hours. After hybridization, the microarray slides were washed in four solutions of increasing stringency (1 × SSC, 0.1% SDS for 4 min; 1 × SSC for 4 min; 0.2 × SSC for 1 min; 0.02 × SSC for 30 sec) and dried by centrifugation at 500 × g for 7 min and vacuum desiccation for 30 min.

Microarray imaging and data extraction

Microarrays were scanned using a TECAN LS300 confocal laser microarray scanner at a resolution of 20 μm per pixel with sequential acquisition of 3 images for each microarray slide, using the following laser/emission-filter combinations: 488 nm laser/520 nm filter (for imaging the fluorescent signal from the FAM-labeled polylinker region of the pCR 2.1 TOPO vector); 543 nm laser/590 nm filter (for imaging the fluorescent signal from the hybridized target labeled with Cy-3); 633 nm laser/670 nm filter (for imaging the fluorescent signal from the hybridized target labeled with Cy-5). The use of a third fluorescent dye is not absolutely required and DArT assays can be performed on any two-color scanner as reported in early DArT papers. However, the third dye provides significantly higher sample throughput together with lower assay cost because two samples can be processed on a single array instead of just one as is the case when using a two-color scanner. The signal from the FAM-labeled vector polylinker provided a reference value for quantity of amplified DNA fragment present in each 'spot' of the microarray. The resulting images were analyzed using *DArTSoft* version 7.44, a program created by *Diversity Arrays Technology Pty Ltd* for microarray image data extraction, polymorphism detection, and marker scoring (Cayla *et al*. in preparation). *DArTsoft* localized the individual spot features of the microarrays from the 16 bit TIFF images generated by the laser scanner and spots with insufficient or absent reference signals were rejected from further analysis. A relative hybridisation intensity value was then calculated for all accepted spots as log [Cy-3 signal/FAM signal] for the targets labelled with Cy-3, and log [Cy-5 signal/FAM signal] for targets labelled with Cy-5. *DArTSoft* then compared the relative intensity values obtained for each clone across all slides/targets to detect the presence of clusters of higher and lower values corresponding to marker scores of '1' and '0' respectively. Targets with relative intensity values that could not be assigned to either of the clusters were recorded as unscored. For each clone, the software gener-

ated a range of quality parameters to assist in selection of polymorphic clones. The quality parameters used in this study were: Call Rate (percentage of targets that could be scored as '0' or '1') and a Reproducibility value (reproducibility of scoring between replicated target assays). Two replicates per clone were spotted on each array. The operational array has 15,360 spots in total, comprising two randomly positioned spots for each one of the 7,680 clones. The DArT array is available to the public through Diversity Arrays Technology Pty Ltd http://www.diversitarrays.com/.

Additional material

Additional file 1 Genome complexity reduction with seven restriction enzymes. Results of the seven restriction enzyme combinations tested for genome complexity reduction in *Eucalyptus grandis* and *Eucalyptus globulus*. **Top panel**: Gel photo showing the digestion of the same pooled DNA sample of *E. grandis* and *E. globulus* with different restriction enzyme combinations: 2-3 *Pst*I(*Taq*I), 4-5 *Pst*I(*Bst*NI), 6-7 *Pst*I(*Msp*I), 8-9 *Pst*I(*Hpa*II), 10-11 *Pst*I(*Ban*II), 12-13 *Pst*I(*Mse*I), 14-15 *Pst*I(*Alu*I). **Bottom panel**: Fluorescence intensity profile of the digested *E. grandis* DNA obtained with each complexity reduction method. The molecular sizing standard (100 bp ladder) is showed in red; track 2 (dark blue *Pst*I/*Taq*I) with a smoother profile was selected as the best complexity reduction method.

Competing interests

JC and AK are employees of Diversity Arrays Technology Pty Ltd which offers genome profiling service with the product of this report and therefore can potentially benefit from this work.

Authors' contributions

CPS, CDP, DAS and JC performed the laboratory work, and most data analysis and interpretation; DAS and CJH participated in initial library constructions; AK, DAS, REV and AAM contributed to the design of the study; AK supervised the study, participated in data analysis and interpretation and edited the manuscript. CPS, CDP and JC drafted the initial version of the manuscript; DG and REV substantially edited the manuscript and participated in data interpretation, analysis and organization. All authors read, edited and approved the final manuscript.

Acknowledgements

The study was supported by the following: (1) Australian Research Council (DP0770506); (2) Brazilian Ministry of Science and Technology (CNPq Project 474645/2007-9); (3) Mondi and Sappi through the Wood and Fiber Molecular Genetics Programme (University of Pretoria, South Africa); (4) CRC for Forestry (Australia); (5) Forestal Mininco S.A. (Chile); (6) Oji Paper; (7) L'Institut National de la Recherche Agronomique (INRA, France). We acknowledge the help and technical support from many employees of Diversity Arrays Technology Pty Ltd where the work was performed. We also express thanks to Tim Sexton, Dean Nicolle, Cristina Marques, Dean Williams and Kelsey Joyce for supplying tissue and/or DNA samples. CPS and CDP hold doctoral fellowships from CAPES (Brazilian Ministry of Education), CJH a Cuthbertson Tasmania Graduate Research Scholarship and DG a productivity research fellowship from CNPq.

Author Details

[1]Plant Genetics Laboratory, EMBRAPA Genetic Resources and Biotechnology - EPqB, 70770-910 Brasilia, Brazil, [2]Dep. Cell Biology, Universidade de Brasilia - 70910-900 Brasília - DF, Brazil, [3]Diversity Arrays Technology Pty Ltd, 1 Wilf Crane Crescent, Yarralumla, ACT 2600, Australia, [4]School of Plant Science and Cooperative Research Centre for Forestry, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia, [5]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa and [6]Genomic Sciences Program - Universidade Católica de Brasília - SGAN, 916 modulo B, 70790-160 Brasília - DF, Brazil

Received: 16 April 2010 Accepted: 30 June 2010
Published: 30 June 2010

References

1. Grattapaglia D, Kirst M: ***Eucalyptus*** **applied genomics: from gene sequences to breeding tools.** *New Phytologist* 2008, **179:**911-929.
2. Myburg AA, Potts BM, Marques CM, Kirst M, Gion JM, Grattapaglia D, Grima-Pettenati J: **Eucalyptus.** In *Genome mapping and molecular breeding in plants Volume 7*. Edited by: C K. Forest trees. New York, NY, USA: Springer; 2007:115-160.
3. Steane DA, Nicolle D, Vaillancourt RE, Potts BM: **Higher-level relationships among the eucalypts are resolved by ITS-sequence data.** *Australian Systematic Botany* 2002, **15:**49-62.
4. Steane D, Conod N, Jones R, Vaillancourt R, Potts B: **A comparative analysis of population structure of a forest tree, *Eucalyptus globulus* (Myrtaceae), using microsatellite markers and quantitative traits.** *Tree Genetics & Genomes* 2006, **2:**30-38.
5. Payn KG, Dvorak WS, Janse BJH, Myburg AA: **Microsatellite diversity and genetic structure of the commercially important tropical tree species *Eucalyptus urophylla*, endemic to seven islands in eastern Indonesia.** *Tree Genetics & Genomes* 2008, **4:**519-530.
6. Grattapaglia D, Ribeiro VJ, Rezende GD: **Retrospective selection of elite parent trees using paternity testing with microsatellite markers: an alternative short term breeding tactic for *Eucalyptus*.** *Theor Appl Genet* 2004, **109:**192-199.
7. Byrne M, Murrell JC, Allen B, Moran GF: **An integrated genetic linkage map for eucalypts using RFLP, RAPD and isozyme markers.** *Theoretical and Applied Genetics* 1995, **91:**869-875.
8. Brondani R, Williams E, Brondani C, Grattapaglia D: **A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus.** *BMC Plant Biology* 2006, **6:**20.
9. Thamarus K, Groom K, Murrell J, Byrne M, Moran G: **A genetic linkage map for *Eucalyptus globulus* with candidate loci for wood, fibre and floral traits.** *Theor Appl Genet* 2002, **104:**379-387.
10. Grattapaglia D, Bertolucci FL, Penchel R, Sederoff RR: **Genetic mapping of quantitative trait loci controlling growth and wood quality traits in *Eucalyptus grandis* using a maternal half-sib family and RAPD markers.** *Genetics* 1996, **144:**1205-1214.
11. Freeman JS, Whittock SP, Potts BM, Vaillancourt RE: **QTL influencing growth and wood properties in *Eucalyptus globulus*.** *Tree Genetics & Genomes* 2009, **5:**713-722.
12. Junghans DT, Alfenas AC, Brommonschenkel SH, Oda S, Mello EJ, Grattapaglia D: **Resistance to rust (*Puccinia psidii* Winter) in *Eucalyptus*: mode of inheritance and mapping of a major gene with RAPD markers.** *Theor Appl Genet* 2003, **108:**175-180.
13. Thamarus K, Groom K, Bradley A, Raymond CA, Schimleck LR, Williams ER, Moran GF: **Identification of quantitative trait loci for wood and fibre properties in two full-sib properties of *Eucalyptus globulus*.** *Theor Appl Genet* 2004, **109:**856-864.
14. Myburg AA, Griffin AR, Sederoff RR, Whetten RW: **Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F_1 hybrid based on a double pseudo-backcross mapping approach.** *Theor Appl Genet* 2003, **107:**1028-1042.
15. Jaccoud D, Peng K, Feinstein D, Kilian A: **Diversity arrays: a solid state technology for sequence information independent genotyping.** *Nucleic Acids Res* 2001, **29(4):**E25.
16. Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A: **Diversity Arrays Technology (DArT) for whole-genome profiling of barley.** *Proc Natl Acad Sci USA* 2004, **101:**9915-9920.
17. Wittenberg A, Lee T, Cayla C, Kilian A, Visser R, Schouten H: **Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*.** *Molecular Genetics and Genomics* 2005, **274:**30-39.
18. Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang S, Uszynski G, Mohler V, Lehmensiek A, Kuchel H, *et al.*: **Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome.** *TAG Theoretical and Applied Genetics* 2006, **113:**1409-1420.
19. Xia L, Peng K, Yang S, Wenzl P, Carmen de Vicente M, Fregene M, Kilian A: **DArT for high-throughput genotyping of cassava (*Manihot esculenta*) and its wild relatives.** *TAG Theoretical and Applied Genetics* 2005, **110:**1092-1098.
20. Tinker NA, Kilian A, Wight CP, Heller-Uszynska K, Wenzl P, Rines HW, Bjornstad A, Howarth CJ, Jannink JL, Anderson JM, *et al.*: **New DArT**

markers for oat provide enhanced map coverage and global germplasm characterization. *BMC Genomics* 2009, **10**:39.
21. Doyle JJ, Doyle JL: **Isolation of plant DNA from fresh tissue.** *Focus* **12**:13-15.
22. Kilian A, Huttner E, Wenzl P, Jaccoud D, Carling J, Caig V, Evers M, Heller-Uszynska K, Cayla C, Patarapuwadol S, *et al.*: **The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement.** In *International Congress In the Wake of the Double Helix: From the Green Revolution to the Gene Revolution: May 27-31 2003 Volume 2003*. Bologna, Italy: Avenue Media; 2005:443-461.
23. Suat Hui Yeoh, Maintz J, Foley WJ, Moran GF: **Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways.** *BMC Genomics* 2009, **10**:452.
24. Vekemans X: **AFLP-SURV version 1.0. Laboratoire de Genetique et Ecologie Vegetale.** University Libre de Bruxelles, Belgium; 2002.

doi: 10.1186/1746-4811-6-16
Cite this article as: Sansaloni *et al.*, A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus *Plant Methods* 2010, **6**:16

Molecular Phylogenetics and Evolution 59 (2011) 206–224

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Population genetic analysis and phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome-wide genotyping

Dorothy A. Steane [a,b,*], Dean Nicolle [c], Carolina P. Sansaloni [d,e], César D. Petroli [d,e], Jason Carling [f], Andrzej Kilian [f], Alexander A. Myburg [g], Dario Grattapaglia [d,e,h], René E. Vaillancourt [a,b]

[a] School of Plant Science, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia
[b] CRC for Forestry, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia
[c] Currency Creek Arboretum, PO Box 808, Melrose Park, South Australia 5039, Australia
[d] Plant Genetics Laboratory, EMBRAPA Genetic Resources and Biotechnology, EPqB, 70770-910 Brasilia, Brazil
[e] Department of Cell Biology, Universidade de Brasília, 70910-900 Brasília, DF, Brazil
[f] Diversity Arrays Technology Pty. Ltd., 1 Wilf Crane Crescent, Yarralumla ACT 2600, Australia
[g] Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria 0002, South Africa
[h] Genomic Sciences Program, Universidade Católica de Brasília, SGAN, 916 modulo B, 70790-160 Brasília, DF, Brazil

ARTICLE INFO

Article history:
Received 17 June 2010
Revised 1 February 2011
Accepted 2 February 2011
Available online 16 February 2011

Keywords:
Australia
Diversity Arrays Technology
DArT
Molecular markers
Networks
Parsimony
Plant systematics

ABSTRACT

A set of over 8000 Diversity Arrays Technology (DArT) markers was tested for its utility in high-resolution population and phylogenetic studies across a range of *Eucalyptus* taxa. Small-scale population studies of *Eucalyptus camaldulensis*, *Eucalyptus cladocalyx*, *Eucalyptus globulus*, *Eucalyptus grandis*, *Eucalyptus nitens*, *Eucalyptus pilularis* and *Eucalyptus urophylla* demonstrated the potential of genome-wide genotyping with DArT markers to differentiate species, to identify interspecific hybrids and to resolve biogeographic disjunctions within species. The population genetic studies resolved geographically partitioned clusters in *E. camaldulensis, E. cladocalyx, E. globulus* and *E. urophylla* that were congruent with previous molecular studies. A phylogenetic study of 94 eucalypt species provided results that were largely congruent with traditional taxonomy and ITS-based phylogenies, but provided more resolution within major clades than had been obtained previously. Ascertainment bias (the bias introduced in a phylogeny from using markers developed in a small sample of the taxa that are being studied) was not detected. DArT offers an unprecedented level of resolution for population genetic, phylogenetic and evolutionary studies across the full range of *Eucalyptus* species.

1. Introduction

Eucalyptus is the dominant taxon of many Australian ecosystems from subalpine woodlands, through cool and warm temperate wet and dry forests to tropical savannah (Ladiges, 1997). While the genus is easily recognised by characteristic leaf, floral and fruit morphologies, there is a huge range of quantitative variation and homoplasy (convergence/parallelism) in phenotypic characters, both among and within species (Pryor and Johnson, 1971, 1981). To further complicate matters, there is incomplete reproductive isolation of morphological species that can produce interspecific hybrids, morphological clines and hybrid swarms (Pryor and Johnson, 1971, 1981; Griffin et al., 1988), although clines can also be produced by primary differentiation (Holman et al., 2003). As a result of these factors, reconstructing the phylogenetic history of *Eucalyptus* species has been problematic for systematists, even with the application of molecular techniques. Eucalypt researchers have tested a range of molecular techniques (see below), but none has proven to be suitable for resolving relationships among closely related species within sections or between closely related sections. A marker system is needed that can resolve species-level relationships; that can be applied to a large number of samples across a broad taxonomic range; and that is relatively cheap. Diversity Arrays Technology (DArT; Jaccoud et al., 2001), a massively-parallel, array-based genotyping system, may provide the genome-wide coverage, resolution and throughput to meet these requirements.

Allozymes were the first source of molecular markers in eucalypts (Brown et al., 1975). They were used mainly to target population-level questions such as mating system, genetic diversity and population differentiation (reviewed by Moran (1992) and Potts and Wiltshire (1997)). The first phylogenetic study using allozymes in eucalypts was by Burgess and Bell (1983) who examined

* Corresponding author at: School of Plant Science, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia. Fax: +61 (0)3 62262698.
E-mail address: Dorothy.Steane@utas.edu.au (D.A. Steane).

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.02.003

allozyme frequencies in the intergrading species *Eucalyptus grandis* and *Eucalyptus saligna* (subg. *Symphyomyrtus*, sect. *Latoangulatae*, ser. *Transversae*). In this and other studies (e.g., Cook and Ladiges, 1998; House and Bell, 1994, 1996; Wright and Ladiges, 1997), allozymes provided only low to moderate levels of variation within single species or between closely related species. Although allozymes are relatively cheap and simple, they require optimisation for each study and provide very few polymorphic loci per species and few alleles per locus and, therefore, are not suitable for high-resolution population genetic studies nor for large-scale phylogenetic studies.

DNA-based studies of eucalypts began in the early 1990s (Steane et al., 1992) with the then state-of-the-art technology of restriction fragment length polymorphism (RFLP) analysis of chloroplast DNA (cpDNA). Because of the expense, a paucity of markers and the large amount of labour involved, only small numbers of samples and markers were used, providing rather coarse resolution of phylogenetic relationships among higher eucalypt taxa (genera and subgenera; Sale et al., 1993, 1996). As DNA analytical methods progressed and became cheaper, fine-scale restriction site analysis of cpDNA was tested as a means to resolve relationships among closely related species within ser. *Viminales* (sect. *Maidenaria*, subgenus *Symphyomyrtus*) (Steane et al., 1998). Although this methodology provided improved resolution of clades, it became apparent that cpDNA haplotypes were not species-specific in *Eucalyptus* and hence not useful for phylogenetic resolution at that low taxonomic level (they were, however, useful for studies of phylogeography; e.g., Byrne and Hines, 2004; Byrne and Macdonald, 2000; Jackson et al., 1999; Wheeler and Byrne, 2006).

In contrast, RFLP analysis of nuclear loci proved to be effective for many genetic studies within species or complexes of a few closely related species (e.g., Butcher et al., 2002; Byrne et al., 1998; Byrne, 1999; Elliott and Byrne, 2003; Elliott and Byrne, 2004; Glaubitz et al., 2003; Hines and Byrne, 2001; Wheeler et al., 2003). Nuclear RFLPs have not, however, yielded any useful phylogenetic data at taxonomic levels higher than the species level within *Eucalyptus*, because of issues associated with homology assessment and increasing risk of character state homoplasy with increasing taxonomic distances. Furthermore, membrane-based RFLP techniques did not lend themselves to studies requiring high-throughput analysis of large numbers of individuals.

In the early 1990s, 5S ribosomal DNA sequence variation was tested for use in phylogenetic resolution of *Eucalyptus* taxa (Udovicic et al., 1995), but the approach was only informative at high taxonomic levels (genera and subgenera). In the mid-1990s sequencing technology improved rapidly and by the end of the decade, PCR-based sequencing and automated DNA analysers allowed the production of relatively large and informative sequence data sets, with cost being the main limiting factor to the size of a study. Steane et al. (1999, 2002, 2007) and Whittock et al. (2003) used sequence data from the internal transcribed spacer (ITS) of the nuclear ribosomal DNA region to explore phylogenetic relationships across all subgenera of *Eucalyptus* and related eucalypt genera (*Corymbia* and *Angophora*). They found that ITS data provided good resolution of sections and higher taxa, but did not contain enough polymorphism to resolve effectively species-level relationships between and within sections. Furthermore, some of the higher-level relationships between eucalypt genera depicted by ITS sequence data caused consternation among the taxonomic community. For example, ITS sequence data, cpDNA RFLPs and chloroplast restriction site data all suggested that *Corymbia* was paraphyletic; this assertion was countered by evidence from other sources such as the external transcribed spacer (ETS) of the nuclear ribosomal DNA region (Parra-O et al., 2006), microsatellites (Ochieng et al., 2007b) and a pseudogene of ITS (Ochieng et al., 2007a). One problem with using sequence data from functional regions of DNA (such as ITS and ETS) comes from the functional constraints imposed on cistrons that might prevent "neutral" change of nucleotides during evolution. Furthermore, there are many copies of ribosomal RNA genes in a genome and this introduces a risk of comparing paralogous loci (Bayly and Ladiges, 2007). Despite the limitations of ribosomal and chloroplast DNA for resolution of species-level relationships, Gibbs et al. (2009) successfully used ITS, ETS and cpDNA sequence data in combination with morphological characters to resolve relationships among species within subgenus *Eudesmia*. Although none of the data sets in isolation produced a well-resolved phylogeny of the eudesmids there were elements of congruence in a combined analysis that provided the basis of a sound system of subdivision for that subgenus.

Because of complications associated with paralogy in multiple-copy regions of DNA (e.g., nuclear ribosomal DNA), researchers turned to low-copy number nuclear genes for phylogenetic and phylogeographic analyses. McKinnon et al. (2005) used the cinnamoyl-CoA reductase (*CCR*) gene to gain insights into the evolutionary history of *Eucalyptus globulus*. Two highly divergent lineages of the *CCR* gene were identified within *E. globulus*, one of which was also found in 16 other species in subg. *Symphyomyrtus*, sect. *Maidenaria*. The other lineage was unique to *E. globulus* among the *Maidenaria* taxa, but showed homology to *CCR* in *E. saligna* (subg. *Symphyomyrtus*, sect. *Latoangulatae*), suggesting either incomplete lineage sorting or reticulate evolution. Poke et al. (2006) investigated this further and found more evidence of intersectional hybridisation in *Eucalyptus*. The authors concluded that using (single-copy, functional) nuclear genes for phylogeny reconstruction of eucalypt taxa would be problematic unless recombination was taken into account.

A genome-wide approach to phylogeny reconstruction, preferably using "neutral" loci (the evolution of which was unconstrained by functional requirements) that could be analysed with a combination of population genetic and phylogenetic approaches, could circumvent complications experienced with single locus analyses in *Eucalyptus*. The development of microsatellite primers for eucalypt taxa (Brondani et al., 1998, 2006; Byrne et al., 1996; Glaubitz et al., 2001; Jones et al., 2001; Ottewell et al., 2005; Shepherd et al., 2006; Steane et al., 2001; Thamarus et al., 2002) opened the door for reliable genome-wide genotyping of a relatively large number of samples. Microsatellite markers gave researchers the power to examine genetic relationships within and among populations of one (e.g., Butcher et al., 2009; Elliott and Byrne, 2003; Jones et al., 2007; Payn et al., 2008; Rathbone et al., 2007; Steane et al., 2006; Walker et al., 2009; see also Byrne (2008) and references therein) or a few closely related species (e.g., Holman et al., 2003; Le et al., 2009; Shepherd et al., 2008; Stokoe et al., 2001). While microsatellites were developed initially for mapping and population genetic studies, Ochieng et al. (2007b) found them helpful for phylogenetic resolution of eucalypt genera. Microsatellite loci are selected by researchers to be highly polymorphic within species and their use for taxonomic purposes between closely related species is limited by the unreliable transferability of these markers across species boundaries (e.g., see Nevill et al., 2008) and by the risk of high levels of homoplasy that might be encountered (e.g., Barkley et al., 2009; Curtu et al., 2004). Hence, while microsatellites have the potential to provide phylogenetic resolution at high taxonomic levels (between genera) and are very useful for population-level studies within species, they are impractical for phylogenetic reconstruction between taxonomic extremes. Furthermore, combining datasets from different studies can be problematic; all samples need to be scored concurrently (or at least a subset of samples should be common to all studies) in order to ensure consistency of microsatellite bin sizes.

Arbitrarily amplified dominant (AAD) markers such as RAPD (randomly amplified polymorphic DNA), ISSR (inter-simple se-

quence repeats) and AFLP (amplified fragment length polymorphism) have had limited use in population and phylogenetic studies of *Eucalyptus*. AAD markers have a high potential for homoplasy, so their application to phylogenetic analysis requires careful consideration of heritability, homology and homoplasy. Bussell et al. (2005) recommended AAD markers for phylogenetic and systematic studies of closely related species and non-reticulating, subspecific lineages, since below these taxonomic levels, population genetic effects (reticulation) may swamp hierarchical signal in the data, while at higher taxonomic levels homoplasy is likely to be significant. AAD markers have been used to examine genetic structure within and between populations of individual species of *Eucalyptus* (e.g., RAPD – Nesbitt et al., 1995; Gaiotto et al., 1997; Li, 2000; ISSR – Okun et al., 2008; AFLP – Gaiotto et al., 1997; Poltri et al., 2003) but also to examine relationships among closely related (reticulating) species. In accordance with the findings of Bussell et al. (2005), McKinnon et al. (2008) could not separate closely related species within sect. *Maidenaria* (subg. *Symphyomyrtus*), with AFLP markers, but they were able to resolve series and subseries. However, the task of checking homology and repeatability of 930 AFLP markers across 84 samples was time-consuming (and hence, expensive). When analysing AFLP data, all the data for a particular study need to be scored and binned at the same time; it is not possible to score a number of data sets separately and then combine them, unless the data are checked manually. The transferability of AFLP markers across projects (and laboratories) is also problematic. Clearly, a more robust, high-throughput method for studies of closely related species would be preferable.

We recently developed a set of Diversity Arrays Technology (DArT) markers for *Eucalyptus* (Sansaloni et al., 2010) that has the potential to provide a rich source of phylogenetic information across a range of species at various taxonomic levels. DArT markers are highly variable genome-wide binary markers, the diversity of which is derived from restriction site polymorphism (Jaccoud et al., 2001). Polymorphism is detected by DNA–DNA hybridization on microarrays, allowing rapid analysis of large numbers of samples through a stream-lined automated production line (see http://www.diversityarrays.com/molecularprincip.html). We developed the markers from 65 species of *Eucalyptus* from across the taxonomic range (see Sansaloni et al., 2010) with a view to producing generic markers that would be useful in a large proportion of the 700+ species of the genus. Because DArT marker fragments are cloned (and most have been sequenced and mapped on genetic linkage maps; Petroli et al., in preparation), they do not suffer from the issues of homology assessment that exist in anonymous AAD markers. Furthermore, because of the genome-wide coverage of coding and non-coding regions (Petroli et al., in preparation), DArT has the potential to provide insights into the regions of the genome that are involved in adaptation, speciation and evolution in *Eucalyptus*.

In this study we test over 8000 DArT markers for their transferability across species and their utility in population genetics and phylogeny reconstruction. Only one other study has explored the utility of DArT markers for studies of evolution in wild populations (James et al., 2008), but that study focused on relationships between populations within two species of cryptogam (a fern, *Asplenium viride* and a moss, *Garovaglia elegans* ssp. *dietrichiae*) rather than inter-specific relationships within a genus. James et al. (2008) found that DArT markers could be highly informative about relationships among populations of cryptogam species. The present study is the first in which DArT markers have been designed for cross-species applications and applied to genus-wide studies of populations and phylogeny. Our aim in this study was to determine the degree to which DArT data can be used for: (1) studies of differentiation within and between species; (2) hybrid identification; and (3) phylogenetic reconstruction in wild populations of *Eucalyptus*.

2. Materials and methods

2.1. Genotyping with DArT markers

Three plates (94 samples per plate) of *Eucalyptus* DNA were genotyped with DArT markers. DArT-genotyping was carried out by DArT P/L (http://www.diversityarrays.com) following the procedure described by Sansaloni et al. (2010). Genotypes were scored as presence/absence of DArT markers and were formatted as binary matrices. It should be noted that this study took place in 2008 during the development of the DArT marker array for *Eucalyptus* (Sansaloni et al., 2010) and before the operational *Eucalyptus* DArT array (comprising 7680 markers) had been designed and become publicly available (mid-2009). Hence, because each plate was involved in a different phase of the DArT marker development process, the suite of DArT markers with which each plate of DNA samples was genotyped differed to some degree (Table 1). Since then, more libraries have been made and screened for polymorphism. Some of the markers that were used in the present study have now been replaced on the array with new markers that decrease redundancy. Despite these changes, we anticipate that results from the final array would be comparable to those reported in this paper, because using subsets of the available markers in this study gave results that were comparable to the other subsets and to the full set of markers (see Section 3).

2.2. Plant material for species and population differentiation surveys

Two microtitre plates of 94 samples (i.e., 188 samples in total) were genotyped with DArT markers to assess the efficacy of the markers in differentiating (1) species of *Eucalyptus* and (2) populations within a species. Plate 1 included DNA from seven species of *Eucalyptus* and one species of *Corymbia*, a close relative (formerly a subgenus) of *Eucalyptus* and was screened with 7052 DArT markers (see Table 1). Plate 2 included a larger representation of four of the most valuable timber and pulp species, *E. grandis*, *Eucalyptus urophylla s.l.* (following Brooker (2000)), *E. globulus* and *Eucalyptus nitens*; it was screened with 4684 DArT markers (Table 1). The samples came from as divergent a set of provenances for each species as could be obtained. Provenance details for these samples are provided in Appendices A and B, respectively.

2.3. Sampling for phylogenetic study

Ninety-four samples (Table 2) from across the taxonomic range of *Eucalyptus sensu stricto* (following the infrageneric classification of Brooker (2000), excluding *Corymbia* Hill and Johnson (see Hill and Johnson, 1995) and *Angophora* Cav. and treating these as gen-

Table 1
Summary of taxa represented on each plate (number of samples given in parentheses) and the number of DArT markers used in the screening of each plate. See Appendices A and B and Table 2 for details.

Plate name	Species (no. samples) on plate	No. DArT markers
Plate 1	*Eucalyptus globulus (12)*, *E. nitens (11)*, *E. grandis (12)*, *E. urophylla (12)*, *E. camaldulensis (11)*, *E. cladocalyx (12)*, *E. pilularis (12)*, *Corymbia variegata (12)*	7052
Plate 2	*E. globulus (49)*, *E. nitens (6)*, *E. nitens* × *globulus (4)*, *E. grandis (13)*, *E. urophylla (15)*, *E. grandis* × *urophylla (7)*	4684
Phylogeny	94 species	8354

Table 2
Samples used in phylogeny trial of DArT markers. Samples with superscripts were included in previous studies (ITS-based phylogenies). "Code" gives an abbreviation of the subgenus and section (where applicable) names. CCA – Currency Creek Arboretum; NSW – New South Wales; NT – Northern Territory; Qld – Queensland; SA – South Australia; Tas – Tasmania; Vic – Victoria; WA – Western Australia.

Species	Subgenus	Section	Code	Provenance	Origin or Herbarium number (ITS GenBank Accession No.)
E. albens	*Symphyomyrtus*	*Adnataria*	SA	West of Wagga Wagga, NSW	DN 2898 (HM596031)
E. amygdalina	*Eucalyptus*	*Aromatica*	EA	Kingston, SE Tas	Bridport Pole 33 (HM596032)
E. arenicola[a,h]	*Eucalyptus*	*Aromatica*	EA	Holey Plains, Gippland, SE Vic	CCA 32,13 (AF058499)
E. baileyana	*Eudesmia*	*Reticulatae*	UR	B/n Grafton & Baryulgil, NSW	DN 665 (HM596033)
E. balladoniensis ssp. *balladoniensis*	*Symphyomyrtus*	*Bisectae (II)*	SB2	Nr Mt Ney, WA	DN 3602 (HM596034)
E. bicostata	*Symphyomyrtus*	*Maidenaria*	SM	Bruthen, E Vic	UTAS 4075 (HM596035)
E. biterranea[f]	*Symphyomyrtus*	*Latoangulatae*	SL	Iron Range, Cape York Peninsula, Qld	DN 2518 (HM596036)
E. brassiana	*Symphyomyrtus*	*Exsertaria*	SE	W of Cooktown, Qld	DN 1316 (HM596037)
E. brevistylis[b]	*Eucalyptus*	*Pedaria*	EPd	E of Mt Frankland, WA	CCA 76,24: seedling from DN 1141 (AF390527)
E. brockwayi[b]	*Symphyomyrtus*	*Bisectae (II)*	SB2	NW of Norseman, WA	CCA 15,15: seedling from DN 136 (AF390505)
E. camaldulensis[g]	*Symphyomyrtus*	*Exsertaria*	SE	Palmer, Qld.	B10626 (From DArT Phase 1, Plate 1) (HM596038)
E. cladocalyx[c]	*Symphyomyrtus*	*Sejunctae*	SSj	Port Lincoln, Eyre Peninsula, SA	DN 4134 (progeny of DN 3182)(EF488228)
E. cloeziana	*Idiogenes*		Id	Isla Gorge NP, Qld	DN696 (no sequence available)
E. coolabah	*Symphyomyrtus*	*Adnataria*	SA	SE of Wilcannia, NSW	DN 2957 (HM596039)
E. cordata ssp. *cordata*	*Symphyomyrtus*	*Maidenaria*	SM	Cape Queen Elizabeth	UTAS 1475 (HM596040)
E. cornuta	*Symphyomyrtus*	*Bisectae (I)*	SB1	Nr Bremer Bay, WA	DN 3748 (HM596041)
E. cosmophylla[c]	*Symphyomyrtus*	*Incognitae*	SIn	Kingscote, Kangaroo Is., SA	CCA: DN 819 (EF488226)
E. crebra[b]	*Symphyomyrtus*	*Adnataria*	SA	NE of Tara, Qld	CCA 51, 01: seedling from DN 680 (AF390503)
E. croajingolensis[a]	*Eucalyptus*	*Aromatica*	EA	Holey Plains, Gippsland, SE Vic	TU: 29/16 (AF058497)
E. curtisii	*Acerosae*		Ac	Nr Beerwah, Qld	DN 2108 (HM596042)
E. dalrympleana[a]	*Symphyomyrtus*	*Maidenaria*	SM	Central Plateau, Tas	TU: 458/DA1 (AF058466)
E. deglupta	*Minutifructus*	*Equatoria*	ME	Philippines	Flecker BG, Cairns (HM596043)
E. delegatensis ssp. *tasmaniensis*[a]	*Eucalyptus*	*Cineraceae*	ECn	Mt Wellington, SE Tas	TU: 636 (AF058480)
E. delicata	*Symphyomyrtus*	*Bisectae (II)*	SB2	Peak Charles, WA	DN 2262 (HM596044)
E. deuaensis	*Eucalyptus*	*Capillulus*	ECp	Deua Nat Park, Southern Tablelands, NSW	DN 1769 (HM596045)
E. diversicolor[b]	*Symphyomyrtus*	*Inclusae*	SI	Walpole, WA	CCA 76, 18: seedling from DN 1142 (AY039754)
E. dives[a]	*Eucalyptus*	*Aromatica*	EA	Gembrook, S Vic.	TU: GEM4 (AF058503)
E. dundasii[b]	*Symphyomyrtus*	*Bisectae (I)*	SB1	Nr Fraser Range, WA	CCA 11, 15: seedling from DN 129 (AF390501)
E. dunnii[b]	*Symphyomyrtus*	*Maidenaria*	SM	Nr Legume, NSW	CCA 89, 31: seedling from DN 1257 (AF390510)
E. falcata[b]	*Symphyomyrtus*	*Bisectae (II)*	SB2	NE of Hopetoun, WA	CCA 17, 30: seedling from DN 198 ((AF390506)
E. gamophylla[a]	*Eudesmia*	*Limbatae*	ULm	Road to Kings Canyon, NT	CCA 09, 11 (HM596046)
E. glaucescens	*Symphyomyrtus*	*Maidenaria*	SM	St Guinear, NSW	St Guinear (HM596047)
E. glaucina	*Symphyomyrtus*	*Exsertaria*	SE	Nr Paterson, NSW	DN 2085 (HM596048)
E. globulus	*Symphyomyrtus*	*Maidenaria*	SM	Strzelecki	UTAS 1160 (HM596049)
E. gomphocephala[c]	*Symphyomyrtus*	*Bolites*	SBo	Bunbury, West Coast, WA	CCA: DN 1148 (EF488231)
E. gongylocarpa[b]	*Eudesmia*	*Limbatae*	ULm	Great Victoria Desert, WA	CCA 40, 25: seedling from DN 519 (AF390466)
E. grandis	*Symphyomyrtus*	*Latoangulatae*	SL	South Africa	#17 Zander Myburg, S. Afr. (HM596050)
E. gunnii ssp. *gunnii*[a]	*Symphyomyrtus*	*Maidenaria*	SM	Snug, SE Tas	TU: 460 (AF058464)
E. hallii[b]	*Symphyomyrtus*	*Exsertaria*	SE	Nr Goodwood, Qld	CCA 58, 06: seedling from DN 716 (AF390512)
E. houseana[b]	*Symphyomyrtus*	*Exsertaria*	SE	March Fly Glen, Kimberley, WA	CCA 134, 22: seedling from DN 1911 (AF390487)
E. howittiana[d]	*Minutifructus*	*Domesticae*	MD	Greenvale, Qld	CCA: DN 2526 (EF694709)
E. insularis	*Eucalyptus*	*Longistylus*	ELn	Mt LeGrand, S. Coast, WA	DN 1637 (HM596051)
E. jacksonii[b]	*Eucalyptus*	*Longistylus*	ELn	Valley of the Giants, WA	CCA 76, 10: seedling from DN 1140 (AF390529)
E. latisinensis[b]	*Eucalyptus*	*Amentum*	EAm	Goodwood, Qld	CCA 58, 33: seedling from DN 715 (AF390532)
E. leucophloia ssp. *leucophloia*	*Symphyomyrtus*	*Platysperma*	SP	Round Hill, W of Capricorn Roadhouse, WA	DN 539 (HM596052)
E. lockyeri ssp. *lockyeri*[b]	*Symphyomyrtus*	*Exsertaria*	SE	NW of Ravenshoe, Qld	CCA: seedling from DN 1323 (AF390488)
E. longifolia[c]	*Symphyomyrtus*	*Similares*	SSi	Eden, South Coast, NSW	CCA: DN 1750 (EF488224)
E. lucasii[b]	*Symphyomyrtus*	*Adnataria*	SA	W of Wiluna, WA	CCA 39, 12: seedling from DN 545 (AF390494)
E. maidenii	*Symphyomyrtus*	*Maidenaria*	SM	Mt Myrtle, NSW	UTAS 3125 (HM596053)
E. marginata ssp. *thalassica*[b]	*Eucalyptus*	*Longistyla*	ELn	Gingin, WA	CCA 26, 29: seedling from DN 246 (AF390530)
E. megacarpa[b]	*Eucalyptus*	*Longistylus*	ELn	Two people's Bay, WA	CCA 70, 37: seedling from DN 1137 (AF390528)
E. michaeliana[b]	*Symphyomyrtus*	*Racemus*	SR	Nr Hillgrove, NSW	CCA: seedling from DN 843 (AF390484)
E. microcorys[d]	*Alveolata*		Al	Johns River, N NSW	CCA: DN 1238 (EF694714)
E. morrisbyi	*Symphyomyrtus*	*Maidenaria*	SM	Calverts Hill, Tas	UTAS 2307 (HM596054)
E. nitens	*Symphyomyrtus*	*Maidenaria*	SM	Heyfield, Vic.	8-185 (From Phase I, Plate 1 of DArT) (HM596055)
E. nitida[a]	*Eucalyptus*	*Aromatica*	EA	Flinders Is. NE Tas	TU trial 90/1; TU N42 (AF058481)
E. obliqua[a]	*Eucalyptus*	*Eucalyptus*	EE	Mt Nelson, SE Tas	TU: 634 (AF058484)
E. obtusiflora	*Symphyomyrtus*	*Dumaria*	SD	S of Shark Bay, WA	DN 1173 (HM596056)
E. optima	*Symphyomyrtus*	*Bisectae (II)*	SB2	B/n Balladonia and Norseman, WA	DN 2154 (HM596057)
E. ovata var. *ovata*[b]	*Symphyomyrtus*	*Maidenaria*	SM	Kingston, Tas	TU 204 (AF390480)
E. pachycalyx	*Symphyomyrtus*	*Bisectae (II)*	SB2	B/n Herberton and Irvinebank, Qld	DN 1307 (HM596058)
E. pachyphylla[b]	*Symphyomyrtus*	*Bisectae (II)*	SB2	E of Mt. Webb, Gibson Desert, WA	CCA 73, 28: seedling of DN 1203 (AF390473)
E. paludicola[c]	*Symphyomyrtus*	*Incognitae*	SIn	Ashbourne, Fleurieu Peninsula, SA	CCA: DN 69 (EF488227)
E. pauciflora ssp. *pauciflora*[a]	*Eucalyptus*	*Cineraceae*	ECn	Tomahawk, NE Tas	TU: 638 (AF058489)
E. perriniana[b]	*Symphyomyrtus*	*Maidenaria*	SM	Kosciusko, NSW	UTAS 688 (AF390476)
E. pilularis[b]	*Eucalyptus*	*Pseudophloia*	EPs	Domain, Sydney, NSW (RBGS 19739)	NA (AF390533)

(*continued on next page*)

Table 2 (*continued*)

Species	Subgenus	Section	Code	Provenance	Origin or Herbarium number (ITS GenBank Accession No.)
E. piperita ssp. *urceolaris*[a]	*Eucalyptus*	*Cineraceae*	ECn	Nowra, SE NSW	DN 610 (AF058485)
E. aff. platyphylla[b,e]	*Symphyomyrtus*	*Exsertaria*	SE	E of Kupiano, Papua New Guinea	CSIRO 13400B (AF390485)
E. polyanthemos ssp. *polyanthemos*[b]	*Symphyomyrtus*	*Adnataria*	SA	Nr Rylston, NSW	CCA 46, 36: seedling from DN 742 (AF390513)
E. populnea ssp. *populnea*	*Symphyomyrtus*	*Adnataria*	SA	SE of Kogan, Qld	DN 679 (HM596059)
E. pseudoglobulus	*Symphyomyrtus*	*Maidenaria*	SM	Wiben's Hill, Vic	UTAS 627 (HM596060)
E. pulchella[a]	*Eucalyptus*	*Aromatica*	EA	Mt Nelson, SE Tas	TU: 633 (AF058487)
E. pulverulenta	*Symphyomyrtus*	*Maidenaria*	SM	Cult. Uni of Tas (Regent St)	UTAS 6065 (HM596061)
E. pumila[c]	*Symphyomyrtus*	*Pumilio*	SPu	Broken Back Range, NSW	CCA: DN 636 (EF488232)
E. raveretiana[a]	*Minutifructus*	*Domesticae*	MD	Oaky Ck, NW of Mingela, Qld	DN 1297 (HM596062)
E. regnans	*Eucalyptus*	*Eucalyptus*	EE	Leslie Vale, SE Tas	Cambridge Arboretum, Rep 1, Row F, Col. 6, Serpentine (HM596063)
E. risdonii[a]	*Eucalyptus*	*Aromatica*	EA	Meehan Range, SE Tas	TU: MRTHC1 (AF058493)
E. rubida ssp. *rubida*	*Symphyomyrtus*	*Maidenaria*	SM	Fingal/South Esk, N Tas	UTAS 693 (HM596064)
E. rubiginosa	*Primitiva*		P	Isla Gorge, Qld	DN 2114 (HM596065)
E. salmonophloia[b]	*Symphyomyrtus*	*Bisectae (II)*	SB2	Great Victoria Desert, WA	CCA 42, 14: seedling from DN 341 (AF390509)
E. scoparia[b]	*Symphyomyrtus*	*Maidenaria*	SM	Mt Norman, Qld	CCA 72, 34: seedling from DN 672 (AF390479)
E. sieberi[a]	*Eucalyptus*	*Cineraceae*	ECn	Freycinet Peninsula, E. Tas. (TU trial 90/1)	TU SIBFBR (AF058495)
E. staeri[b]	*Eucalyptus*	*Longistylus*	ELn	Wellstead, WA	CCA 70, 34: seedling from DN 1133 (AF3905531)
E. stoatei[b]	*Symphyomyrtus*	*Dumaria*	SD	SE Raventhorpe, WA	CCA 41, 34: seedling from DN 181 (AF390498)
E. tenuipes[b]	*Cuboidea*		Cb	Auburn Rd, 44 km N of Warrego Hwy	RBGS 842705 (AF390523)
E. tereticornis ssp. *tereticornis*	*Symphyomyrtus*	*Exsertaria*	SE	B/n Helidon and Crows Nest, Qld	DN 2937 (HM596066)
E. tetrodonta	*Eudesmia*	*Complanatae*	UCm	Darwin, NT	DN 5157 (HM596067)
E. tindaliae[b]	*Eucalyptus*	*Capillulus*	ECp	SE of Grafton, NSW	CCA 138, 3: seedling from DN 1243 (AF390534)
E. torquata[b]	*Symphyomyrtus*	*Dumaria*	SD	NW of Norseman, WA	CCA 43, 3: seedling from DN135 (AF390499)
E. umbra[a,h]	*Eucalyptus*	*Amenta*	EAm	NSW/Qld	Waite Arboretum, #1537 (AF058505)
E. urophylla	*Symphyomyrtus*	*Latoangulatae*	SL	Domesticated, South Africa	#15 Zander Myburg, S. Afr. (HM596068)
E. viminalis ssp. *viminalis*	*Symphyomyrtus*	*Maidenaria*	SM	Leith, NW Tas	UTAS 923 (HM596069)
E. wandoo ssp. *wandoo*[b]	*Symphyomyrtus*	*Bisectae (I)*	SB1	Stirling Range NP, WA	CCA 23, 31: seedling from DN 230 (AF390497)
E. woodwardii[a]	*Symphyomyrtus*	*Dumaria*	SD	Southern WA	Waite Arboretum, #136 (AF058479)

Samples with superscripts were included in previous studies (ITS-based phylogenies):
[a] Steane et al. (1999).
[b] Steane et al. (2002).
[c] Steane et al. (2007).
[d] Whittock et al. (2003).
[e] This sample was listed as *E. alba* by Steane et al. (2002), a close relative of *E. platyphylla* (same series) but not found in Papua New Guinea.
[f] Brooker (2000) includes *E. biterranea* in the better-known *E. pellita*.
[g] *E. camaldulensis* ssp. *acuta* or ssp. *simulata* - both subspecies occur in this region but not enough morphological information is available to determine the classification of this sample.
[h] *E. arenicola* was listed as *E. willisii* ssp. *willisii* by Steane et al. (1999).

era) were genotyped with 8354 DArT markers. Most DNA samples came from previous phylogenetic studies (McKinnon et al., 2008; Steane et al., 1999, 2002, 2007; Whittock et al., 2003) and the set of DNAs used in Plates 1 and 2 (Table 1), but several fresh leaf samples were collected from Currency Creek Arboretum (South Australia; http://www.dn.com.au/) and a *Eucalyptus* arboretum (SeedEnergy Pty. Ltd., Cambridge, Tasmania).

2.4. DNA extraction

At least 1 μg of DNA was extracted from fresh or frozen leaf tissue, using a CTAB extraction protocol (Doyle and Doyle, 1990) with several modifications (McKinnon et al., 2004). DNA was resuspended in a low-EDTA TE buffer (0.1 mM EDTA, 10 mM Tris pH 8.0). DArT analysis requires high molecular weight, restrictable DNA and all samples in a plate need to be of a uniform concentration (a total of 500–1000 ng DNA at a concentration of 50–100 ng/μl). The DNA concentration of each sample was measured using a *Pico*fluor™ handheld fluorometer (Turner designs, CA, USA) and checked by running 1 μl of DNA on a 0.8% agarose gel alongside a series of standard concentrations (10, 25, 50, 75 and 100 ng) of undigested λ DNA. The gel was post-stained with GoldView (Guangzhou Geneshun Biotech Ltd., China) and visualised using a Molecular Imager® GelDoc™ XR imaging system (BioRad Laboratories Inc.). The concentration of every sample was adjusted to approx. 75 ng/μl and re-checked on a 0.8% agarose gel. The quality of every DNA sample was tested by restriction of 2 μl DNA (ca. 150 ng) with a six-cutter enzyme, either *Eco* RV or *Hin*d III (New England Biolabs), according to the recommendations of the manufacturer. Digests were visualised on a 0.8% agarose gel, as described above. DNA preparations that did not digest properly were discarded and the samples were re-extracted.

2.5. Analysis of genetic diversity within and between species

The output of a DArT genotyping comprises "presence/absence" data along with a range of statistics that provide insight into the information content of each marker and the reliability of the data derived from each marker in that particular analysis. The stringency of an analysis can be increased by excluding data on the basis of, for example, "Reproducibility" and/or "Call Rate". As replicated individuals should give identical results, replicated points are expected to fall into the same cluster (i.e., "presence" vs. "absence"). "Reproducibility" is a measure of the consistency of scoring technical replicates (2–4 assays per sample per marker). It measures how often the replicates fall into the same cluster. This

value tends to be kept above 98.9% in the data sets, but can be adjusted upwards to 100% by eliminating markers with low reproducibility. A complementary measure of reproducibility is "Discordance" and this latter measure is usually included in the data set that is provided by DArT P/L. Discordance expresses the overall variation of scores within the technical replicates (see above). Hence, maximising the Reproducibility or minimising the Discordance will have similar effects on a data set. The Call Rate value is an expression of reliability of the final scores for each marker. It represents the percentage of samples that could be scored as 0 or 1. For Plate 2 of this study, in which there were four species (and hybrids) being genotyped with markers derived predominantly from those same species (but also some from a few other species), the call rate always exceeded 80%. In the analysis of the data from Plates 1 and 2, increasing the stringency of the data by reducing Discordance to 0% and increasing the Call Rate to 95% did not greatly affect the overall results. Hence, all data were included in the analyses presented here.

In order to determine the efficacy of DArT markers in differentiating groups of eucalypts at different taxonomic levels, several types of analysis were undertaken. Analysis of Molecular Variance (AMOVA) and a distance-based Principal Coordinates Analysis (PCoA) were done on Plate 1 data using *GenAlex 6.1* (Peakall and Smouse, 2006) to determine how the genetic diversity was partitioned among samples and whether DArT markers could be used to differentiate species, sections and subgenera. *Splitstree4* (Huson, 1998; Huson and Bryant, 2006) was used to generate relationship networks from Plate 1 and Plate 2 data sets, using the default settings of the software.

To check for potential bias in the number of polymorphic loci generated from DArT markers of different origin, the proportion of markers that gave polymorphic results on Plate 2 (two species from subg. *Symphyomyrtus*, sect. *Maidenaria* and two species from subg. *Symphyomyrtus*, sect. *Latoangulatae*) were calculated for each marker source (i.e., the taxon from which the markers were developed).

2.6. Comparison of ITS sequences of samples in the DArT phylogenetic study and samples in ITS-based phylogenies

ITS sequences were generated from 39 samples (GenBank HM596031–HM596069; Table 2) that had not been included in an ITS-based phylogenetic analysis previously (see Steane et al., 1999, 2002, 2007; Whittock et al., 2003). ITS sequences from all 94 samples were included in a phylogenetic analysis of all available ITS sequences (from the UTAS database) to confirm that the samples used in the DArT analysis were genetically comparable to other samples of the same species and/or section, and to provide a phylogeny that would be comparable to phylogenies derived from DArT data. Aligned ITS sequences were analysed as described previously (Steane et al., 2002), using both parsimony (PAUP*4.0b10; Swofford, 1999) and Bayesian (MrBayes 3.1; Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) methods using both desktop computers and the freely available University of Oslo Bioportal computer (www.bioportal.uio.no). The Bayesian analysis used the TrN + I + G model of nucleotide substitution, with parameters calculated by Modeltest Ver. 3.7 (Posada and Crandall, 1998). Two runs of a Bayesian analysis were started from a random tree and were run simultaneously to convergence over 4×10^6 generations, using four incrementally heated Markov chains, employing the default heating values. The Markov chains were sampled each 100th generation, yielding 40,001 trees, of which the first 10,000 were discarded as "burnin". The remaining sample points were used to generate a consensus tree.

2.7. Phylogenetic analysis of DArT data

Bayesian and cladistic approaches were taken to the phylogenetic analysis of DArT data using both desktop computers and the freely available University of Oslo Bioportal computer (www.bioportal.uio.no).

The Bayesian analysis, using MrBayes 3.1 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), adopted a restriction site (binary) model of evolution. An analysis that was conducted as described above for the ITS sequence data failed to reach convergence, so the analysis was modified so that the temperature was lowered from the default temperature of 0.2 to 0.1, and the number of swaps between chains was increased from the default value of 1 to 2 per swapping event. However, even with these modifications the parallel runs failed to converge. Bayesian analysis of DArT data was abandoned at this stage.

Maximum parsimony analysis of the DArT data set (using PAUP*4.0; Swofford, 1999) comprised a heuristic search using 10,000 replicates of a random stepwise addition sequence, TBR branch swapping and steepest descent in effect. Bootstrapping comprised 1×10^6 replicates of the "fast step-wise" algorithm.

The DArT markers used in the phylogenetic study were derived from seven species, four sections and two subgenera of *Eucalyptus* (Table 3). The characters in the phylogenetic data set were divided into subsets of DArT markers that had been derived from particular taxa (e.g., subgenus *Symphyomyrtus* sections *Maidenaria* and *Latoangulatae*; subgenus *Symphyomyrtus* and subgenus *Eucalyptus*). The Incongruence Length Difference (ILD) test of Farris et al. (1994; as implemented in PAUP*4.0b10 (Swofford, 1999) as the partition homogeneity test) was used to test whether each partition produced results that differed from results generated by random partitions of the data, and therefore whether such partitions

Table 3
Sources of DArT markers that were used to genotype 94 species of *Eucalyptus* for phylogenetic analysis.

Source species	Subgenus [Section]	No. clones
E. globulus	*Symphyomyrtus* [*Maidenaria*]	2754
E. nitens	*Symphyomyrtus* [*Maidenaria*]	861
[Subtotal]	[*Maidenaria*]	[3615]
E. grandis	*Symphyomyrtus* [*Latoangulatae*]	2517
E. urophylla	*Symphyomyrtus* [*Latoangulatae*]	882
E. grandis × *urophylla*	*Symphyomyrtus* [*Latoangulatae*]	456
[Subtotal]	[*Latoangulatae*]	[3855]
E. camaldulensis	*Symphyomyrtus* [*Exsertaria*]	429
[Subtotal]	[*Symphyomyrtus*]	7899
E. pilularis	*Eucalyptus* [*Pseudophloius*]	455
[Subtotal]	[*Eucalyptus*]	455
Total		8354

Table 4
Proportion (percentage) of samples on each plate for which ⩾95%, ⩾90%, ⩾85% and ⩾80% of DArT markers were scorable (i.e., hybridisation between the sample being genotyped and the DArT marker could be scored unambiguously as either "present" or "absent"). For example, in Plate 1, 100% of samples had scorable binary data (i.e., 0, 1) for at least 90% of the markers used in the screening of that plate; in other words, 100% of samples had less than 10% missing data.

	No. taxa	No. markers	Percentage scorable data				Total missing data per plate
			⩾95%	⩾90%	⩾85%	⩾80%	
Plate 1	7	7052	78%	100%			4.4%
Plate 2	4	4684	33%	86%	100%		6.0%
Phylogeny plate	94	8354	50%	95%		99%[a]	5.6%

[a] *E. gamophylla* had 24% missing data.

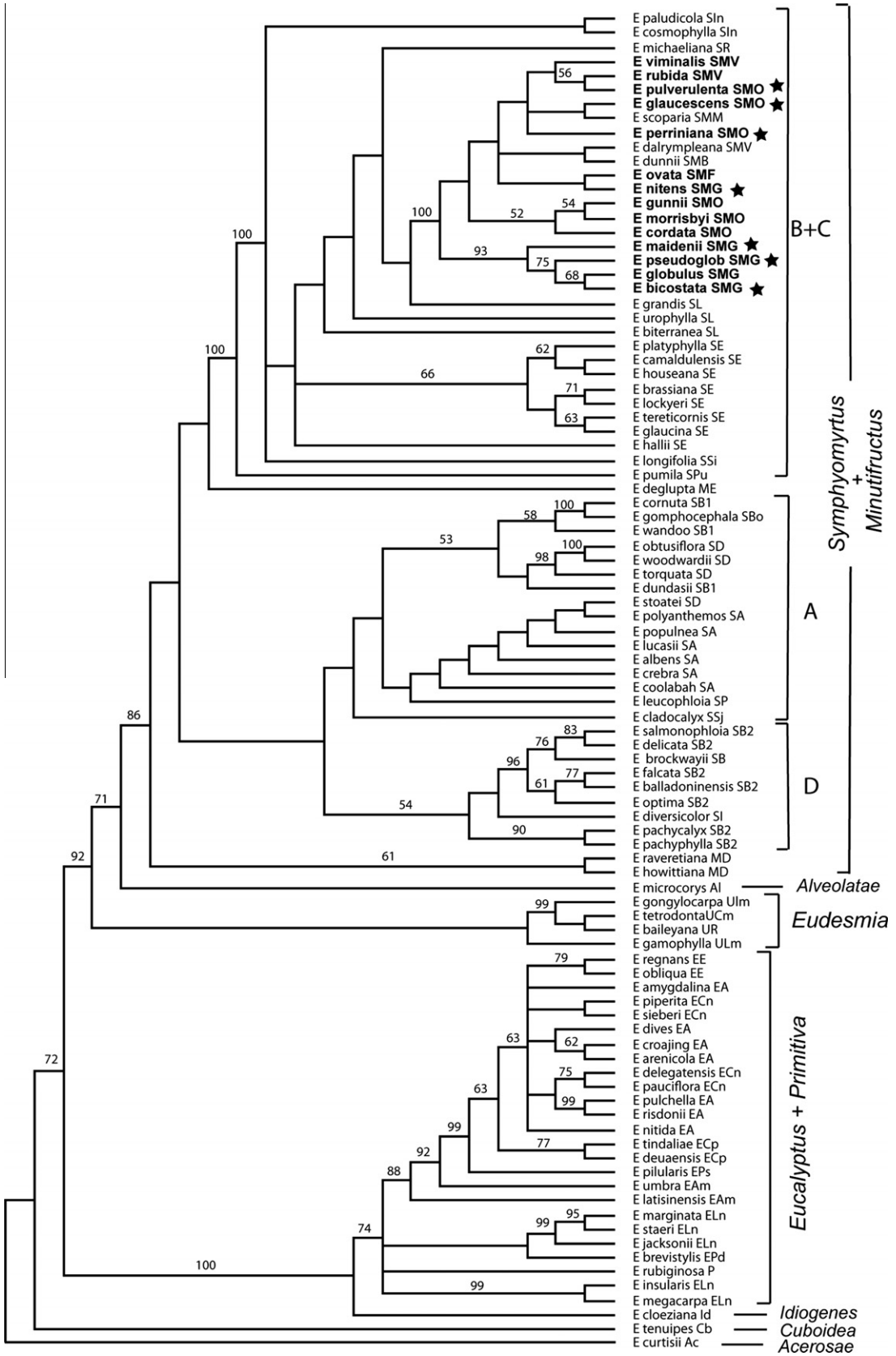


Fig. 1. Strict consensus of 10 trees derived from DArT data using PAUP*4.0b10 (length (including ten autapomorphies) = 74,861; CI (excl. autapomorphies) = 0.111; RI = 0.624). The cladograms were rooted on *E. curtisii* (subg. *Aceroceae*) on the basis of previous studies (Drinnan and Ladiges, 1991; Steane et al., 2002). Numbers above branches represent bootstrap values greater than 50%. Although resolution within subgenus *Symphyomyrtus* is good, many nodes have poor bootstrap support. Clades A, B, C and D refer to clades within subgenus *Symphyomyrtus* that were identified in phylogenies based on ITS sequence data (Steane et al., 1999, 2002, 2007). Refer to Table 2 of main text for taxon abbreviations (subgenus, section) after each species name. Series within section *Maidenaria* (SM) are shown: B – *Bridgesianae*; F – *Foveolatae*; G – *Globulares*; M – *Microcarpae*; O – *Orbiculares*; V – *Viminales*. Species relating to the AFLP study of McKinnon et al. (2008) are shown in bold, with species from mainland Australia marked with a star (see Section 4).

Table 5
Comparison of levels of polymorphism within sections *Maidenaria* and *Latoangulatae* (Plate 2) in DArT markers from different sources. The call rate is the percentage of samples that could be scored as "0" or "1" (i.e., not missing).

Marker source	Total number of markers from source used to screen Plate 2	No. (%) of markers polymorphic in *Maidenaria*	No. markers (%) with a call rate of 100% across *Maidenaria* samples	No. (%) of markers polymorphic in *Latoangulatae*	No. markers (%) with call rate of 100% across *Latoangulatae* samples
Section *Maidenaria*	1293	1006 (78%)	451 (35%)	1026 (79%)	370 (29%)
Section *Latoangulatae*	2425	1708 (70%)	981 (41%)	2107 (86%)	802 (33%)
Section *Exsertaria*	132	99 (75%)	35 (27%)	110 (83%)	38 (29%)
Subgenus *Eucalyptus*	109	89 (82%)	24 (22%)	93 (85%)	30 (28%)
Phylogeny plate	669	489 (73%)	245 (37%)	588 (88%)	204 (31%)
Corymbia	56	40 (71%)	29 (52%)	47 (84%)	26 (46%)
Mean percentage		74%	36%	84%	33%

would be expected to influence phylogenetic outcomes. One thousand replicates of the ILD test were conducted using a heuristic search strategy. The heuristic search used 10 replicates of random (stepwise) addition sequence and TBR branch swapping. The partition homogeneity test for the markers derived from subgenus *Eucalyptus* was carried out using the freely available University of Oslo Bioportal (www.bioportal.uio.no).

We also tested whether results changed when we excluded less stringent data (e.g., data that showed >0% discordance; data with call rates less than 90% or less than 95%) or when we genotyped the samples with markers derived from a particular taxon. The DArT characters were partitioned into nine subsets. Each subset was used in a *Splitstree4* analysis of the 94 species in the phylogenetic study:

1. No Discordance (i.e. excluding all characters with discordance value > 0%) (7536 characters included)
2. Call rate ⩾90% (i.e., excluding all characters with a call rate less than 90%) (6232 characters included)
3. Call rate ⩾95% (i.e., excluding all characters with a call rate less than 95%) (4094 characters included)
4. No Discordance, Call rate ⩾90% (5937 characters included)
5. No Discordance, Call rate ⩾95% (3996 characters included)
6. Subg. *Symphyomyrtus* markers only (7899 characters included)
7. Subg. *Eucalyptus* markers only (455 characters included)
8. Subgenus *Symphyomyrtus*, section *Maidenaria* markers only (3615 characters included)
9. Subgenus *Symphyomyrtus*, section *Latoangulatae* markers only (2973 characters included)

Resulting cladograms were compared to the results derived from analysis of the full DArT data set (8354 characters) to detect differences in topologies. For subsets 8 and 9, special attention was paid to the relationships among taxa in sects. *Maidenaria* and *Latoangulatae*.

3. Results

3.1. Call rate, repeatability, polymorphism and missing data

The proportion of data missing from each sample in each of the three plates was quantified (Table 4). In Plate 1 greater than 95% of the markers could be scored unambiguously (i.e., there were fewer than 5% missing data) in 78% of the samples; all samples had greater than 90% scorable markers. In Plate 2, all samples had fewer than 15% missing data (i.e., greater than 85% unambiguously scored markers). In the phylogeny study, 95% of all samples had more than 90% unambiguously scored markers and 98% of the samples had more than 80% unambiguously scored markers. In the phylogeny study, one sample, *Eucalyptus gamophylla*, had 24% missing data (reason not apparent); this might account for the anomalous position of *E. gamophylla* as sister to the rest of subgenus *Eudesmia* (Fig. 1; and see Gibbs et al., 2009). Overall the proportion of missing data for each plate was low, ranging from 4.4% in Plate 1 to 6.0% in Plate 2.

Table 5 shows the numbers of DArT markers from each marker source that were polymorphic in Plate 2 samples from subg. *Symphyomyrtus*, sect. *Maidenaria* (*E. globulus* and *E. nitens*) and compares them to the polymorphism observed in subg. *Symphyomyrtus*, sect. *Latoangulatae* (*E. grandis* and *E. urophylla*). There was a 10% difference overall in the proportion of markers that were polymorphic in the two sections, with the balance tipped in favour of sect. *Latoangulatae* (84% compared to 74% in sect. *Maidenaria*). However, the call rate (i.e., the proportion of samples for which the DArT markers provided scorable data) tended to be a little higher for sect. *Maidenaria* than for sect. *Latoangulatae*.

3.2. Differentiation within and between species

Early on in the development of the DArT markers it became clear that there was poor transferability of DArT markers between *Eucalyptus* and *Corymbia* and, because our focus was on *Eucalyptus*, we abandoned the development of DArT markers for *Corymbia* (Sansaloni et al., 2010). Hence, no results for *Corymbia* are reported.

An AMOVA analysis of the seven *Eucalyptus* species (up to 12 samples per species) in Plate 1 indicated that 36% of the DArT var-

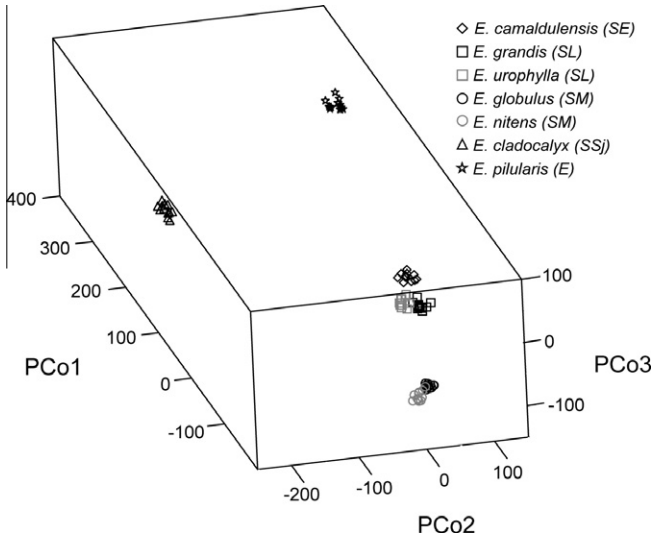


Fig. 2. Three dimensional representation of the first three Principal Components for DArT variation in seven species of *Eucalyptus*, calculated using *GenAlEx* (Peakall and Smouse 2006). The first three coordinates explained 91.9% (68.9%, 16.1% and 6.9% for axes 1, 2 and 3, respectively) of the variation among samples. Subgenera and sections are given in parentheses after species name. Subgenus *Symphyomyrtus* sections are as follows: SE = *Exsertaria*, SL = *Latoangulatae*, SM = *Maidenaria* and SSj = *Sejunctae*. Subgenus *Eucalyptus* is represented by E.

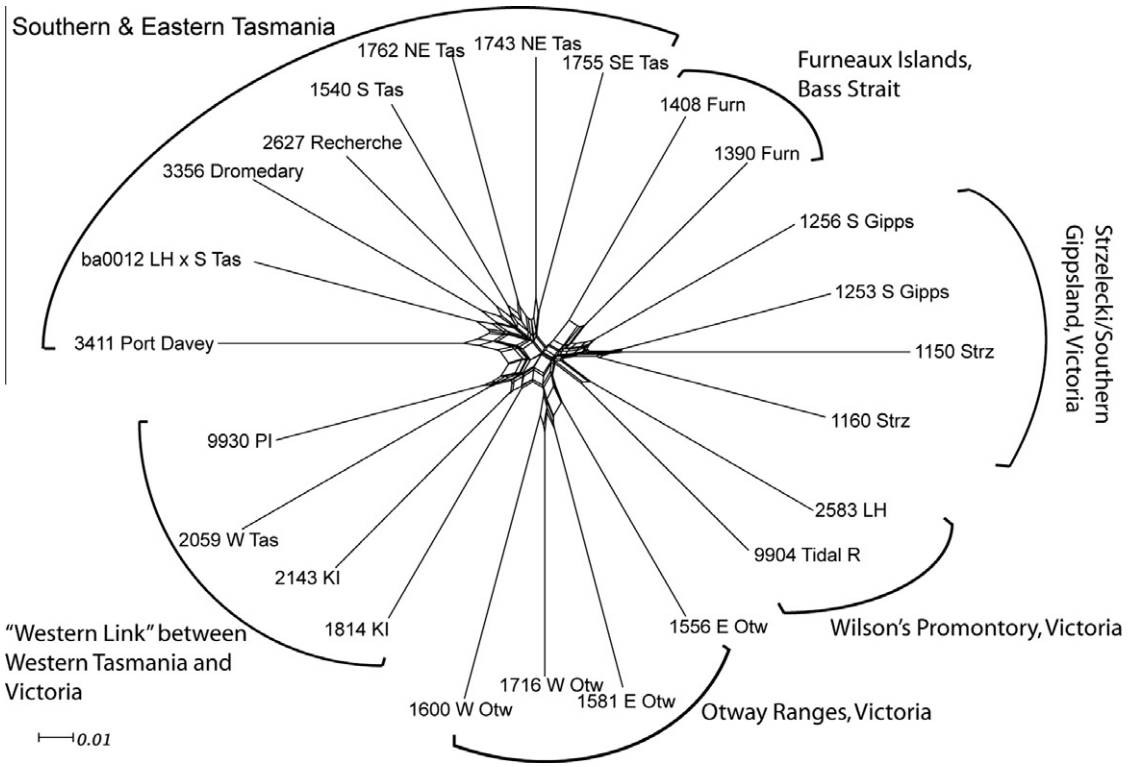


Fig. 3. Network generated by *Splitstree4* showing relationships among races of *Eucalyptus globulus* (based on DArT results from Plate 2). Abbreviations: E Otw – eastern Otway Ranges (Victoria); Furn – Furneaux group of islands, eastern Bass Strait; KI – King Island, western Bass Strait; LH – Wilson's Promotory Lighthouse (Victoria); NE Tas – northeastern Tasmania; PI – Phillip Island (Victoria); S Gipps – Southern Gippsland (Victoria); S Tas – southern Tasmania; SE Tas – southeastern Tasmania; Strz – Strzelecki Ranges (Victoria); Tidal R – Tidal River, Wilson's Promontory (Victoria); W Otw – western Otway Ranges (Victoria); W Tas – western Tasmania. Scale bar shows *Uncorrected P* genetic distance equivalent to 0.01.

iation occurred within species and 64% occurred among species. The three principal coordinates in a PCoA of Plate 1 data (Fig. 2) explained 91.9% of the variation among samples and provided complete separation of each species, with *Eucalyptus pilularis* (subgenus *Eucalyptus*, sect. *Pseudophloius*) being well-separated in PCo1 and PCo2 from a group of species from subgenus *Symphyomytus* (i.e., *E. globulus*, *E. nitens*, *E. grandis*, *E. urophylla*, *Eucalyptus camaldulensis* and *Eucalyptus cladocalyx*). The spatial distribution of the species from subgenus *Symphyomyrtus* is congruent with taxonomic relationships (see Steane et al., 2002, 2007): *E. camaldulensis* (sect. *Exsertaria*) clusters with *E. grandis* + *E. urophylla* (sect. *Latoangulatae*); *E. globulus* + *E. nitens* form a separate "*Maidenaria*" cluster.

Examination of single species in isolation demonstrated the potential of DArT to identify the provenance of individual samples. For example, DArT markers resolved population-level relationships in *E. globulus*. Fig. 3 shows a *Splitstree4* network derived from Plate 2 data. Although networks generated by *Splitstree4* are somewhat complex to look at, they do impart an understanding of the complexity of DArT data and provide a useful summary of agreement and conflict among the data. Huson and Bryant (2006) provided a clear explanation of how to interpret network figures. Briefly, the parallel lines represent two-way splits in the data. If you cut the figure across the parallel lines you can visualise to which of two groups each sample belongs. The longer the line associated with a split, the more evidence there is to support that split. This method of depicting the data is an effective way of expressing the considerable homoplasy (character conflict) in DArT data sets. The network in Fig. 3 shows the geographic differentiation of Victorian, Furneaux, Eastern Tasmanian and "Western Link" provenances of *E. globulus*, results that are congruent with microsatellite studies (Steane et al., 2006). Also in agreement with other population genetic studies, geographic structuring could be seen in *E. camaldulensis*, *E. urophylla* and *E. cladocalyx* (Supplementary material Fig. S1). *E. camaldulensis* samples from Queensland (probably mostly ssp. *acuta* but possibly also one or two samples of ssp. *simulata*; McDonald et al., 2009; David Lee, DEEDI, Queensland, pers. comm.) and Victoria (ssp. *camaldulensis*; McDonald et al., 2009) formed separate clusters (Supplementary material Fig. S1A), supporting the microsatellite results (Butcher et al., 2009) and subspecific taxonomy of the species (McDonald et al., 2009). Samples of *E. urophylla* clustered according to their Indonesian island of origin (Supplementary material Fig. S1B), in agreement with the microsatellite study of Payn et al. (2008). Geographic partitioning of DArT variation in *E. cladocalyx* (Supplementary material Fig. S1C), a species with several disjunct populations in South Australia, was similar to results obtained by McDonald et al. (2003) in an allozyme analysis of the species.

3.3. Hybrid identification

Results from the screening of the four commercially important species in Plate 2 (*E. globulus*, *E. nitens*, *E. grandis* and *E. urophylla*) demonstrated the potential of DArT markers to identify plants of hybrid origin. Fig. 4 shows the intermediate position of *E. nitens* × *globulus* hybrids between the parent species. In contrast, hybrid progeny of *E. urophylla* × *grandis* emerged within the *E. urophylla* cluster, rather than between the two species clusters. Without detailed pedigree information about the parent plants, it is impossible to say whether this is a DArT artefact or whether the *E. urophylla* parent of these hybrids was, for example, an F1 or backcross hybrid, rather than pure *E. urophylla*. Such situations can occur in breeding populations. For example, one *E. grandis* sample (labelled on Fig. 4 as "putative hybrid, South Africa") was originally believed to be pure *E. grandis*. However, when the DArT

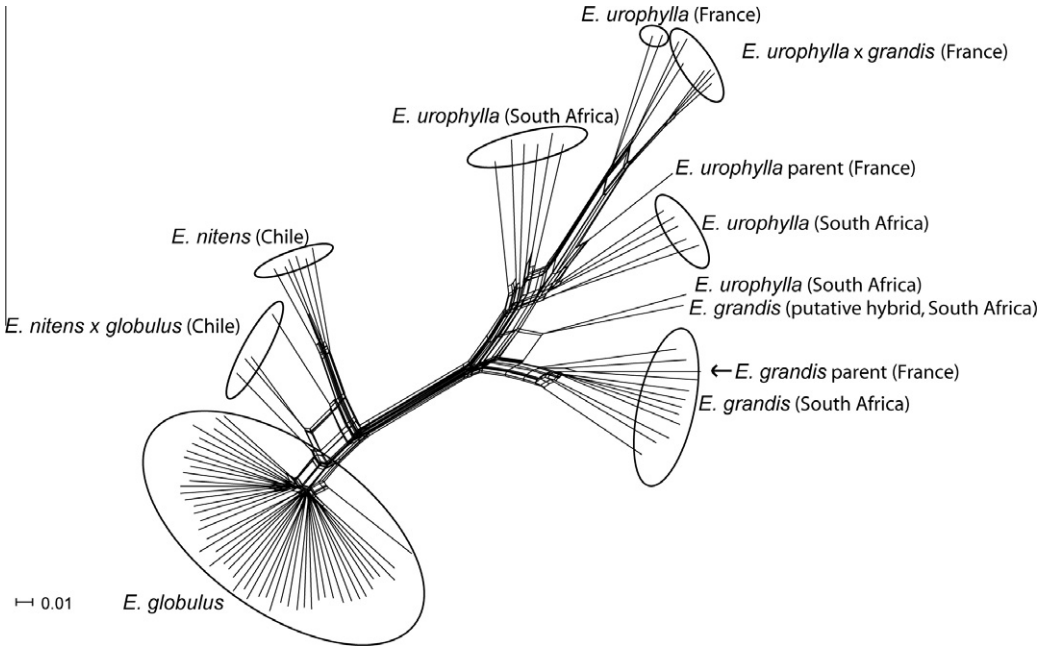


Fig. 4. *Splitstree4* network showing the positions of eucalypt hybrids relative to their parent species. While the *E. nitens* × *globulus* hybrids appear to be intermediate between the two parent species, the *E. urophylla* × *grandis* hybrids that were included in this study appear to be more closely related to *E. urophylla* than to *E. grandis* at the genome-wide level.

results showed this tree to be an outlier relative to the other *E. grandis* samples, the tree was re-examined morphologically and atypical juvenile epicormic foliage demonstrated that the tree was, in fact, not pure *E. grandis* after all and probably originated from a *E. grandis* × *E. urophylla* hybrid. Similarly, a sample of *E. urophylla* clustered with this putative hybrid, raising concern that this sample may not be pure *E. urophylla*. Unfortunately, pedigree data for these samples were not available.

3.4. Phylogenetic analysis

3.4.1. ITS sequence data

ITS sequences were generated for all samples that had not been included in previous ITS-based phylogenetic analyses (Table 2). These sequences were added to the existing database, yielding a dataset comprising 140 operational taxonomic units and 680 characters (659 aligned nucleotide characters and 21 indels scored as presence/absence data). Maximum parsimony (MP; Supplementary material Fig. S2) and Bayesian analyses (not shown) showed that all new samples but one clustered in an appropriate clade, i.e., alongside samples of the same species or section. The ITS sequence from the sample of *E. grandis* (subg. *Symphyomyrtus*, sect. *Latoangulatae*) that had been included in the phylogenetic plate of the DArT study did not cluster with the existing three samples of that species (in Clade B; see Steane et al., 2002), but emerged in a clade with sect. *Maidenaria* (Clade C; Steane et al., 2002) (see Supplementary material Fig. S2A). Despite this result, the sample of *E. grandis* (from a South African trial of Australian provenances) was retained in the DArT analysis for comparative purposes.

3.4.2. Checking the robustness of subsets of DArT data

The final DArT data set for the phylogeny analysis comprised 94 taxa and 8354 binary characters. Fig. 1 shows the strict consensus of ten trees derived from cladistic analysis of the full data set. Each subset of DArT data of different technical robustness yielded similar results. Obviously, exclusion of characters resulted in shorter branch lengths in the trees (less support for some clades), but the relationships inferred from the different character sets remained reasonably stable. Exclusion of different sets did affect some of the finer details within clades, but the changes usually affected the same taxa in each instance (i.e., in subgenus *Symphyomyrtus*, *Eucalyptus hallii* (sect. *Exsertaria*, monotypic ser. *Connexentes*), *Eucalyptus dundasii* (sect. *Bisectae*, monotypic ser. *Dundasianae*), *E. cladocalyx* (monotypic sect. *Sejunctae*), *Eucalyptus pumila* (monotypic sect. *Pumilo*) and species of sect. *Latoangulatae* relative to species of sect. *Exsertaria*; in subgenus *Eudesmia*, *E. gamophylla* (ULm); and in subgenus *Eucalyptus*, *Eucalyptus nitida*).

The overall topology of the phylogeny did not change when only DArT markers of *Symphyomyrtus* origin were used in an analysis. Using markers derived from only sect. *Maidenaria* or sect. *Latoangulatae* did not affect the general results, and at the fine scale affected the positioning of the more "mobile" taxa only (e.g., the positioning of species of sect. *Latoangulatae* and *E. hallii* (SE) relative to sect. *Exsertaria*, *E. dundasii* (SB1) and *E. gamophylla* (ULm); see comment above). In contrast, using only the markers derived from subgenus *Eucalyptus* yielded a less well-resolved phylogeny with some differences in topology when compared to the results from the full set of characters. However, considering that the number of characters was reduced to 5.4% of the full complement, the changes were relatively minor (results not shown). The critical influence here may be the sheer number of data that were excluded (i.e., 95% when only markers derived from subgenus *Eucalyptus* were included). The ILD test of Farris et al. (1994) was used to test the null hypothesis that a chosen partition (e.g., the markers derived from subgenus *Eucalyptus*) would not generate results that differed from any similar-sized random subset of the data. The alternative hypothesis in the test was that the partition was not random and that it generated results (in this case, phylogenies) that differed from results generated by similar-sized random subsets of the data. When the DArT markers that were derived from subgenus *Eucalyptus* were nominated as the "partition", the ILD test returned a *P* value of 0.43, indicating that this set of markers contained the same phylogenetic signal as a similar-sized random subset of markers. Hence, the different tree topologies that

Table 6
Probability values for partition homogeneity tests.

Data partition	No. clones in partition	*P* value from ILD test
Subg. *Eucalyptus*	455	0.423
Sect. *Exsertaria*	429	0.115
Sect. *Latoangulatae*	3855	0.003*
Sect. *Maidenaria*	3615	0.013*

* A significant *P* value suggests that a data partition contains phylogenetic signal that is different from that generated by random partitions of the data set.

were generated by the "subg. *Eucalyptus* (monocalypt) markers only" subset of markers compared to those generated from using "*Symphyomyrtus* markers only" were probably a result of the size of the data partition rather than anything else.

Unlike the ILD test for the markers derived from subgenus *Eucalyptus* (above), ILD tests of sect. *Maidenaria* and sect. *Latoangulatae* did not yield non-significant results (Table 6). The results suggest that the source of the markers from these sections may affect the topology of inferred phylogenies. However, this result conflicts with results from such phylogenetic analyses (above), suggesting that the ILD test may be flawed (see Section 4).

3.4.3. Bayesian and maximum parsimony analyses of DArT data

Despite extensive searches that took up to two weeks (or up to 40 h on the University of Oslo Bioportal facility), parallel runs of the Bayesian analyses failed to reach convergence, so this method was abandoned. This appears to be a common problem with Bayesian analysis of data sets comprising large numbers of taxa (see Discussion at http://treethinkers.blogspot.com/2009/04/when-mrbayes-fails.html; viewed January 2011). A lack of convergence may also be a result of conflicting phylogenetic signals within a data set (see Mossel and Vigoda, 2006) arising either (i) from different genes or genomic regions that have different phylogenetic histories or (ii) as a result of interspecific hybridization.

MP analysis (Fig. 1) of the complete DArT data set yielded ten equally most parsimonious trees (length = 74861; consistency index, CI = 0.112; retention index, RI = 0.624) with a strict consensus topology comparable to topologies derived from ITS sequence data in this (see Supplementary material Fig. S2) and previous studies. Because CI is negatively correlated with the number of taxa and characters in an analysis (Forey et al., 1992), it is not surprising that the CI value was so low. This does not necessarily mean that homoplasy was particularly problematic, since the retention index (RI) indicated that 62% of the similarities on the tree were synapomorphic (Farris, 1989). Although the resolution among taxa was high, bootstrap values revealed that many of the relationships depicted in the trees had low statistical support. Reweighting characters on the basis of CI or RI did not greatly alter the gross topology of the MP strict consensus trees derived from the DArT data, but resulted in a loss of resolution within and between the main clades (data not shown).

Because of the limited sampling in this analysis, we were hesitant to draw many inferences from the fine topological details of the MP strict consensus cladogram. In general, the DArT analyses (of the whole data set or of subsets of taxa) produced trees that were congruent with existing phylogenies from DNA sequence data (Steane et al., 1999, 2002, 2007; Gibbs et al., 2009), SSR-based population studies (e.g., *E. globulus* species complex; Jones, 2009), AFLP studies (McKinnon et al., 2008) and morphology-based classifications (e.g., Brooker, 2000). The four main clades (A–D) of subgenus *Symphyomyrtus* that are always found in ITS analyses were apparent. The main difference was the lack of differentiation of Clades B and C and the sister relationship of Clades A and D in the DArT-based analysis (compare Fig. 1 with Supplementary material Fig. S2). Even with the exclusion from the MP analysis of the aberrant sample of *E. grandis* (sect. *Latoangulatae*), the position of *E. urophylla* (sect. *Latoangulatae*) relative to sect. *Maidenaria* (Clade C) and the other representative of sect. *Latoangulatae* (*Eucalyptus biterranea*) that clustered with sect. *Exsertaria* (Clade B) remained unresolved (results not shown). Also in contrast to ITS-based analyses, *E. cladocalyx* (monotypic sect. *Sejunctae*) formed the sister group to the rest of Clade A, rather than being embedded within Clade A.

DArT data provided more resolution than ITS sequence data within subgenus *Eucalyptus*. Some closely related species formed clades (e.g., *Eucalyptus regnans* and *Eucalyptus obliqua* from sect. *Eucalyptus*; *Eucalyptus delegatensis* and *Eucalyptus pauciflora* from sect. *Cineraceae; Eucalyptus pulchella* and *Eucalyptus risdonii* from sect. *Aromatica*, ser. *Insulanae*), but many of Brooker's (2000) sections were not found to be monophyletic. The positions of subgenus *Idiogenes* as sister to subgenus *Eucalyptus* and of subgenus *Primitiva* embedded within subgenus *Eucalyptus* are congruent with previous studies based on nuclear ribosomal DNA sequence data (Steane et al., 1999, 2002; Ladiges et al., 2010).

Subgenus *Eudesmia* formed a clade that was well-supported apart from the positioning of *E. gamophylla* at the base of the clade (Fig. 1), a position that was incongruent with other data (Gibbs et al., 2009) and probably due to a high level of missing data for that species. Monotypic subgenus *Alveolata* (*Eucalyptus microcorys*) emerged as sister to subgenera *Symphyomyrtus* and *Minutifructus*; monotypic subgenus *Cuboidea* (*Eucalyptus tenuipes*) was sister to all other eucalypts except for *Eucalyptus curtisii* (monotypic subgenus *Acerosae*) which was used as the outgroup in this analysis (Fig. 1).

4. Discussion

4.1. Application of DArT to studies of genetic differentiation within and between species

Rigorous statistical analysis of populations requires a large number of individuals and/or large numbers of characters that sample genetic diversity throughout the genome. Although attempts have been made to develop high-throughput marker systems, it has been difficult to generate large numbers of polymorphic markers that can be applied efficiently to large numbers of samples using techniques such as RFLP (e.g., Byrne et al., 1998; Butcher et al., 2002), RAPD (Nesbitt et al., 1995) and even microsatellites (Steane et al., 2001; Brondani et al., 1998, 2006). AFLP markers can be scaled up to produce a high-throughput system (Myburg et al., 2001), but even with these relatively abundant markers, developing large numbers of reliable, high-quality polymorphic markers is laborious and expensive due to the requirement for gel or capillary electrophoresis. The development of an automated DArT marker system for use across many species of *Eucalyptus* allows rapid, high-throughput whole-genome analysis of numerous individuals across a wide range of species.

In the development of the DArT markers, although we focussed on species of commercial importance, we aimed to produce an array that would provide polymorphic markers for use in all species of *Eucalyptus*. Most of the markers were developed from four commercially important species in two sections of subgenus *Symphyomyrtus*: *E. grandis* and *E. urophylla* from sect. *Latoangulatae*, and *E. globulus* and *E. nitens* from sect. *Maidenaria*. In addition, substantial numbers of markers were developed from *Corymbia variegata, E. camaldulensis* (subgenus *Symphyomyrtus*, sect. *Exsertaria*) and *E. pilularis* (subgenus *Eucalyptus*). Libraries were also created from 64 DNA samples from the phylogeny plate, so there is a small set of (ca. 700) markers derived from a mixture of DNAs representing

the full taxonomic range of *Eucalyptus* (see Sansaloni et al., 2010). We explored the possibility that markers derived from one taxon would be more polymorphic in that taxon than in another (the so-called "ascertainment bias" (Clark et al., 2005)). We found a very slight bias towards higher polymorphism in sect. *Latoangulatae* than in sect. *Maidenaria* when using markers derived from sect. *Latoangulatae*, but since a slight bias was also observed in sect. *Latoangulatae* with markers derived from other sections (e.g., sect. *Exsertaria* and a range of taxa from phylogeny plate; Table 5) this may have been an artefact of how the markers were selected (e.g., a slight bias towards selecting markers that were polymorphic in section *Latoangulatae*) rather than an intrinsic taxonomic bias. Thus, there was no evidence of strong ascertainment bias in the eucalypt DArT array. The level of DArT variation within populations and the application of DArT to population-level studies were surveyed in just a small number of species from which most of the markers were developed (*E. cladocalyx*, *E. globulus*, *E. nitens*, *E. grandis*, *E. urophylla* and *E. pilularis*). Although we feel that the success of these studies is a good indication of their wider applicability, further studies involving larger numbers of individuals per population and species will be required to determine the utility of the markers for fine-scale population studies in other taxa.

Each individual tree in our study had a unique genotype, as would be expected from a genome-wide fingerprint comprising thousands of markers. The fact that polymorphism of DArT markers relies primarily on restriction site polymorphism (in most cases polymorphisms come from single nucleotide mutations in restriction sites) means that DArT markers are highly heritable and can be traced readily in pedigrees (Sansaloni et al., 2010). The sample sizes for each species in this study were small, but where known geographic partitioning existed among samples of a species, DArT markers identified this partitioning. For example, DArT markers resolved population-level relationships in *E. globulus* that were consistent with results from previous molecular analyses: the DArT-based clusters of Victorian, Furneaux, Eastern Tasmanian and "Western Link" provenances were supported by both nuclear microsatellite data (Steane et al., 2006) and chloroplast DNA data (Freeman et al., 2001). Geographic structuring was also observed in *E. camaldulensis*, *E. urophylla* and *E. cladocalyx* (Supplementary material, Fig. S1). *E. camaldulensis* is widespread across mainland Australia and has significant geographical variation; McDonald et al. (2009) recognised seven infraspecific taxa, but sampling for this study included samples only from Victoria (ssp. *camaldulensis*) and Queensland (probably all ssp. *acuta*) and these formed discrete geographic clusters in *Splitstree4* analyses. This result is congruent with results from microsatellite data that showed distinct geographic clustering of *E. camaldulensis* populations across the full distribution of the species (Butcher et al., 2009). Timor mountain gum, *E. urophylla s.l.,* is a tropical species comprising a limited number of disjunct populations located on volcanic soils on seven of the Lesser Sunda Islands in eastern Indonesia. DArT results derived from Plate 2 showed distinct geographic clustering of samples from several islands (i.e., Flores, Wetar and Timor), although a few samples from Lembata (previously Lomblen) did not cluster tightly. These results are congruent with those derived from microsatellite data by Payn et al. (2008), who detected subtle island-based geographic structuring of *E. urophylla* within a highly homogeneous gene pool. The sugar gum, *E. cladocalyx,* grows naturally in three disparate regions in South Australia (Kangaroo Island, southern Eyre Peninsula and the southern Flinders Ranges) and displays significant partitioning of allozyme (McDonald et al., 2003) and DArT variation. In other species where our sampling was more-or-less continuous across part or all of the species' distribution, geographic partitioning was not observed (i.e., *E. grandis*, *E. nitens* and *E. pilularis*). This seems surprising in the case of *E. nitens*, where geographic partitioning of genetic variation has been reported previously (Byrne et al., 1998), but our sampling of *E. nitens* in this study (only one or two samples per locality; see Appendices A and B) may have been insufficient to detect such diversity partitioning. A larger sample size with multiple samples from more localities might yield more definitive results. *E. grandis* provides an interesting counterpoise. In this species that has a long and, in places, disjunct latitudinal distribution from far northern Queensland to mid-coast northern New South Wales, one might expect evidence of geographic partitioning of molecular genetic variation. However, such partitioning was not detected either with allozymes (Burgess and Bell, 1983) or with cpDNA sequence data (Jones et al., 2006), lending support to the otherwise perplexing (negative) results of the DArT study. Hence, DArT markers have the potential to be a powerful tool for detecting the geographic substructuring of genetic variation within *Eucalyptus* species. However, caution should be exercised if using DArT for estimating genetic diversity and inbreeding parameters (e.g., *F* statistics). Dominant markers are generally not considered ideal for such studies, but there are algorithms – such as those in AFLP-SURV (Vekemans, 2002) – that allow for calculations of these statistics from dominant markers.

In contrast to other molecular marker systems (including chloroplast and ribosomal DNA sequence data), DArT markers were useful for differentiating closely related species and, to some degree, closely related sections. Hence, DArT could play a role in species identification, especially if taxon-specific markers or suites of markers were identified and incorporated onto a single "taxonomy" array. However, DArT would not be a practical tool to use for "DNA barcoding" (see Kress and Erickson, 2008) of plants generally, since DArT arrays would need to be developed for all plant groups (species or genera) and this would be prohibitively expensive and time-consuming.

4.2. Application of DArT to hybrid identification and studies of introgression

Due to the genome-wide sampling of thousands of marker loci, DArT has the potential to identify interspecific hybrid material. We observed that known *E. nitens* × *globulus* hybrids fell between clusters of the two parent species when genotyped with DArT markers. However, more rigorous studies of hybrids are required to test further the behaviour of DArT markers in hybrids. Progenies of a *E. urophylla* × *grandis* cross that were genotyped in this study did not yield DArT profiles that were intermediate between the two parents, contrary to what might be expected. This might be a result of using non-pure parental material of *E. urophylla* (e.g., the parent might have been a hybrid between *E. urophylla* and a species other than *E. grandis*), or there may be directional segregation distortion of DArT markers occurring in the hybrid progeny that favours *E. urophylla* alleles over *E. grandis* alleles. Large, targeted pedigree studies are required to determine how DArT markers behave in such pedigrees.

Problems with species identification may sometimes arise when historical hybridisation has left traces of one species in morphologically pure material of another species. DArT could be used as a tool to investigate such 'reticulate evolution' among closely related species (e.g., *Eucalyptus cordata* and *E. globulus*; McKinnon et al., 2004, 2008). It may also be used to examine historical introgression among less-closely related taxa, such as taxa in different series or sections. Previous research using the cinnamoyl-CoA reductase (*CCR*) gene (Poke et al., 2006) demonstrated historic recombination among the genomes of sects. *Latoangulatae*, *Exsertaria* and *Maidenaria* (subg. *Symphyomyrtus*). The sample of *E. grandis* included in the phylogenetic analysis in this study appeared to be intermediate in its DArT profile between sect. *Maidenaria* and other representatives of sect. *Latoangulatae*. Hence, this could be

an example of historic genetic recombination among closely related sections, a topic that will be examined in a future study.

4.3. Application of DArT to phylogeny reconstruction

Traditionally DNA-based phylogenetic analyses of plants have utilised sequence data from one or several small region(s) of the genome, for example, chloroplast DNA, the ITS and/or ETS region(s) of nuclear ribosomal DNA, or sundry single copy nuclear genes. More often now, with increasingly economical DNA sequencing technologies, several regions of the genome are combined into a single analysis, but these still represent just a small proportion of the whole genome. There has been considerable debate about how well single-gene phylogenies reflect species phylogenies (see Liu et al. (2009) and references therein) and researchers have often lamented the lack of an efficient method of whole-genome phylogeny reconstruction. To overcome this issue, genome-wide marker systems such as microsatellites (e.g., Ochieng et al., 2007b; Eggert et al., 2009) and AFLP (e.g., McKinnon et al., 2008; Perrie and Shepherd, 2009) have been used for phylogenetic reconstruction (usually in studies of quite closely related species), but these systems have their limitations in terms of labour, numbers of markers, cost of marker development, transferability of markers between laboratories, the ease with which different (linked) data sets can be combined, time-consuming analysis and hierarchical level at which they are effective.

This study was the first to use DArT data to examine phylogenetic relationships among species from across a large and diverse taxonomic group. Two issues concerning the use of the markers for phylogeny reconstruction were the transferability of the markers (i.e. the DArT polymorphisms) across species and the potential for biased results from markers that were developed from one taxon and applied to phylogeny reconstruction in another (ascertainment bias).

Transferability of the markers among species of *Eucalyptus s.s.* was generally good, with polymorphic markers being available for all species; transferability between closely related taxa *Eucalyptus* and *Corymbia* was poor.

We tested a number of subsets of DArT markers (that varied either in technical reproducibility or taxonomic origin) to see whether the results obtained in different groups were markedly different; on the whole, the phylogenies did not change much. We also used the ILD test (Farris et al., 1994) to determine whether a subset of markers from a particular taxon (e.g., all markers from sect. *Maidenaria*) would yield phylogenetic patterns that differed from patterns generated by similar-sized random subsets of the full complement of markers in the data set. While the subsets of markers from subgenus *Eucalyptus* (the monocalypts) and sect. *Exsertaria* did not appear to be biased, we found positive suggestions of bias for sects. *Latoangulatae* and *Maidenaria* even though actual phylogenetic analyses showed no such effect. Ramirez (2006) reviewed numerous problems associated with the ILD test, including the fact that highly significant ILD values can be obtained when there is homoplasy in one of the data partitions and there are characters that are irrelevant to the groups-in-conflict in the other data partition. This may well be the case with DArT data where the proportion of DArT markers (derived from a particular taxon) that provide phylogenetically useful information might decrease as the taxonomic distance (between the source of the DArT marker and the taxon being genotyped) increases. For example, markers from sect. *Latoangulatae* might be more likely to be phylogenetically informative within sect. *Latoangulatae* than outside that group. However, when subsets of the markers (e.g., markers derived from sect. *Latoangulatae* or sect. *Maidenaria*) were used in phylogenetic analyses, the overall topology of the phylogenetic trees did not change greatly, suggesting that the DArT markers are informative across the full taxonomic range of *Eucalyptus* and not just in taxa that are close to the source of the markers.

The results of the DArT analyses of higher-level taxa (subgenera, sections, clades A-D within subgenus *Symphyomyrtus*) were largely concordant with those generated from ITS sequence data, regardless of which analytical method was used for the DArT data. For example, the position of subgenus *Minutifructus* within subgenus *Symphyomyrtus* (Whittock et al., 2003) was supported by the DArT data, as were the close relationships among subgenera *Eucalyptus*, *Idiogenes* and *Primitiva* (Fig. 1). At the species level, there were a few "mobile" samples (see Section 3.4.2) that tended to move around depending on which partition of data or which analytical method was used, but these were mostly taxa whose positions were unresolved in ITS analyses as well. Because of the sparse sampling across the genus for this study, opportunities for direct comparisons of species-level phylogenies derived from the DArT analysis and other studies are limited. There is one published study of AFLP variation across endemic Tasmanian species from subg. *Symphyomyrtus* sect. *Maidenaria* (McKinnon et al., 2008) that can be compared to the DArT data set. McKinnon et al. (2008) assayed 84 samples across 21 species and found that within subsect. *Euryotae*, the *E. globulus* complex (*E. globulus*, *Eucalyptus bicostata*, *Eucalyptus pseudoglobulus* and *Eucalyptus maidenii*; ser. *Globulares*) formed a distinct group and its putative sister-species, *E. nitens* (ser. *Globulares*), was an outlier; *Eucalyptus perriniana* (ser. *Orbiculares*) clustered with most species from ser. *Viminales* (e.g., *Eucalyptus rubida* and *Eucalyptus viminalis*); and the boundary between other species from ser. *Orbiculares* (e.g., Tasmanian endemics *E. cordata*, *Eucalyptus morrisbyi* and *Eucalyptus gunnii*) and subsect. *Triangulares* ser. *Foveolatae* (e.g., *Eucalyptus ovata*) was blurred. The DArT study included a small subset (13) of the samples used by McKinnon et al. (2008), as well as some additional samples from sect. *Maidenaria* from mainland Australia representing ser. *Orbiculares* (*Eucalyptus glaucescens*) and ser. *Bridgesiana* (*Eucalyptus dunnii*) from subsect. *Euryote*, and ser. *Microcarpae* from subsect. *Triangulares*. The DArT study reinforced the observation (McKinnon et al., 2008) that *E. nitens* is not closely related to other species in sect. *Maidenaria* ser. *Globulares* (Fig. 1). Furthermore, both the AFLP study and the DArT study found that the Tasmanian species belonging to ser. *Orbiculares* (e.g., *E morrisbyi*, *E. gunnii* and *E. cordata*) formed a cluster that was distinct from mainland representatives of ser. *Orbiculares* (e.g., *E. glaucescens*, *Eucalyptus pulverulenta*) and the latter tended to cluster with Tasmanian species from ser. *Viminales* (Fig. 1). *Eucalyptus perrininana* (ser. *Orbiculares*) grows on both the island of Tasmania and mainland Australia. The sample of *E. perriniana* in this study was from mainland Australia and grouped with the other mainland samples of ser. *Orbiculares*. However, when the other mainland samples of ser. *Orbiculares* were omitted from the analysis, *E. perrininana* clustered with the Tasmanian species from that series. The AFLP and the DArT studies both suggested that the boundaries between series within section *Maidenaria* are blurred. This may be a biological phenomenon or could reflect a lack of resolution provided by the two marker systems. Much denser sampling within sect. *Maidenaria* would be required for the DArT results to be convincing.

The phylogenetic analysis presented in this study represented only about 12% of all the species of *Eucalyptus* from across the species range. Analysis of the 94 species together gave fine-scale (species-level) topologies that differed slightly from those obtained when different subsets of the taxa were analysed independently (e.g., subg. *Symphyomyrtus* sect. *Maidenaria*; results not shown), most likely because of a reduction in the level of homoplasy among the characters across the taxa. Future phylogenetic studies based on DArT will employ high-density sampling of species within small taxonomic units, for example, subgenus *Eudesmia* (18 species), the genetic clusters identified in this and other studies (e.g., Clades A-

D) and individual sections (e.g., sect. *Maidenaria*). The inclusion of multiple samples of closely related species in a small taxonomic unit may help to increase the accuracy of a DArT-based interpretation of species-level evolution.

5. Conclusion

Our studies have shown that the DArT markers developed for *Eucalyptus* have great potential for studies of (1) genetic differentiation within and among species; (2) hybridisation and introgression; and (3) phylogeny reconstruction at many taxonomic levels within *Eucalyptus*. One of the most appealing aspects of DArT markers is that they are cloned, which means that issues of homology assessment are negligible. Many of the markers have been sequenced and mapped onto linkage maps (e.g., Sansaloni et al., 2010) and soon will be traced onto the completed *Eucalyptus* genome sequence (http://eucalyptusdb.bi.up.ac.za/). Searches on GenBank have indicated that at least 30% of the markers come from coding regions of the genome. Hence, in future studies we will be able to divide our data sets into markers from coding and non-coding regions, allowing us to compare phylogenies derived from regions of the genome that are under selection and regions of the genome that are (presumably) selectively neutral. We also hope to be able to identify mutations in regions of the genome that are diagnostic for particular taxa and that exhibit segregation distortion in F2 hybrid progeny, that might provide insight into genomic regions that are linked to speciation and postzygotic isolation. DArT markers in combination with the complete *Eucalyptus* genome sequence hold much promise for breakthroughs in the understanding of evolution and speciation in this complex genus.

Acknowledgments

We would like to thank the following individuals and organisations for their assistance with providing plant material or DNA: Dean Williams (Forestry Tasmania), Chris Harwood (CSIRO Sustainable Ecosystems), Peter Gore and David Boomsma (SeedEnergy), Merv Shepherd and Tim Sexton (Southern Cross University, NSW), Luke McManus (University of Melbourne), Minique de Castro (University of Pretoria, South Africa), David Lee (Department of Primary Industry, Queensland), Kelsey Joyce (Gunns Ltd.), Christophe Plomion and Jean-Marc Gion (INRA, France), Nigel England and Keiji Tomita (Albany Forestry Research Centre, Oji Paper Co., Ltd.), Cristina Marques (RAIZ, Portugal), Rebeca Sanhueza Herrera (Forestal Mininco, Chile), Kitt Payn (Mondi, South Africa) and Geoff Galloway (Sappi, South Africa). Thanks also to the following people (University of Tasmania) for technical assistance with sample collection, processing and DNA sequence analysis: Scott Nicolls, Robert Barbour, James Marthick, Jules Freeman, James Worth. We are very grateful to Gay McKinnon (formerly of University of Tasmania) for her very constructive comments on the manuscript. This research was funded with a grant from the Australian Research Council (DP0770506) and assistance from the Cooperative Research Centre for Forestry (Australia).

Appendix A

Eucalyptus samples in Plate 1 of DArT study. *Eucalyptus grandis* BRASUZ1 is the tree from which the USA Department of Energy, Joint Genome Institute produced the first complete genome sequence for *Eucalyptus*. Abbreviations: NSW – New South Wales; SC Vic – Southern Central Victoria; SA – South Australia. N/A – not applicable because *Corymbia* does not have subgenera.

Species	Subgenus, Section	Identifier	Provenance
E. globulus	*Symphyomyrtus, Maidenaria*	1138	Strzelecki, Victoria
E. globulus	*Symphyomyrtus, Maidenaria*	1279	Sth Gippsland, Victoria
E. globulus	*Symphyomyrtus, Maidenaria*	1393	Furneaux, eastern Bass Strait
E. globulus	*Symphyomyrtus, Maidenaria*	1424	Furneaux, eastern Bass Strait
E. globulus	*Symphyomyrtus, Maidenaria*	1542	Southern Tas
E. globulus	*Symphyomyrtus, Maidenaria*	1573	Eastern Otways, Victoria
E. globulus	*Symphyomyrtus, Maidenaria*	1704	Western Otways, Victoria
E. globulus	*Symphyomyrtus, Maidenaria*	1707	Western Otways, Victoria
E. globulus	*Symphyomyrtus, Maidenaria*	1723	Northeastern Tasmania
E. globulus	*Symphyomyrtus, Maidenaria*	1768	Southeastern Tasmania
E. globulus	*Symphyomyrtus, Maidenaria*	2071	Western Tasmania
E. globulus	*Symphyomyrtus, Maidenaria*	ba0010	Light House (Vic) X Southern Tasmania
E. nitens	*Symphyomyrtus, Maidenaria*	8-151	Toolangi, SC Victoria
E. nitens	*Symphyomyrtus, Maidenaria*	8-185	Mt Wellington, Victoria
E. nitens	*Symphyomyrtus, Maidenaria*	7-206	Toorongo, SC Victoria
E. nitens	*Symphyomyrtus, Maidenaria*	8-155	Toolangi, SC Victoria
E. nitens	*Symphyomyrtus, Maidenaria*	5-201	Toorongo, SC Victoria
E. nitens	*Symphyomyrtus, Maidenaria*	4-173	Mt Erica, Thomson Valley, SC Victoria
E. nitens	*Symphyomyrtus, Maidenaria*	17-2,5	N. NSW
E. nitens	*Symphyomyrtus, Maidenaria*	8-207	N Toorongo, SC Victoria
E. nitens	*Symphyomyrtus, Maidenaria*	17-9,16	S. NSW
E. nitens	*Symphyomyrtus, Maidenaria*	CP20(1)	Connors Plains, Victoria
E. nitens	*Symphyomyrtus, Maidenaria*	CP186(3)	Connors Plains, Victoria
E. grandis	*Symphyomyrtus, Latoangulatae*	1	Baldy State Forest, Queensland
E. grandis	*Symphyomyrtus, Latoangulatae*	2	Mareeba, Queensland
E. grandis	*Symphyomyrtus, Latoangulatae*	3	Ravenshoe, Queensland
E. grandis	*Symphyomyrtus, Latoangulatae*	4	Townsville, Queensland

(*continued on next page*)

Appendix A (*continued*)

Species	Subgenus, Section	Identifier	Provenance
E. grandis	*Symphyomyrtus, Latoangulatae*	5	Kenilworth, Queensland
E. grandis	*Symphyomyrtus, Latoangulatae*	6	Veteran Gympie, Queensland
E. grandis	*Symphyomyrtus, Latoangulatae*	8	Toonumba, NSW
E. grandis	*Symphyomyrtus, Latoangulatae*	9	Lake Cathie, NSW
E. grandis	*Symphyomyrtus, Latoangulatae*	11	Mt George, Taree, NSW
E. grandis	*Symphyomyrtus, Latoangulatae*	12	Coffs Harbour, NSW
E. grandis	*Symphyomyrtus, Latoangulatae*	13	Parent of mapping population
E. grandis	*Symphyomyrtus, Latoangulatae*	BRASUZ1	DOE-JGI target genome
E. urophylla	*Symphyomyrtus, Latoangulatae*	2	Lere-Baukrenget, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	3	le Nggele, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	4	Jontona, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	5	Labalekan, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	6	Beangonong, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	7	Delaki, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	8	Bonleu, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	9	Mollo, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	10	Pintu Mas, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	11	Apui, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	12	Nesunhuhun, Indonesia
E. urophylla	*Symphyomyrtus, Latoangulatae*	13	Parent of mapping population
E. camaldulensis	*Symphyomyrtus, Exsertaria*	3R33B	Lake Albacutya E, Victoria
E. camaldulensis	*Symphyomyrtus, Exsertaria*	3R32A	Kororoit Ck, Melton, Victoria
E. camaldulensis	*Symphyomyrtus, Exsertaria*	2R32D	Edenhope, Victoria
E. camaldulensis	*Symphyomyrtus, Exsertaria*	2R34B	Edenhope, Victoria
E. camaldulensis	*Symphyomyrtus, Exsertaria*	PET32	Petford, Queensland
E. camaldulensis	*Symphyomyrtus, Exsertaria*	PET116	Petford, Queensland
E. camaldulensis	*Symphyomyrtus, Exsertaria*	CST01009	Lake Albacutya, Victoria
E. camaldulensis	*Symphyomyrtus, Exsertaria*	B10530	Mitchell, Queensland
E. camaldulensis	*Symphyomyrtus, Exsertaria*	B10540	Morehead, Queensland
E. camaldulensis	*Symphyomyrtus, Exsertaria*	B10531	Laura, Queensland
E. camaldulensis	*Symphyomyrtus, Exsertaria*	B10626	Palmer, Queensland
E. cladocalyx	*Symphyomyrtus, Sejunctae*	DN 2569	Horrocks Pass, Flinders Ranges, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	DN 3182	Port Lincoln, Eyre Peninsula, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	K006	Kangaroo Is, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	K046	Kangaroo Is, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	K047	Kangaroo Is, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	K048	Kangaroo Is, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	K051	Kangaroo Is, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	W002	Wirrabara, Flinders Ranges, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	W004	Wirrabara, Flinders Ranges, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	W012	Wirrabara, Flinders Ranges, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	W025	Wirrabara, Flinders Ranges, SA
E. cladocalyx	*Symphyomyrtus, Sejunctae*	W027	Wirrabara, Flinders Ranges, SA
E. pilularis	*Eucalyptus, Pseudophloia*	9025	Goonengerry, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9468	Tamban, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9479	Gallangowan, Queensland
E. pilularis	*Eucalyptus, Pseudophloia*	9659	Olney, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9720	Kiwarrak, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9742	Bulga, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9789	Clouds Creek, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9910	Whain Whain, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9375	Orara East, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9415	Coopernook, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9464	Kerewong, NSW
E. pilularis	*Eucalyptus, Pseudophloia*	9519	Newry, NSW
Corymbia variegata	*N/A, Politaria*	4893	Wedding Bells, NSW
Corymbia variegata	*N/A, Politaria*	7782	Richmond Ranges, NSW
Corymbia variegata	*N/A, Politaria*	7785	Presho, SW of Yeppoon, Queensland
Corymbia variegata	*N/A, Politaria*	7797	Brisbane Forest Park, Queensland
Corymbia variegata	*N/A, Politaria*	7989	Woondum, Queensland
Corymbia variegata	*N/A, Politaria*	7991	Woondum, Queensland

Appendix A (*continued*)

Species	Subgenus, Section	Identifier	Provenance
Corymbia variegata	*N/A, Politaria*	10178	Cherry Tree, NSW
Corymbia variegata	*N/A, Politaria*	10394	Brooyar, W of Gympie, Queensland
Corymbia variegata	*N/A, Politaria*	10406	Ewingar, NSW
Corymbia variegata	*N/A, Politaria*	10262	Cherry Tree, NSW
Corymbia variegata	*N/A, Politaria*	10286	Woondum, Queensland
Corymbia variegata	*N/A, Politaria*	10348	Candole, NSW

Appendix B

Eucalyptus samples used in Plate 2 of DArT study. Abbreviations: WPLH – Wilson's Promontory Lighthouse, Victoria; NSW – New South Wales.

	Species	Identifier	Race
1	*E. globulus*	1150	Strzelecki, Victoria
2	*E. globulus*	1160	Strzelecki, Victoria
3	*E. globulus*	1253	Southern Gippsland, Victoria
4	*E. globulus*	1256	Southern Gippsland, Victoria
5	*E. globulus*	1390	Furneaux, eastern Bass Strait
6	*E. globulus*	1408	Furneaux, eastern Bass Strait
7	*E. globulus*	1540	Southern Tasmania
8	*E. globulus*	1556	Eastern Otways, Victoria
9	*E. globulus*	1581	Eastern Otways, Victoria
10	*E. globulus*	1600	Western Otways, Victoria
11	*E. globulus*	1716	Western Otways, Victoria
12	*E. globulus*	1743	Northeastern Tasmania
13	*E. globulus*	1751	Southeastern Tasmania
14	*E. globulus*	1762	Northeastern Tasmania
15	*E. globulus*	1814	King Island, western Bass Strait
16	*E. globulus*	2059	Western Tasmania
17	*E. globulus*	2143	King Island, western Bass Strait
18	*E. globulus*	2583	Wilson's Promontory Lighthouse, Victoria
19	*E. globulus*	2627	Recherche Bay, southern Tasmania
20	*E. globulus*	3356	Dromedary, southeastern Tasmania
21	*E. globulus*	3411	Port Davey, Western Tasmania
22	*E. globulus*	9904	Tidal River, Wilson's Promontory, Victoria
23	*E. globulus*	9930	Phillip Island, Victoria
24	*E. globulus*	ba0012	WPLH X Southern Tasmania
25	*E. globulus*	OP1	unknown
26	*E. globulus*	OP2	unknown
27	*E. globulus*	OP3	unknown
28	*E. globulus*	OP4	unknown
29	*E. globulus*	OP5	unknown
30	*E. globulus*	OP6	unknown
31	*E. globulus*	OP7	unknown
32	*E. globulus*	OP8	unknown
33	*E. globulus*	OP9	unknown
34	*E. globulus*	OP10	unknown
35	*E. globulus*	OP11	unknown
36	*E. globulus*	OP12	unknown
37	*E. globulus*	Port1	Portugal
38	*E. globulus*	Port2	Portugal
39	*E. globulus*	Ch1	Chivilingo, Chile
40	*E. globulus*	Ch2	Chivilingo, Chile
41	*E. globulus*	Ch3	Manzano Miramar, Chile
42	*E. globulus*	Ch4	CerroAlto, Chile
43	*E. globulus*	Ch5	Araneda, Chile
44	*E. globulus*	Ch6	Araneda, Chile
45	*E. globulus*	Ch7	Manzano Miramar, Chile

(*continued on next page*)

Appendix B (*continued*)

	Species	Identifier	Race
46	*E. globulus*	Ch8	Maquehua, Chile
47	*E. globulus*	Ch9	Araneda, Chile
48	*E. globulus*	Ch10	CerroAlto, Chile
49	*E. globulus*	Ch11	Araneda, Chile
50	*E. nitens*	Ch12	Tallaganda State Forest, NSW
51	*E. nitens*	Ch13	Badja State Forest, NSW
52	*E. nitens*	Ch14	Tallaganda State Forest, NSW
53	*E. nitens*	Ch15	Tallaganda State Forest, NSW
54	*E. nitens*	Ch16	Tallaganda State Forest, NSW
55	*E. nitens*	Ch17	Thomson Valley, Victoria
56	*E. nitens x globulus*	Ch18	Cuatro del recorte, Chile
57	*E. nitens x globulus*	Ch19	Cuatro del recorte, Chile
58	*E. nitens x globulus*	Ch20	Cuatro del recorte, Chile
59	*E. nitens x globulus*	Ch21	La Huina, Chile
60	*E. grandis*	grand 13	Mareeba, Queensland
61	*E. grandis*	grand 14	Townsville, Queensland
62	*E. grandis*	grand 15	Baldy State Forest, Queensland
63	*E. grandis*	grand 16	Woondum/Gympie, Queensland
64	*E. grandis*	grand 17	Kenilworth, Queensland
65	*E. grandis*	grand 18	Belthorpe, Queensland
66	*E. grandis*	grand 19	Wauchope, NSW
67	*E. grandis*	grand20	Lake Cathie, NSW
68	*E. grandis*	grand21	Mt George Taree, NSW
69	*E. grandis*	grand22	Taree, NSW
70	*E. grandis*	grand23	Bulahdelah, NSW
71	*E. grandis*	grand24	Wauchope, NSW
72	*E. urophylla*	uro13	Lasinisir, Indonesia
73	*E. urophylla*	uro14	Rotus, Indonesia
74	*E. urophylla*	uro15	Lasinisir, Indonesia
75	*E. urophylla*	uro16	Lembata, Indonesia
76	*E. urophylla*	uro17	Padekluwa Lembata, Indonesia
77	*E. urophylla*	uro18	Lembata, Indonesia
78	*E. urophylla*	uro19	Padekluwa Lembata, Indonesia
79	*E. urophylla*	uro20	Lelobatan, Timor, Indonesia
80	*E. urophylla*	uro21	Leloboko, Timor, Indonesia
81	*E. urophylla*	uro22	Kilawair Ille, Wodong, Flores, Indonesia
82	*E. urophylla*	uro23	Hokeng, Flores, Indonesia
83	*E. urophylla*	uro24	Leloboko, Timor, Indonesia
84	*E. urophylla x grandis*	Fr1	France
85	*E. urophylla x grandis*	Fr3	France
86	*E. urophylla x grandis*	Fr5	France
87	*E. urophylla x grandis*	Fr7	France
88	*E. urophylla x grandis*	Fr11	France
89	*E. urophylla x grandis*	Fr13	France
90	*E. urophylla x grandis*	Fr15	France
91	*E. urophylla*	Fr17	France
92	*E. grandis*	Fr24	France
93	*E. urophylla*	Fr20	France
94	*E. urophylla*	Fr21	France

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.02.003.

References

Barkley, N.A., Krueger, R.R., Federici, C.T., Roose, M.L., 2009. What phylogeny and gene genealogy analyses reveal about homoplasy in citrus microsatellite alleles. Plant Syst. Evol. 282, 71–86.

Bayly, M.J., Ladiges, P.Y., 2007. Divergent paralogues of ribosomal DNA in eucalypts (Myrtaceae). Mol. Phylogenet. Evol. 44, 346–356.

Brondani, R., Brondani, C., Tarchini, R., Grattapaglia, D., 1998. Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. Theor. Appl. Genet. 97, 816–827.

Brondani, R.P.V., Williams, E.R., Brondani, C., Grattapaglia, D., 2006. A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers from the genus. BMC Plant Biol. 6, 20.

Brooker, M.I.H., 2000. A new classification of the genus *Eucalyptus* L'Hér. (Myrtaceae). Aust. Syst. Bot. 13, 79–148.

Brown, A.H.D., Matheson, A.C., Eldridge, K.G., 1975. Estimation of mating system of *Eucalyptus obliqua* L'Herit. by using allozyme polymorphisms. Aust. J. Bot. 23, 931–949.

Burgess, I., Bell, J.C., 1983. Comparative morphology and allozyme frequencies of *Eucalyptus grandis* Hill ex Maiden and *Eucalyptus saligna* SM. Aust. Forest Res. 13, 133–149.

Bussell, J.D., Waycott, M., Chappill, J.A., 2005. Arbitrarily amplified DNA markers as characters for phylogenetic inference. Perspect. Plant Ecol. Evol. Syst. 7, 3–26.

Butcher, P.A., Otero, A., McDonald, M.W., Moran, G.F., 2002. Nuclear RFLP variation in *Eucalyptus camaldulensis* Dehnh. from northern Australia. Heredity 88, 402–412.

Butcher, P.A., McDonald, M.W., Bell, J.C., 2009. Congruence between environmental parameters, morphology and genetic structure in Australia's most widely distributed eucalypt, *Eucalyptus camaldulensis*. Tree Genet. Genomes 5, 189–210.

Byrne, M., 1999. High genetic identities between three oil mallee taxa, *Eucalyptus kochii* ssp. *kochii*, ssp. *plenissima* and *E. horistes*, based on nuclear RFLP analysis. Heredity 82, 205–211.

Byrne, M., 2008. Phylogeny, diversity and evolution of eucalypts. In: Sharma, A.K., Sharma, A. (Eds.), Plant Genome Biodiversity and Evolution. Part E. Phanerogams – Angiosperms, vol. 1. Science Publishers, Berlin, pp. 303–346.

Byrne, M., Hines, B., 2004. Phylogeographical analysis of cpDNA variation in *Eucalyptus loxophleba* (Myrtaceae). Aust. J. Bot. 52, 459–470.

Byrne, M., Macdonald, B., 2000. Phylogeography and conservation of three oil mallee taxa, *Eucalyptus kochii* ssp. *kochii*, ssp. *plenissima* and *E. horistes*. Aust. J. Bot. 48, 305–312.

Byrne, M., Marques-Garcia, M.I., Uren, T., Smith, D.S., Moran, G.F., 1996. Conservation and genetic diversity of microsatellite loci in the genus *Eucalyptus*. Aust. J. Bot. 44, 331–341.

Byrne, M., Parrish, T.L., Moran, G.F., 1998. Nuclear RFLP diversity in *Eucalyptus nitens*. Heredity 81, 225–233.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., Nielsen, R., 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15, 1496–1502.

Cook, I.O., Ladiges, P.Y., 1998. Isozyme variation in *Eucalyptus nitens* and *E. denticulata*. Aust. J. Bot. 46, 35–44.

Curtu, A.L., Finkeldey, R., Gailing, O., 2004. Comparative sequencing of a microsatellite locus reveals size homoplasy within and between European oak species (*Quercus* spp.). Plant Mol. Biol. Rep. 22, 339–346.

Doyle, J.J., Doyle, J.L., 1990. Isolation of plant DNA from fresh tissue. Focus 12, 13–15.

Drinnan, A.N., Ladiges, P.Y., 1991. Floral development and systematic position of *Eucalyptus curtisii*. Aust. Syst. Bot. 4, 539–552.

Eggert, L.S., Beadell, J.S., McClung, A., McIntosh, C.E., Fleischer, R.C., 2009. Evolution of microsatellite loci in the adaptive radiation of Hawaiian honeycreepers. J. Hered. 100, 137–147.

Elliott, C., Byrne, M., 2003. Genetic diversity within and between natural populations of *Eucalyptus occidentalis* (Myrtaceae). Silvae Genet. 52, 169–173.

Elliott, C.P., Byrne, M., 2004. Phylogenetics and the conservation of rare taxa in the *Eucalyptus angustissima* complex in Western Australia. Conserv. Genet. 5, 39–47.

Farris, J.S., 1989. The retention index and the rescaled consistency index. Cladistics 5, 417–419.

Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1994. Testing significance of incongruence. Cladistics 10, 315–319.

Forey, P.L., Humphries, C.J., Kitching, I.J., Scotland, R.W., Siebert, D.J., Williams, D.M., 1992. Cladistics. A Practical Course in Systematics. The Sytematics Association Publication No. 10. Oxford University Press, New York.

Freeman, J.S., Jackson, H.D., Steane, D.A., McKinnon, G.E., Dutkowski, G.W., Potts, B.M., Vaillancourt, R.E., 2001. Chloroplast DNA phylogeography of *Eucalyptus globulus*. Aust. J. Bot. 49, 585–596.

Gaiotto, F.A., Bramucci, M., Grattapaglia, D., 1997. Estimation of outcrossing rate in a breeding program of *Eucalyptus urophylla* with dominant RAPD and AFLP markers. Theor. Appl. Genet. 95, 842–849.

Gibbs, A.K., Udovicic, F., Drinnan, A.N., Ladiges, P.Y., 2009. Phylogeny and classification of *Eucalyptus* subgenus *Eudesmia* (Myrtaceae) based on nuclear ribosomal DNA, chloroplast DNA and morphology. Aust. Syst. Bot. 22, 158–179.

Glaubitz, J., Emebiri, L., Moran, G., 2001. Dinucleotide microsatellites from *Eucalyptus sieberi*: inheritance, diversity, and improved scoring of single-base differences. Genome 44, 1041–1045.

Glaubitz, J.C., Murrell, J.C., Moran, G.F., 2003. Effects of native forest regeneration practices on genetic diversity in *Eucalyptus consideniana*. Theor. Appl. Genet. 107, 422–431.

Griffin, A.R., Burgess, I.P., Wolf, L., 1988. Patterns of natural and manipulated hybridisation in the genus Eucalyptus L'Herit. – a review. Aust. J. Bot. 36, 41–66.

Hill, K.D., Johnson, L.A.S., 1995. Systematic studies in the eucalypts. 7. A revision of the bloodwoods, genus *Corymbia* (Myrtaceae). Telopea 6, 185–504.

Hines, B., Byrne, M., 2001. Genetic differentiation between mallee and tree forms in the *Eucalyptus loxophleba* complex. Heredity 87, 566–572.

Holman, J.E., Hughes, J.M., Fensham, R.J., 2003. A morphological cline in *Eucalyptus*: a genetic perspective. Mol. Ecol. 12, 3013–3025.

House, A.P.N., Bell, J.C., 1994. Isozyme variation and mating system in *Eucalyptus urophylla* ST Blake. Silvae Genet. 43, 167–179.

House, A.P.N., Bell, J.C., 1996. Genetic diversity, mating system and systematic relationships in two red mahoganies, *Eucalyptus pellita* and *E. scias*. Aust. J. Bot. 44, 157–174.

Huelsenbeck, J.P., Ronquist, F.R., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.

Huson, D.H., 1998. SplitsTree: a program for analyzing and visualizing evolutionary data. Bioinformatics 14, 68–73.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Jaccoud, D., Peng, K., Feinstein, D., Kilian, A., 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. Nucl. Acids Res. 29, e25.

Jackson, H.D., Steane, D.A., Potts, B.M., Vaillancourt, R.E., 1999. Chloroplast DNA evidence for reticulate evolution in *Eucalyptus* (Myrtaceae). Mol. Ecol. 8, 739–751.

James, K.E., Schneider, H., Ansell, S.W., Evers, M., Robba, L., Uszynski, G., Pedersen, N., Newton, A.E., Russell, S.J., Vogel, J.C., Kilian, A., 2008. Diversity Arrays Technology (DArT) for pan-genomic evolutionary studies of non-model organisms. PLoS ONE 3, e1682. doi:10.1371/journal.pone.0001682.

Jones, R.C., 2009. Molecular Evolution and Genetic Control of Flowering in the *Eucalyptus globulus* Species Complex. PhD Thesis, University of Tasmania.

Jones, M.E., Stokoe, R.L., Cross, M.J., Scott, L.J., Maguire, T.L., Shepherd, M., 2001. Isolation of microsatellite loci from spotted gum (*Corymbia variegata*) and cross-species amplification in *Corymbia* and *Eucalyptus*. Mol. Ecol. Notes 1, 276–278.

Jones, M.E., Shepherd, M., Henry, R.J., Delves, A., 2006. Chloroplast DNA variation and population structure in the widespread forest tree, *Eucalyptus grandis*. Conserv. Genet. 7, 691–703.

Jones, T.H., Vaillancourt, R.E., Potts, B.M., 2007. Detection and visualization of spatial genetic structure in continuous *Eucalyptus globulus* forest. Mol. Ecol. 16, 697–707.

Kress, W.J., Erickson, D.L., 2008. DNA barcodes: genes, genomics and bioinformatics. Proc. Natl. Acad. Sci. USA 105, 2761–2762.

Ladiges, P.Y., 1997. Phylogenetic history and classification of eucalypts. In: Williams, J.E., Woinarski, J.C.Z. (Eds.), Eucalypt Ecology: Individuals to Ecosystems. Cambridge University Press, Cambridge, pp. 16–29.

Ladiges, P.Y., Bayly, M.J., Nelson, G.J., 2010. East-west continental vicariance in *Eucalyptus* subgenus *Eucalyptus*. In: Williams, D.M., Knapp, S. (Eds.), Beyond Cladistics: The Branching of a Paradigm. University of California, pp. 267–302.

Le, S., Nock, C., Henson, M., Shepherd, M., 2009. Genetic differentiation among and within three red mahoganies (ser. *Annulares*), *Eucalyptus pellita*, *E. resinifera* and *E. scias* (Myrtaceae).. Aust. Syst. Bot. 22, 332–343.

Li, C.Y., 2000. RAPD analysis of genetic variation in *Eucalyptus microtheca F. Muell* populations. Hereditas 132, 151–156.

Liu, L., Yu, L., Kubatko, L., Pearl, D.K., Edwards, S.V., 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53, 320–328.

McDonald, M.W., Rawlings, M., Butcher, P.A., Bell, J.C., 2003. Regional divergence and inbreeding in *Eucalyptus cladocalyx* (Myrtaceae). Aust. J. Bot. 51, 393–403.

McDonald, M.W., Brooker, M.I.H., Butcher, P.A., 2009. A taxonomic revision of *Eucalyptus camaldulensis* (Myrtaceae). Aust. Syst. Bot. 22, 257–285.

McKinnon, G.E., Vaillancourt, R.E., Steane, D.A., Potts, B.M., 2004. The rare silver gum, *Eucalyptus cordata*, is leaving its trace in the organellar gene pool of *Eucalyptus globulus*. Mol. Ecol. 13, 3751–3762.

McKinnon, G.E., Potts, B.M., Steane, D.A., Vaillancourt, R.E., 2005. Population and phylogenetic analysis of the cinnamoyl coA reductase gene in *Eucalyptus globulus* (Myrtaceae). Aust. J. Bot. 53, 827–838.

McKinnon, G.E., Vaillancourt, R.E., Steane, D.A., Potts, B.M., 2008. An AFLP marker approach to lower-level systematics in *Eucalyptus* (Myrtaceae). Am. J. Bot. 95, 368–380.

Moran, G.F., 1992. Patterns of genetic diversity in Australian tree species. New Forests 6, 49–66.

Mossel, E., Vigoda, E., 2006. Limitations of Markov Chain Monte Carlo algorithms for Bayesian inference of phylogeny. Ann. Appl. Probab. 16, 2215–2234.

Myburg, A.A., Remington, D.L., O'Malley, D.M., Sederoff, R.R., Whetten, R.W., 2001. High-throughput AFLP analysis using infrared dye-labeled primers and an automated DNA sequencer. Biotechniques 30, 348–357.

Nesbitt, K.A., Potts, B.M., Vaillancourt, R.E., West, A.K., Reid, J.B., 1995. Partitioning and distribution of RAPD variation in a forest tree species, *Eucalyptus globulus*. Heredity 74, 628–637.

Nevill, P.G., Reed, A., Bossinger, G., Vaillancourt, R.E., Larcombe, M., Ades, P.K., 2008. Cross-species amplification of *Eucalyptus* microsatellite loci. Mol. Ecol. Resour. 8, 1277–1280.

Ochieng, J.W., Henry, R.J., Baverstock, P.R., Steane, D.A., Shepherd, M., 2007a. Nuclear ribosomal pseudogenes resolve a corroborated monophyly of the eucalypt genus *Corymbia* despite misleading hypotheses at functional ITS paralogs. Mol. Phylogenet. Evol. 44, 752–764.

Ochieng, J.W., Steane, D.A., Ladiges, P.Y., Baverstock, P.R., Henry, R.J., Shepherd, M., 2007b. The genus *Corymbia* (Myrtaceae) is monophyletic at microsatellite loci. Genet. Mol. Biol. 30, 1125–1134.

Okun, D.O., Kenya, E.U., Oballa, P.O., Odee, D.W., Muluvi, G.M., 2008. Analysis of genetic diversity in *Eucalyptus grandis* (Hill ex Maiden) seed sources using inter simple sequence repeats (ISSR) molecular markers. Afr. J. Biotechnol. 7, 2119–2123.

Ottewell, K.M., Donnellan, S.C., Moran, G.F., Paton, D.C., 2005. Multiplexed microsatellite markers for the genetic analysis of *Eucalyptus leucoxylon* (Myrtaceae) and their utility for ecological and breeding studies in other *Eucalyptus* species. J. Hered. 96, 445–451.

Parra-O, C., Bayly, M., Udovicic, F., Ladiges, P., 2006. ETS sequences support the monophyly of the eucalypt genus *Corymbia* (Myrtaceae). Taxon 55, 653–663.

Payn, K.G., Dvorak, W.S., Janse, B.J.H., Myburg, A.A., 2008. Microsatellite diversity and genetic structure of the commercially important tropical tree species *Eucalyptus urophylla*, endemic to seven islands in eastern Indonesia. Tree Genet. Genomes 4, 519–530.

Peakall, R., Smouse, P.E., 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol. Ecol. Notes 6, 288–295.

Perrie, L.R., Shepherd, L.D., 2009. Reconstructing the species phylogeny of *Pseudopanax* (Araliaceae), a genus of hybridizing trees. Mol. Phylogenet. Evol. 52, 774–783.

Petroli, C.D., Sansaloni, C.P., Carling, J., Hudson, C., Steane, D.A., Myburg, A.A., Vaillancourt, R.E., Kilian, A., Grattapaglia, D., in preparation. A high-density sub-centi-Morgan DArT/microsatellite genetic linkage map for species of *Eucalyptus* based on 2450 markers. BMC Plant Biology.

Poke, F.S., Martin, D.P., Steane, D.A., Vaillancourt, R.E., Reid, J.B., 2006. The impact of intragenic recombination on phylogenetic reconstruction at the sectional level in *Eucalyptus* when using a single copy nuclear gene (cinnamoyl CoA reductase). Mol. Phylogenet. Evol. 39, 160–170.

Poltri, S.N.M., Zelener, N., Traverso, J.R., Gelid, P., Hopp, H.E., 2003. Selection of a seed orchard of *Eucalyptus dunnii* based on genetic diversity criteria calculated using molecular markers. Tree Physiol. 23, 625–632.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Potts, B.M., Wiltshire, R.J.E., 1997. Eucalypt genetics and genecology. In: Williams, J., Woinarski, J. (Eds.), Eucalypt Ecology: Individuals to Ecosystems. Cambridge University Press, Cambridge, pp. 56–91.

Pryor, L.D., Johnson, L.A.S., 1971. A Classification of the Eucalypts. Australian National University Press, Canberra.

Pryor, L.D., Johnson, L.A.S., 1981. *Eucalyptus*, the universal Australian. In: Keast, A. (Ed.), Ecological Biogeography of Australia. Dr W. Junk, The Hague, pp. 499–536.

Ramirez, M.J., 2006. Further problems with the incongruence length difference test: ''hypercongruence'' effect and multiple comparisons. Cladistics 22, 289–295.

Rathbone, D.A., McKinnon, G.E., Vaillancourt, R.E., Steane, D.A., Potts, B.M., 2007. Microsatellite and cpDNA variation in island and continental populations of a regionally rare eucalypt, *Eucalyptus perriniana* (Myrtaceae). Aust. J. Bot. 55, 513–520.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Sale, M.M., Potts, B.M., West, A.K., Reid, J.B., 1993. Relationships within *Eucalyptus* using chloroplast DNA. Aust. Syst. Bot. 6, 127–138.

Sale, M.M., Potts, B.M., West, A.K., Reid, J.B., 1996. Relationships within *Eucalyptus* (Myrtaceae) using PCR-amplification and southern hybridization of chloroplast DNA. Aust. Syst. Bot. 9, 273–282.

Sansaloni, C.P., Petroli, C.D., Carling, J., Hudson, C., Steane, D.A., Myburg, A.A., Grattapaglia, D., Vaillancourt, R.E., Kilian, A., 2010. A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in *Eucalyptus*. Plant Meth. 6, 16.

Shepherd, M., Kasem, S., Lee, D., Henry, R., 2006. Construction of microsatellite genetic linkage maps for *Corymbia*. Silvae Genet. 55, 228–238.

Shepherd, M., Kasem, S., Ablett, G., Ochieng, J., Crawford, A., 2008. Genetic structuring in the spotted gum complex (genus *Corymbia*, section *Politaria*). Aust. Syst. Bot. 21, 15–25.

Steane, D.A., West, A.K., Potts, B.M., Ovenden, J.R., Reid, J.B., 1992. Restriction fragment length polymorphisms in chloroplast DNA from six species of *Eucalyptus*. Aust. J. Bot. 39, 399–414.

Steane, D.A., Byrne, M., Vaillancourt, R.E., Potts, B.M., 1998. Chloroplast DNA polymorphism signals complex interspecific interactions in *Eucalyptus* (Myrtaceae). Aust. Syst. Bot. 11, 25–40.

Steane, D.A., McKinnon, G.E., Vaillancourt, R.E., Potts, B.M., 1999. ITS sequence data resolve higher-level relationships among the eucalypts. Mol. Phylogenet. Evol. 12, 215–233.

Steane, D.A., Vaillancourt, R.E., Russell, J., Powell, W., Marshall, D., Potts, B.M., 2001. Development and characterization of microsatellite loci in *Eucalyptus globulus* (Myrtaceae). Silvae Genet. 50, 89–91.

Steane, D.A., Nicolle, D., McKinnon, G.E., Vaillancourt, R.E., Potts, B.M., 2002. Higher level relationships among the eucalypts are resolved by ITS-sequence data. Aust. Syst. Bot. 15, 49–62.

Steane, D.A., Conod, N., Jones, R.C., Vaillancourt, R.E., Potts, B.M., 2006. A comparative analysis of population structure of a forest tree, *Eucalyptus globulus* (Myrtaceae), using microsatellite markers and quantitative traits. Tree Genet. Genomes 2, 30–38.

Steane, D.A., Nicolle, D., Potts, B.M., 2007. Phylogenetic positioning of anomalous eucalypts by using ITS sequence data. Aust. Syst. Bot. 20, 402–408.

Stokoe, R.L., Shepherd, M., Lee, D.J., Nikles, D.G., Henry, R.J., 2001. Natural inter-subgeneric hybridisation between *Eucalyptus acmenoides* Schauer and *Eucalyptus cloeziana F. Muell* (Myrtaceae) in southeast Queensland. Ann. Bot. 88, 563–570.

Swofford, D.L., 1999. 'PAUP*. Phylogenetic Analysis Using Parsimony (*and other Methods), Version 4. Sinauer, Sunderland, MA.

Thamarus, K.A., Groom, K., Murrel, J., Byrne, M., Moran, G.F., 2002. A genetic linkage map for *Eucalyptus globulus* with candidate loci for wood, fibre and floral traits. Theor. Appl. Genet. 104, 379–387.

Udovicic, F., McFadden, G., Ladiges, P.Y., 1995. Phylogeny of *Eucalyptus* and *Angophora* based on 5S rDNA spacer sequence DNA. Mol. Phylogenet. Evol. 4, 247–256.

Vekemans, X., 2002. AFLP-SURV Version 1.0. Distibuted by the Author, Laboratoire de Genetique et Ecologie Vegetale, Universite Libre de Bruxelles, Belgium.

Walker, E., Byrne, M., Macdonald, B., Nicolle, D., McComb, J., 2009. Clonality and hybrid origin of the rare *Eucalyptus bennettiae* (Myrtaceae) in Western Australia. Aust. J. Bot. 57, 180–188.

Wheeler, M.A., 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. Aust. J. Bot. 54, 17–26.

Wheeler, M.A., Byrne, M., McComb, J.A., 2003. Little genetic differentiation within the dominant forest tree, *Eucalyptus marginata* (Myrtaceae) of South-Western Australia. Silvae Genet. 52, 254–259.

Whittock, S., Steane, D.A., Vaillancourt, R.E., Potts, B.M., 2003. Molecular evidence shows that the tropical boxes (*Eucalyptus* subgenus *Minutifructus*) are over-ranked. Trans. Roy. Soc. S. Aust. 127, 27–32.

Wright, I.J., Ladiges, P.Y., 1997. Geographic variation in *Eucalyptus diversifolia* (Myrtaceae) and the recognition of new subspecies *E. Diversifolia* subsp. *hesperia* and *E. diversifolia* subsp. *megacarpa*. Aust. Syst. Bot. 10, 651–680.

Sansaloni et al. BMC Proceedings 2011, 5(Suppl 7):P54
http://www.biomedcentral.com/1753-6561/5/S7/P54

BMC Proceedings

POSTER PRESENTATION Open Access

Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*

Carolina Sansaloni[1*], Cesar Petroli[1], Damian Jaccoud[2], Jason Carling[2], Frank Detering[2], Dario Grattapaglia[3], Andrzej Kilian[2]

From IUFRO Tree Biotechnology Conference 2011: From Genomes to Integration and Delivery
Arraial d'Ajuda, Bahia, Brazil. 26 June - 2 July 2011

Background

Wider genome coverage and higher throughput genotyping methods have become increasingly important to meet the resolution and speed necessary for a variety of applications in genomics and molecular breeding of forest trees. Developed more than 10 years ago [1], the Diversity Arrays Technology (DArT) has experienced an increasing interest worldwide for it has efficiently satisfied the requirements of throughput, genome coverage and inter-specific transferability for over 40 different plant species to date, including *Eucalyptus*[2] and recently *Pinus* (Dione Alves-Freitas, this meeting). DArT is based on genome complexity reduction using restriction enzymes, followed by hybridization to microarrays to simultaneously assay hundreds to thousands of markers across a genome. Genome complexity reduction for genotyping has now been taken to another level when combined to next generation sequencing (NGS) technologies. Such a strategy has been used for rapid SNP discovery in different organisms [3], and proposed as a way to genotype with RAD (Restriction-associated DNA) sequencing [4]and recently by a similar method generally termed GbS (Genotyping-by-Sequencing)[5]. In this work we assessed the power of the now well established DArT marker platform in combination with Illumina short read sequencing to generate a linkage map for a segregating outcrossed F1 population derived from *E. grandis* BRASUZ1, the donor of the *Eucalyptus* reference genome.

Methods

A segregating population of 89 individuals derived from the intra-specific cross BRASUZ1 x M4D31 was provided by Suzano company. Correct parentage of all individuals was certified by microsatellite genotyping. DNA samples of parents and progeny were processed for the conventional array-based DArT genotyping as described earlier [2]to generate marker data for comparative analysis with the NGS based DArT data. For the sequencing based DArT genotyping two complexity reduction methods optimized for several other plant species at DArT PL were used: PstI_ad/TaqI/HpaII_ad and PstI_ad/TaqI/HhaI_ad with TaqI restriction enzyme used to eliminate a subset of PstI -HpaII and PstI-HhaI fragments, respectively. PstI-site specific adapter was tagged with 92 different barcodes enabling encoding a plate of DNA samples to run within a single lane on an Illumina GAIIx. PstI adapter included also a sequencing primer, so that the tags generated were always reading into the genomic fragments from the PstI sites. After the sequencing run the FASTQ files (full reads of 77 bp) were quality filtered using the threshold of 90% confidence for at least 50% of the bases and in addition filtered more stringently for barcode sequences. The filtered data were split into their respective target (individual) data using barcode splitting script. After producing various QC statistics and trimming of the barcode the sequences were aligned against the reference *Eucalyptus grandis* genome available in Phytozome. The

* Correspondence: carols@cenargen.embrapa.br
[1]EMBRAPA Genetic Resources and Biotechnology - EPqB Final W5 Norte 70770-917 Brazilia DF and Dep. Cell Biology, Universidade de Brazilia – UnB, Brazil
Full list of author information is available at the end of the article

output files from alignment (generated using Bowtie software) were processed using an analytical pipeline developed by DArT PL to produce "DArT score" tables and "SNP" tables. A linkage maps was constructed with JoinMap 3.0 [6]using the microarray-based DArT markers, DArT NGS markers, and 40 microsatellites of known map position as anchors. A parallel analysis exclusively meant to estimate the total number of potential SNPs within the short read tags was carried out using CLC genomic workbench v4.6 software [7] with a minimum read coverage of 6 and minimum variant frequency of 25%.

Results

The microarray-based DArT platform yielded 1,088 high quality markers of which 505 (46.4%) segregated in a 1:1 pseudo-testcross while the remaining 583 (53.6%) segregated 3:1. This relatively lower number of markers when compared to other Eucalyptus mapping populations was expected. Not only it is an intra-specific cross but also involves BRASUZ1, a know self (S1) individual with a lower level of sequence heterozygosity. DArT genotyping using NGS technology yielded 2,835 polymorphic presence/absence markers, almost three-fold the number produced by the microarray platform. Of these, 2,449 markers mapped to the 11 chromosome scaffolds with an average of 222 markers per scaffold, while the remaining 386 markers fall out of the 11 scaffolds, potentially allowing the localization of a fraction of the still unassembled smaller genome scaffolds. In total, an integrated linkage map with 564 DArT markers, 1,930 DArT-NGS and 29 microsatellites was preliminarily built. Furthermore, from the 148 million reads generated (~10.5 Gb), 83.6 million (6.1 Gb) were successfully mapped on the Eucalyptus reference genome. Although a very large number of SNPs can be identified when all reads combined are mapped, only a fraction that displays sufficient coverage allows robust scoring at the individual level. Still, over 1,500 SNPs could be confidently genotyped providing a further advantage of adding co-dominant markers to the already large number of dominant markers obtained.

Conclusion

These initial results show that the combined use of DArT as a robust genome complexity reduction method with optimized barcoded NG sequencing protocol provides at least three fold more dominant markers than the conventional microarray-based DArT method and an additional set of co-dominant SNPs. We are now genotyping a much larger set of distantly related individuals of a training population to be used for Genomic Selection (GS). The possibility of delivering large numbers of both dominant and co-dominant markers with the same platform will enable fitting dominance effect in predictive models therefore increasing the selection accuracy.

Author details
[1]EMBRAPA Genetic Resources and Biotechnology - EPqB Final W5 Norte 70770-917 Brazilia DF and Dep. Cell Biology, Universidade de Brazilia – UnB, Brazil. [2]Diversity Arrays Technology Pty Ltd - PO Box 7141, Yarralumla, ACT 2600, Australia. [3]EMBRAPA Genetic Resources and Biotechnology – Estação Parque Biológico, 70770-910, Brazilia, DF and Genomic Sciences Program - Universidade Católica de Brasília - Brazilia, Brazil.

Published: 13 September 2011

References

1. Jaccoud D, Peng K, Feinstein D, Kilian A: **Diversity arrays: a solid state technology for sequence information independent genotyping.** *Nucleic Acids Res* 2001, **29(4)**:e25.
2. Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D, Vaillancourt RE, Kilian A: **A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus.** *Plant Methods* 2010, **6.**
3. Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS: **SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries.** *Nature Methods* 2008, **5(3)**:247-252.
4. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA: **Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers.** *Plos One* 2008, **3(10)**.
5. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species.** *PLoS One* 2011, **6(5)**:e19379.
6. Van Ooijen J, Voorrips R: **JoinMap 3.0 software for the calculation of genetic linkage maps.** Wageningen, the Netherlands: Plant Research Internation; 2001.
7. Knudsen B, Knudsen T, Flensborg M, Sandmann H, Heltzen M, Andersen A, Dickenson M, Bardram K, Steffensen PJ, Mønsted S, *et al*: **CLC Genomics Workbench.**bio C , 4.6.1.

doi:10.1186/1753-6561-5-S7-P54
Cite this article as: Sansaloni *et al.*: Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. *BMC Proceedings* 2011 **5**(Suppl 7):P54.