LEANDRO GOMIDE NEVES

DETECTION AND MAPPING OF SINGLE FEATURE POLYMORPHISMS (SFP) ON A HIGH-DENSITY SHORT OLIGONUCLEOTIDE ARRAY FOR *Eucalyptus* spp.

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-graduação em Genética e Melhoramento, para obtenção do título de *Magister Scientiae*.

VIÇOSA

MINAS GERAIS – BRASIL

2009

LEANDRO GOMIDE NEVES

DETECTION AND MAPPING OF SINGLE FEATURE POLYMORPHISMS (SFP) ON A HIGH-DENSITY SHORT OLIGONUCLEOTIDE ARRAY FOR *Eucalyptus* spp.

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-graduação em Genética e Melhoramento, para obtenção do título de *Magister Scientiae*.

APROVADA: 05 de agosto de 2009

Dr. Dario Grattapaglia	Dr. Marcos Deon Vilela de Resende
(Co-Orientador)	(Co-Orientador)
Prof. Matias Kirst	Pesq. Danielle Assis de Faria

Prof. Acelino Couto Alfenas

(Orientador)

À minha mãe, Clara, ao meu pai, Ary, e aos meus irmãos Cláudia e Túlio, dedico.

Chamamos "explicação" o que nos distingue dos graus de conhecimento e de ciência mais antigos, mas isso não passa de "descrição". Sabemos descrever melhor – explicamos igualmente pouco como nossos predecessores.

FRIEDRICH NIETZSCHE, em A Gaia Ciência.

AGRADECIMENTOS

São muitas as pessoas que merecem menção por contribuírem para minha formação pessoal e profissional. Em especial, agradeço:

À Universidade Federal de Viçosa e ao programa de Pós-Graduação em Genética e Melhoramento, por apoiarem a realização deste trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos durante o curso.

Ao professor Acelino Couto Alfenas, por me aconselhar e apoiar suportar minhas decisões.

Ao Dr. Dario Grattapaglia, a quem sou grato não só pelos ensinamentos técnicos, mas também, e principalmente, pela amizade e aconselhamentos.

Ao professor Matias Kirst, pela confiança em integrar-me ao seu grupo de pesquisa na Universidade da Flórida, permitindo que a maior parte do componente experimental fosse realizada.

Ao professor Giancarlo Pasquali e sua equipe, por receberem-me em seu laboratório na UFRGS durante a fase inicial do projeto.

Ao Dr. Marcos Deon, por ajudar-me em momentos importantes da análise dos dados do experimento.

Ao amigo, tio e professor Wagner Campos Otoni, por contribuir para o meu desenvolvimento sempre com a boa vontade e o esforço que lhe são característicos.

À Márcia Brandão, por muitas ajudas, sempre com dedicação e paciência.

À Aracruz Celulose S.A. e ao Projeto Genolyptus por fornecerem e auxiliarem na coleta em campo do material biológico utilizado neste trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Empresa Brasileira de Pesquisa Agropecuária (Embrapa) pelo financiamento para a realização dos experimentos.

Àqueles que por muito tempo me acompanham com amizade incondicional, particularmente Thiago Reis, Vinícius, Leonardo, Tiago Soares, Luciano e Douglas.

A toda minha família pelo constante amparo - pais, irmãos, tios, primos e especialmente à minha avó Maria Cornélia.

Aos amigos do Laboratório de Genética Vegetal e da Embrapa, Juliana, Marco, Túlio, Rodrigo, Ediene, Leonardo, Samuel, Marília, César e Carol, por ajudarem em minha estadia em Brasília e amenizarem a rotina de trabalho.

Um agradecimento especial a Danielle Faria, por constantemente ajudar no desenvolvimento deste trabalho, mas principalmente pela amizade.

Aos amigos e colegas do mestrado, pelas horas compartilhadas estudando e descansando em Viçosa, particularmente ao Márcio e ao Ricardo.

A todos os amigos que me ajudaram enquanto estive na Flórida, entre eles Derek, Cynthia, Matthew, Evandro, Carol, Tania, Kathy, Gigi e Chris.

A todos aqueles que, apesar de não citados, participaram significativamente desta etapa, aconselhando, ensinando e encorajando-me nos passos para a conclusão do curso.

BIOGRAFIA

Leandro Gomide Neves, filho de Ary Rodrigues das Neves e Clara Maria de Brito Gomide, nasceu no dia 21 de janeiro de 1984 na cidade de Viçosa, Minas Gerais.

Obteve sua formação acadêmica inicial em Viçosa, concluindo o ensino médio no Colégio Universitário (COLUNI) em 2001 e graduando-se em Engenharia Florestal pela Universidade Federal de Viçosa em 2007. Seu interesse por Genética e Melhoramento iniciou-se após cursar um semestre de sua graduação na *University of Minnesota Crookston* e consolidou-se com a realização do trabalho de conclusão do curso sob a orientação do pesquisador Dario Grattapaglia na Embrapa – Cenargen.

Em agosto de 2007 iniciou o mestrado em Genética e Melhoramento na Universidade Federal de Viçosa, com pesquisa realizada em intercâmbio com a Embrapa – Cenargen e a *University of Florida*, submetendo-se à defesa de tese em agosto de 2009.

CONTEÚDO

RESUMO

NEVES, Leandro Gomide. M.Sc., Universidade Federal de Viçosa, agosto de 2009. **Detecção e mapeamento de polimorfismo de sequência única em uma população segregante de *Eucalyptus* spp.** Orientador: Acelino Couto Alfenas. Co-orientadores: Dario Grattapaglia e Marcos Deon Vilela de Resende.

O gênero *Eucalyptus* apresenta ampla variabilidade genética, relacionada a características economicamente importantes, passível de ser explorada pela integração de métodos clássicos de melhoramento genético e genômica. Apesar dos avanços obtidos pelo melhoramento genético, espécies de *Eucalyptus* ainda estão nos estágios iniciais de domesticação apresentando, portanto, ampla oportunidade de ganhos pela seleção direcional. O desenvolvimento de mapas genéticos de alta resolução, enriquecidos com informação de genes, é uma estratégia com possibilidades de impactar futuras aplicações de melhoramento molecular. No entanto, com exceção de organismos para os quais o genoma encontra-se sequenciado, as técnicas atuais para se localizar genes em um mapa-referência têm mostrado eficiência limitada. Recentemente foi demonstrado em organismos modelo que microarranjos de oligonucleotídeos desenvolvidos para estudos de expressão podem ser utilizados para a detecção de polimorfismos de sequência, gerando marcadores denominados polimorfismos de elementos individuais (SFP, do inglês *Single Feature Polymorphisms*). As sondas no microarranjo detectam sítios polimórficos de SNPs ou "*indels*" nas regiões expressas fornecendo marcadores específicos de genes que, ao demonstrarem comportamento

mendeliano, podem ser mapeados. O objetivo deste trabalho foi aplicar o princípio de descoberta, genotipagem e mapeamento de SFPs em uma população segregante F1 de *E. urophylla* x *E. grandis*. SFPs foram detectados com sucesso utilizando microarranjos de oligonucleotídeos contendo sequências derivadas de genes únicos gerados a partir de sequências-consenso de ESTs de diferentes espécies de *Eucalyptus*. Visto que essa classe de marcadores representa regiões gênicas, mapeamento de SPFs se torna uma abordagem eficiente para mapeamento em larga escala de genes em organismos com genoma não sequenciado. O uso da estratégia de pseudo cruzamento teste permitiu a detecção de marcadores dominantes segregando 1:1 e 3:1 em uma amostra da população de mapeamento. Um mapa genético saturado foi gerado com 884 SFPs sobre um mapa de referência pré-existente de 180 microssatélites, atingindo uma densidade média de um marcador a cada 1,2 cM em 11 grupos de ligação. O uso de um delineamento experimental que permite a detecção de SFPs segregando 3:1, além de aumentar a densidade do mapa e o número de genes mapeados, possibilitou o aumento da qualidade do mapa final. Os resultados também demonstraram que aumentando-se o número de sondas testadas por gene aumenta-se a probabilidade de detectar SFPs no gene alvo e consequentemente de mapeá-lo. Este é o primeiro trabalho de desenvolvimento e mapeamento de SFPs em *Eucalyptus*. A possibilidade de mapear genes em larga escala de forma rápida e acessível permite a condução de análises de co-localização destes genes com QTLs, abrindo espaço para a descoberta de potenciais genes candidatos que controlam esses QTLs para futuramente serem testados em experimentos de genética de associação.

ABSTRACT

NEVES, Lenadro Gomide. M.Sc., Universidade Federal de Viçosa, agosto de 2009. **Detection and mapping of Single Feature Polymorphisms (SFP) on a high-density short oligonucleotide array for *Eucalyptus* spp**. Adviser: Acelino Couto Alfenas. Co-Advisers: Dario Grattapaglia and Marcos Deon Vilela de Resende.

The genus *Eucalyptus* presents a broad natural genetic diversity for economically important traits that could be explored through an integration of classical breeding and genomic approaches. In despite of the advances obtained by classical breeding, species of *Eucalyptus* are still in the early stages of domestication and, therefore, present several opportunities of gain through forward selection. The development of high-resolution genetic maps enriched with gene information is a strategy that can possibly impact future molecular breeding applications. However, except for organisms where genome sequence information is available, current techniques to allocate genes on the reference map have shown limited efficiency. Recently, it has been demonstrated in model organisms that oligonucleotide microarrays developed to assay gene expression can be used to detect sequence polymorphisms, generating markers termed Single Feature Polymorphism (SFP). The probes of the microarray detect polymorphic loci containing SNPs or indels on expressed regions, creating gene-specific markers that can be mapped. The objective of this study was to apply the principle of SFP discovery, genotyping and mapping in a segregating F1 population of *E. urophylla* x *E. grandis*. SFPs were successfully detected using oligonucleotide microarrays with sequences of unigenes generated by sequencing ESTs from different species of *Eucalyptus*.

Since this class of markers samples gene regions, SFP mapping becomes an efficient approach for large-scale gene mapping in organisms with unsequenced genomes. The use of pseudo-testcross strategy allowed the detection of dominant markers segregating 1:1 and 3:1 in a subset of the mapping population. A saturated gene-rich genetic map was generated with 884 SFPs on a previous reference map of 180 microsatellites, with an average density of one marker every 1.2 cM in 11 linkage groups. The use of an experimental design that enables SFPs segregating 3:1 to be detected not only increased map density and number of mapped genes but, more importantly, improved the overall quality of the final map. The results also demonstrated that increasing the number of probes designed per unigene resulted in a higher probability of SFP detection for the targeted gene and, thus, of ultimately mapping it. This is the first report of SFP detection and genotyping for *Eucalyptus*. The possibility to map genes in large-scale in a quick and inexpensive way allow for the conduction of co-localization analysis of these genes to QTLs, creating possibilities to discover potential candidate genes for these QTLs to be tested in association genetics experiments.

1. INTRODUCTION

Development of genetic maps has been a continuous and important step on the study of relevant biological phenomenon and on breeding applications. The types of markers used and the density of the genetic map are major characteristics that have been evolving over time. From phenotypic mutations that followed a Mendelian segregation [1] to the introduction of molecular markers, such as restriction fragment length polymorphisms (RFLPs) [2], greater marker density has been achieved. Among the class of molecular marker to be used, some explore random genomic regions while others explore pre-selected regions, a difference predominantly dependent on the degree of previous genomic information required by the technique. Although this might not seem limiting for model species, it is a crucial point for less studied organisms, which represent the majority of important commercial and ecological species and where genetic maps may be mostly useful.

This is the case of *Eucalyptus*, a genus comprising more than 700 tree and shrub species with original occurrence in Australia and adjacent islands, where some of these species have gained increasing silvicultural relevance to become one of the world most widely planted hardwood tree species. In spite of all its importance, breeding of selected genetic backgrounds is recent and the genera can yet be considered largely undomesticated, showing vast natural genetic diversity susceptible to be explored by forward genetics approaches concomitantly to classical breeding methodologies [3].

Along with an increase in marker density, a desirable advance for practical applications would be to include gene information on such maps. However, current methods for mapping genes, such as single-strand

conformation polymorphism (SSCP), cleaved amplified polymorphic sequence (CAPS), denaturing gradient gel electrophoresis (DGGE), RFLP, microsatellite mining from expressed sequence tags (ESTs) and even SNP genotyping have limited throughput on organisms where genome is not fully sequenced or with high level of polymorphism resulted from outcrossing. As a consequence, few successful applications of these approaches are available for *Eucalyptus*, mostly allowing the mapping of a few dozen candidate genes.

For instance, Gion *et al.* [4] developed SSCP and CAPS markers and incorporated only eight lignin and symbiosis regulated genes to a reference RAPD map. As specific primers have to be designed for each gene, these methods are limited by the need to optimize polymerase chain reactions (PCR) and electrophoresis conditions to detect polymorphism and map the genes. Similarly, RFLP is a laborious process and design of probes on coding sequence to detect polymorphism also lacks throughput and, for example, only 31 cambium-specific ESTs and 14 known function genes were mapped by Thamarus *et al.* [5]. Comparable low-throughput results were also reported for pine species, another non-model outcrossing genera, even after considerable effort was employed on mapping genes [6].

The elevated nucleotide diversity is another characteristic that makes those approaches limited. Yet assuming that microsatellite regions are present on EST sequences and that sufficient flanking region exists to design specific primers, there is a chance that sequence polymorphisms forbid these primers to hybridize to the genome or that they are accidentally designed on an exon-intron border. Moreover, screening, PCR and multiplexing conditions are still required to be optimized for all markers prior to its application. Finally, an analogous problem happens with SNP genotyping, since currently available high-throughput technologies, *e.g.* GoldenGate™ assay, requires 60 bp of SNP-free sequence flanking the SNP that will be genotyped to design specific primers. Therefore, there is a great chance that many diverse coding regions will not overcome this requirement. Such consequence can be observed in the recent study in loblolly pine where only 27 candidate genes were mapped from a pre-selected 384 OPA [7].

On the other hand, DNA microarray technology has demonstrated to be a reliable platform for genomic studies. Since its development from complementary DNA microarrays [8], two key aspects allowed their application to less studied organisms, being (i) the ability to *in situ* synthesize oligonucleotide arrays [9] and (ii) the recent possibility to design custom arrays from some manufactures (as reviewed by [10]). Also, the principle that oligonucleotide arrays could be used to detect genetic differences between genotypes were first speculated [9] and then proved to be possible in the simple genome of yeast [11]. Nevertheless, only five years later it was demonstrated for the more complex genome of *Arabidopsis* by Borevitz *et al.* [12], who also termed this class of polymorphism as Single Feature Polymorphism (SFP).

The principle of SFP detection relies on the disruption of hybridization signal resulted from the hybridization of a sample that contains polymorphic loci between its genome and the reference sequence used to design the probes present on the array. If distinct genotypes are hybridized to the same array, sequence differences between them can therefore be ultimately detected as their hybridization patterns change [11].

Initially, genomic DNA was used as a hybridization source in yeast [11], *Arabidopsis* [12, 13], and rice [14], with the advantage that hybridizing equal amounts of genomic DNA for all samples suggests that every difference in signal is likely to be an SFP. However, due to a higher complexity, larger genomes tend to incorporate more noise when genomic DNA is hybridized to expression array.

Ronald *et al.* [15], Rostoks *et al.* [16] and Cui *et al.* [17] were probably the first authors to expand the SFP principle and hybridize RNA (in fact cDNA or cRNA) to the expression microarray of yeast and barley, with a rationale that polymorphisms present in DNA are transcribed into the messenger RNA and can also reduce the hybridization intensity signal. This approach has several improvements over the previous scenario, mainly allowing for genome complexity reduction that enhances the signal-to-noise ratio, and making it possible to assess gene expression levels and sequence polymorphism genotyping on the same single assay. Moreover, another class of polymorphism

based on differential expression of the gene, called gene expression markers (GEM), is also possible to be obtained [18].

Regardless of the source of polymorphism, generally called SFP unless otherwise stated, genotyping such markers in a segregating family and extracting those with Mendelian behavior ultimately should allow the development of high-density genetic map where the mapped markers represent genes for which the probe sets were originally designed. Consequently, SFP mapping can be summarized as a two-fold task involving the detection of probes that reveal putative SFP and their evaluation as Mendelian markers in a structured mapping population.

A major advance has been made to develop SFPs in model, self-pollinated, homozygous species. In these studies, identification of the putative SFPs has been made based on signal differences between the two original inbred lines. Subsequently, one would search and test for a bimodal distribution of these candidate markers in a sufficiently large mapping population typically made up of recombinant inbred lines or backcross progeny and genotype every individual using as a reference the signal intensity of the inbred parents [18-20].

Considering only the literatures where SFPs were fully genotyped and mapped, the perspectives are that this approach outperforms any previous method develop to incorporate genes on the genetic map of less characterized species. For instance, Singer *et al.* [13] saturated a *Arabidopsis* genetic map with 676 genes using SFP detected from DNA hybridization, but the actual number of genes they could have positioned on the map if that was the interest could have been close to 8,000. A different study on *Arabidopsis* that hybridized RNA was able to map 187 and 968 genes respectively by the genotyping of GEMs and SFPs, with the gene order consistent to the sequenced genome [18]. In the more complex genome of barley, Luo *et al*. [19] reported that they were able to map 1504 and 1523 SFPs when leaf and embryo tissue were separately used for SFP genotyping of a small subset of 30 doubled-haploid lines. On the only study that attempted to genotype microarray markers on outcrossing species, Drost *et al.* [21] mapped 324 SFPs and 117 GEMs segregating 1:1 using a pseudo-backcross progeny of 154 individuals. Very recently, 71 RILs of

hexaploid wheat were genotyped and had 877 SFPs allocated to a genetic map along with 269 microsatellites [22].

Motivated by the potential downstream applications that large scale mapping of genes might have both in gene discovery and molecular breeding, the objective of this study was to apply the principle of SFP discovery, genotyping and mapping in a reference population of *Eucalyptus*. We maximized the experimental efficiency by a combination of screening and selective mapping to localize close to a thousand genes with sequence information derived from EST libraries. We showed that when compared to self pollinating crops, outcrossing species can benefit from a different experimental design based on the pseudo-testcross mapping strategy [23] where information from a subset of the progeny allows clear-cut detection and screening of SFPs.

2. LITERATURE REVIEW

2.1. Genetic maps for biology and breeding applications

After the rediscovery of Mendel's work on inheritance, it was realized by Punnet and Bateson that pairs of alleles might not independently segregate as proposed by Mendel, with those originally present in the parents happening more frequently in the progeny. The phenomenon was further studied by Thomas Hunt Morgan and this observation was attributed to the physical linkage of those loci on the same chromosome, guiding his group to postulate the principles of developing a genetic map. To explain differences in the proportion of alleles in the individuals of progenies from different crosses of *Drosophila*, Morgan speculated that crossing over of the chromosomes could be resulting in such recombination and that the more apart two genes are, the more likely crossing over is expected to occur. Following this assumption, Morgan's student Alfred Sturtevant was the first to develop a genetic map [1].

The implications that genetic maps have had on biology and breeding since those pioneer works are enormous. The first markers used were phenotypic mutations that followed a Mendelian segregation and therefore were mapped in a large F2 population. Perhaps the first application resulting from genetic mapping, and indeed a very interesting one, was that Morgan predicted *Drosophila* to have 7,500 factors (genes) [1]. His educated guess was not only closer to the ~14,000 currently annotated genes [24] because the resolution of their map was not big enough due to a low density of markers.

Marker density was increased with the use of biochemical markers such as isozymes, but it was the development of molecular markers that allowed

genetic mapping to became more widely explored in biology and breeding, particularly after the development of restriction fragment length polymorphisms (RFLPs) [2]. Although these are very informative codominant markers, the development of polymerase chain reaction (PCR) let other classes of molecular markers to be generated, requiring less DNA and being less labor intensive. Random amplified polymorphic DNA (RAPD) were largely explored during the last decade and for many species, such as *Eucalyptus*, it was the first class of marker employed to create genetic maps, mostly because it does not require previous genomic information [23]. Shortly after, microsatellites became the marker of choice for genetic analysis resulting from its elevated polymorphic information content, transferability between different genetic backgrounds, capability of multiplexing several markers in a single reaction and easy detection on capillarity sequencers [25]. Early this decade single nucleotide polymorphisms (SNPs) have arisen as possible gold standard markers due to their large occurrence and evenly distribution on the genome, even though a single SNP is not a multiallelic marker [7].

As the generation of relatively dense maps became possible through the use of molecular markers, their applications on breeding started to be speculated. Genomic regions responsible for controlling a significant portion of the phenotypic variation, termed QTLs (quantitative trait locus) were identified and even major responsible genes were cloned underneath such QTLs [26, 27]. Practical applications of molecular markers, such as genotype identification and discrimination, germplasm characterizations and assessments of genetic diversity have also been reported in the literature (reviewed for forest species by [28]. Nevertheless, a more comprehensive application of such maps to marker assisted selection (MAS) is still lacking [3] and new approaches requiring an even greater marker density have been proposed, as is the case of genome-wide association (GWA) studies and genome-wide selection (GWS) [29, 30].

Only recently, new methodologies that are less labor intensive started to be available for some model species, allowing genotyping of hundreds or thousands of markers at a relatively low cost per individual. In general, they no longer rely on PCRs but explore instead nucleic acid hybridization technologies

to detect sequence polymorphisms. This is the case of DArT (Diversity Arrays Technology) markers, a robust class of molecular marker that does not require previous genomic information but, as a consequence, interrogates anonymous regions for polymorphism [31]; and also of SFP (single feature polymorphisms) markers, which make use of some previous available genome information to design oligonucleotide DNA microarrays and interrogate these pre-selected regions for polymorphisms [12].

1.2. Genetic maps on *Eucalyptus*

The first genetic map was developed on *Eucalyptus* with RADP markers using a pseudo-testcross strategy [23]. RAPD markers were also used to fine map and detect a major gene responsible for rust resistance in a large full sib family of *E. grandis* [32], and preliminary studies involving QTL detection in *Eucalyptus* were also initially developed using this class of molecular markers [33, 34]. The use of RAPD markers make such results essentially not comparable across genetic backgrounds and consequently there was a shift towards the use of microsatellites for genetic and breeding analysis.

The codominant nature of 20 microsatellites developed by Brondani *et al.* [35] was explored to generate an integrated map for a *E. grandis* and *E. urophylla* hybrid. Many more microsatellite markers were further characterized and a much denser consensus genetic map was developed for *Eucalyptus* [25].

A first *Eucalyptus*' map that concerned with the incorporation of gene information focused on those genes related to lignin biochemical pathways with a motivation to propose candidate genes for QTL regions [4]. The single-strand conformation polymorphism (SSCP) technique used by those authors requires gene-specific PCR primers to be developed and these sequences to be amplified through PCR. The polymorphism is then detected by analyzing mobility shifts of the amplified single-stranded DNA fragments on non-denaturing polyacrylamide gel electrophoresis owing to conformational differences between genotypes. Clearly, this approach lacks throughput because optimum PCR amplification and SSCP electrophoresis migration

conditions have to be tested and optimized for each gene before ultimately detecting a mappable polymorphism. Therefore, those authors report the positioning of only eight genes on a RAPD map of *Eucalyptus*.

Thamarus *et al.* [5] made a more comprehensive effort to map genes on *E. globulus* pedigree essentially using the RFLP technology using probes for known genes. Probes were either already available for some of these genes from genomic and cDNA libraries developed by earlier studies, and some probes were derived for specific sequences by PCR amplification on genomic DNA. RFLP is a laborious methodology and those authors were able to map only 31 cambium-specific expressed sequence tags (ESTs) and 14 known function genes.

Another way of confidentially mapping genes is the *in silico* screening of EST libraries for microsatellite markers. This is a relatively efficient approach and markers keep the advantages intrinsic to microsatellites (*et al.* transferability, data quality, multiallelic). The method obviously requires microsatellites to be present in the ESTs and primers are developed flanking these regions, assuming that enough sequence is available for the EST to accommodate primers on both sides of the microsatellite. Regardless of these limitations, the amplification of the microsatellite can simply not work because the primers might have accidently been designed on an exon-intron border or because there is a sequence polymorphism between the primer sequence and the sample, which will preclude the hybridization of the primer to the predicted site. Furthermore, all the putative markers have to be screened for segregation and as a consequence high-throughput is not achieved, although one can map many genes depending on the EST library coverage and level of polymorphism of the pedigree [36].

Detection of SNPs on the sequence of genes can also successfully add few candidate genes on a reference map and different strategies exist to genotype such detected SNPs. In an exploratory study, our group resequenced the genomic region of some candidate ESTs and genes by Sanger sequencing and used the haplotypes created by two neighbor SNPs to genotype a subset of the mapping population and anchor some genes to a previous microsatellite

map [37]. Similar to other low throughput technologies, the technique employed allowed mapping only a few genes. Other methods of SNP genotyping via high-density platform, such as the Illumina GoldeGate assay can also be used to genotype SNPs previously discovered in target genes [38], but its efficiency on highly polymorphic species of *Eucalyptus* with a nucleotide diversity above 1% has yet to be tested.

1.3. Microarray technology

With the development of genome sequencing projects of model species, such as yeast (*Saccharomyces cerevisiae*) and *Arabidopsis thaliana*, the physical structure of the genome became available for specific reference genotypes [39, 40]. Concomitantly, it was also realized that other functional information would be required for the understanding of the biology behind the sequences, particularly the role of the predicted genes in the organism. Although this is a very complex and yet unresolved question [41], understanding the transcription patterns of genes, for example the relative difference in expression between tissues or biological treatments, became possible with the advancement of DNA microarray technology [8].

Shortly, DNA microarray is an evolution of the classical southern blot procedure, where, instead of a few genes, it is possible to assign the expression level of every single gene of the organism. Another difference is that the DNA sequences representing the genes are now covalently bound to a small rectangle glass matrix. For example, Agilent's microarray (Agilent technologies Inc.) measures around 7.5 cm by 3 cm, similarly to a microscope slide, and can fit up to one million features. On this surface, genes are allocated in a XY coordinate plane that allows the system to track back the signal to its relative gene [10].

Although whole genome sequence information is available for more than 180 organisms ([42], accessed on July 19, 2009), allowing the design of high quality probes to be included on the microarray, this information was not available at the birth of the technology nor is available today for all species.

Besides, sequencing a unique genotype obviously does not represent all polymorphisms present in the species, which is particularly relevant for species where interespecific hybridization is commonly used in breeding. This raises the question of where to extract the sequence information that will represent the genes and populate the microarray.

When DNA microarray technology was first developed, the pioneering work of Schena *et al.* [8], carried out even before the *Arabidopsis*' genome was sequenced, used cDNA representing 64 genes spotted on glass by an automatic robot to represent the genes. In the forestry scenario, Kirst *et al.* [43] working with *Eucalyptus* also used spotted cDNA comprising 2,608 genes to populate the microarray, selecting them from a library of ESTs.

An interesting characteristic of cDNA microarrays resides in the fact that cDNAs are long fragments representing the transcribed part of the gene, being therefore robust to variation in signal due to length polymorphisms and SNPs between the sample and the library used to obtain the cDNAs. This means that if one is simply willing to measure gene expression levels in highly polymorphic species, the results would theoretically be less influenced by such class of polymorphism [44]. Nevertheless, spotted cDNA have the drawbacks of being labor intensive to produce, presenting several significant sources of variation that might confound the expression analysis and being more potentially affected by cross-hybridization problems [10].

The concept of high-density, *in-situ* synthesized, oligonucleotide microarrays arose to overcome some of these disadvantages associated with cDNA microarrays [9]. Instead of pre-synthesizing the sequences that will be present in the microarray, generally by PCR of the cDNA clones, the DNA sequences representing the genes are synthesized straight on the glass surface that will constitute the microarray, allowing a better control of the quantity and quality of the DNA present on the microarray. The design of these genes representations require sequence information for the species to allow the selection of several short probes (typically between 25 and 60 bases long), based on uniqueness scores, GC content and other empirical criteria to sample the 3' end of each open reading frame (ORF), defining a probeset for a

particular gene. In other words, all the desired target genes would be represented by a probeset. One of the first versions of this format was a probeset of 20 probes, each one 25 bases long, constituting the Affymetrix GeneChip Array [9]. Wodicka *et al.* [9] were also the first to raise the possibility of using microarrays to detect genetic differences between genotypes.

A final significant evolution in the microarray technology came with the possibility to customize the probes that will be present on the array. This gives the researcher mobility to choose the regions of the genome one wants to consider most and the degree of coverage and redundancy allowed for each probeset (*i.e.* the number of probes per probeset). Longer probes can also be customized, up to 85-mer, although this size is constantly being improved. Such customization process allowed the construction of high-density *in situ* synthesized oligonucleotide arrays based on EST library information for species with unsequenced genome (reviewed by [10]).

Microarray analysis is prone to confounding sources of variance biasing the expression signal information. Appropriate experimental designs and statistical analyses have to be considered to account for these issues, allowing for the most accurate estimation of levels of gene expression. Even after controlling for these variations, inherent “noise” is a peculiar characteristic of microarray data, making confidence in the estimated parameters an important information extracted from appropriate statistical analysis. Kerr and Churchill [45] were some of the first authors to ask such questions and connect microarray experiments to classical experimental theory. For example, they illustrated the sources of variance commonly seen in a microarray experiment. They proposed the analysis of variance (ANOVA) as a way to separate biological variation from technical confounding effects and suggested new experimental designs based on classical experimental theory.

In the early days of large scale expression analysis via microarrays, one contrasting treatments (*e.g.* pathogen infection; different mutated strains) would have its transcription level compared to a common control (*e.g.* non-infected genotype; wild strain). Then, RNA samples of the treated genotype would be labeled with red and those from the control sample with a green dye, mixed and

hybridized together to the array. Based solely on the relative red to green intensity signal a gene would be called differentially expressed (see picture 1 of [46].

It was quickly recognized that microarrays experiments could be designed in a factorial fashion and mixed-model analysis of variance could be used to test for significance in the main and interaction effects [46, 47]. Furthermore, the sources of variation in microarray data became largely understood and could be considered in a more appropriate experimental design, incorporating the principles of randomization, replication and blocking. For example, every microarray or slide is said to be an incomplete block with space for two treatments (the two dyes) and the genes or probes considered to be the fundamental experimental units [45, 48].

During the laboratorial part of a microarray experiment, several steps might introduce variation for specific genotypes that are neither biologically meaningful nor inherent to the microarray technique. Ideally, tissue collection, RNA extraction and every other subsequent reaction (labeling, hybridization, washing and scanning) should be performed in the same moment under the same exact conditions. However, when the experiment requires dealing and collecting samples from hundreds of genotypes this obviously becomes an unfeasible task and even more variation is incorporated. These examples of experimentally introduced variation that are not related to the biology, sometimes called “batch effects”, are a subset of variation that can be controlled by using appropriate experimental practices, as shown by Yang *et al* [49]. These authors presented comparable experimental data from four different laboratories to discuss that if batch and biological effects are confounded, it is impossible to extract reliable conclusions out of the data. As a result, reproducibility is compromised and comparisons between laboratories are inappropriate.

In a classical microarray experiment there are four main sources of variation and their interactions. Array or slide effect represent situations where the overall signal is variable for particular arrays. Dye effect refers to one dye being consistently different than the other, evidenced when using Cy3 (higher

signal) and Cy5 (lower signal) dyes. Genotype (or treatment) effect indicates a case where particular genotypes have an overall signal difference from the remaining genotypes for the genes under study. At last, gene effect will show circumstances where certain genes show higher or lower signal compared to other genes [48]. Although it is believed that all these sources will contribute to variation in the data, the higher quality of the *in situ* synthesized oligonucleotide array tends to reduce their relative influence, for example reducing problems of gene effects, since equal amounts of DNA are synthesized on every microarray spot.

When measuring gene expression level, the second order interaction of genotype x gene is the effect of interest, indicating genes that have their signal intensities varying according to the genotype. Other combinations of second order interaction are due to technical variation and the higher third and fourth order interactions are assumed to be irrelevant [48]. Finally probe effects are present when a gene-level analysis is performed in oligonucleotide arrays, and significant genotype x probe interaction indicates genes with putative single feature polymorphism (see below).

To estimate components of variance for every effect as well as significance for effects of interest, a mixed-model analysis of variance is usually suggested, with slide effect and its interactions being considered random and the remaining effects considered fixed, except for the mean and the error, which are always assumed to be fixed and random, respectively [50, 51].

A logical implication of identifying the sources of variation that most likely introduce variation in the data is that they can be considered in experimental designs. The first design used was the reference design. Here, one reference sample (a genotype or treatment) is hybridized to every array with each of the other treatments at a time, allowing pair-wise comparisons between the treatments and the reference. One of the problems of this approach is that genotype effect is completely confounded with dye effect, which may create further inaccurate results. Kerr and Churchill [45] proposed the loop design as a practical alternative to the reference design, where a reference sample is no longer used and the samples of interest are independently labeled with both

dyes. The loop starts when the first array is hybridized with the repetition two of the sample 1 and the repetition one of the sample 2; followed by the second array hybridized with the repetition two of the sample 2 and the repetition one of the sample 3; and the loop is continued until the last array, which is hybridized with the second repetition of the sample *n* and the repetition one of the sample 1 [52].

When compared to the reference design, the loop design is a robust experimental design for two-color microarrays that increases the number of observations for the samples of interest and has no confounding effects. It also provides degrees of freedom for estimating the error variance [48]. A final remark about the loop design is that it requires every sample to be labeled twice, which might increase the costs of the experiment without increasing the number of biological repetitions. Therefore, adaptations of this method to accommodate the experimental budget is possible if one assumes that the resulting confounding effects (if any) will not influence the analysis [53].

1.4. SFP technology

With the development of oligonucleotide arrays, the idea that sequence polymorphism would affect hybridization intensities, allowing for the detection of genetic differences between genotypes, was first speculated and later proved in yeast [9, 11]. The application of this principle to species with more complex genomes was only proved to be possible five years later by Borevitz *el al.* [12], to the yet relatively simple genome of *Arabidopsis*. Their work triggered a series of studies using DNA microarrays to detect polymorphism and this type of polymorphism was termed Single Feature Polymorphism (SFP).

Hybridization of genomic DNA was first used in studies involving SFP detection, however always on organisms with simple genome [11-14]. As larger genomes tend to have a greater proportion of repetitive sequences that can incorporate bias to the analysis, soon there was a change towards the hybridization of RNA for more complex species [15-17]. Concomitantly, new statistical methods to detect probes behaving as SFPs had to be developed, as

differential expression between genotypes influence the hybridization signal and, therefore, complicates the detection of SFPs. For example, Wang *et al.* [20] demonstrated that the classical method developed by Winzeler *et al.*[11] is not the most efficient when expression is incorporated. Other advantages of using RNA rather than DNA are also reported on the literature, particularly the possibility to use the data for both expression analysis and genotyping and to obtain another class of polymorphism based on differential expression, called gene expression markers (GEM) [18].

While SFP is an inherent characteristic of a single probe of a probeset, a GEM is expected to change the expression profile of all the probes of that particular probeset. This can be explained by analyzing the genetic basis of each type of polymorphism. An SFP has its signal altered due to a SNP, insertion or deletion, or even a polymorphism generated during mRNA processing (e.g. alternative splicing and polyadenylation) [16]. On the other hand, the constitutive signal variation of the probes belonging to a probeset affected by a GEM is primarily a consequence of *cis* and *trans*-acting regulators of gene expression [19].

The position and number of polymorphisms present in the region explored by the probes play a major role on SFP detection. It has been reported that a greater chance of SFP detection occurs when the polymorphism is present on central regions when compared to the border of the probes [15, 17] and for cases where multiple polymorphisms exist [16]. The number of polymorphisms present along a probeset may also influence the SFP detection depending on the statistical method employed and most of the methods developed so far assume that there is one or only a few polymorphic probes per probeset [15, 54].

Although a major advance has been made to develop SFPs in model, self-pollinated, homozygous species, recent works have proposed its applicability to other less studied and complex organisms, such as poplar [21] and wheat [22]. Not all studies published on SFP ultimately genotyped a mapping population with those markers, but instead several of them were restrict to the detection of SFPs [14, 54, 55]. On the other hand, considering

those where efforts were made to genotype the detected SFPs, number of mapped markers varies from hundreds to thousands of SFPs [13, 18, 19, 21, 22].

2. OBJECTIVES

The general objective of our work was to apply the SFP technology to large-scale mapping of genes in a reference map of *Eucalyptus*. This general objective can be broken down into four more specific objectives:

i) Design an *in situ* oligonucleotide microarray for *Eucalyptus* from previously available EST resources;
ii) Identify and genotype SFPs on a subset of a mapping population using the pseudo-testcross strategy;
iii) Develop a saturated gene-rich genetic map of *Eucalyptus* based on SFP genotyping;
iv) Propose an optimum SFP screening and mapping approach that could be used to maximize the number of genes mapped for other less genomically characterized outcrossing plant species.

3. MATERIAL AND METHODS

3.1. *Eucalyptus* pedigree selection

An interespecific cross between *E. urophylla* (U15) and *E. grandis* (G38) was selected for this study. The whole family comprising 250 individuals was planted in southern Brazil (State of Rio Grande do Sul) at Aracruz S.A. The experimental design used was in single tree plots with five blocks so that each individual was clonally replicated and five ramets were available for the study. A linkage map with 220 microsatellites markers was available for 188 individuals of this same family (Mamani *et al.* unpublished).

From the collected field grown trees, 28 biologically replicated individuals and 134 unique individuals were available for the subsequent analysis, all of them matching the microsatellite data set and with parentage and clonal confirmation after analysis of six highly informative microsatellites (Additional file 1).

To optimize experimental costs while extracting the best linkage information, 68 individuals were selected based on the distribution of recombination breakpoints observed in the microsatellite dataset using the selective mapping approach implemented by the software MapPop 1.0 [56]. As the mapping data for the microsatellites was derived from a cross between two heterozygous individuals, and the software only deals with dataset derived from crosses between inbred parents with known linkage phase, the selection had to be carried out for each parental map separately. From the 68 individuals selected from each parental map data, 41 could be selected that overlapped

between the two selected datasets and the remaining ones were taken in equal proportions from the non-overlapping to equally represent each parental map.

3.2. Microarray design

A *Eucalyptus* transcriptome custom array was ordered to perform the present study. Sequence information was available from the Genolyptus interespecific unigene set generated from ESTs derived from four species (*E. grandis*, *E. urophylla*, *E. globulus*, *E. pellita*) involving a total of 21,428 unique sequences [57].

The 21,428 unique sequences were submitted through Agilent's eArray software tool to generate a total of 214,218 25-mer probes at a rate of 10 probes for every sequence. Shortly, the algorithm selects the 10 best sequences of 25 bases for each consensus according to an appropriate GC content, melting temperature and cross-hybridization probabilities to optimize hybridization conditions for the designed microarray. Also, a note with base composition (BC) and non-self perfect match (NSPM) scores is created as a quality index for every designed probe.

The microarray used for the SFP genotyping was developed in the same way done to screen informative molecular markers. A preliminary step involved screening a large number of probes for the largest number of genes possible within the affordable microarray format. From the 214,218 designed probes, we attempted to select five high-quality probes for each unigene. For a random subset of 1,308 unigenes 10 probes were selected while for 2,868 unigenes there were less than five high-quality probes available. The final screening array had 103,000 probes representing 20,726 unigenes with a variable distribution of probes per unigene (Additional file 2). Twenty-six negative control probes were included on the array for background expression normalization with their sequences reported elsewhere [58]. The screening step was carried out with 14 arrays in a 2 x 105K Agilent slide format with a set of 28 progeny individuals with two biological replicates. From the results of the screening step a smaller array containing only selected expressed probes, *i.e.* with signal above a given

threshold, was designed to be used to hybridize the remaining progeny set previously sampled by selective mapping based on the microsatellite linkage data.

The microarray used for the full scale progeny genotyping included 43,777 probes selected from the results of the screening step. These probes represented 15,698 genes and were selected based on a set of criteria (see results). As a result, the full scale genoypying experiment used 17 Agilent customized arrays in a 4x44K slide format. The 26 negative control probes were again included on the array.

The sequence information for both screening and genotyping array is available upon request.

3.3. Tissue collection, DNA and RNA preparation and expression profiling

Differentiating xylem and expanded leaves was collected from each tree during a period of four consecutive days respectively for RNA and DNA extraction. An area of approximately 20 cm x 10 cm had the bark removed and the exposed tissue was collected by scraping this area (Additional file 3). The trees were 54 months old, growing under standard silvicultural conditions. Tissue collection was carried out in January 2008 during the active growing season to maximize transcript abundance. Immediately upon collection, tissues were stored in 50 mL sterile tubes under dry ice until they were lyophilized for further transportation and storage at room temperature. The parental trees were not planted on the area and could not be sampled. Nevertheless, frozen leaves were available for DNA extraction.

For microsatellite genotyping for parentage and identity verification, genomic DNA was extracted from leaves following standard procedures [23] and diluted to 2 ng/μL. A multiplex PCR reaction was carried out using the 5X Multiplex PCR Kit (Qiagen, Valencia, CA, USA), with the following volume modifications: 2.5 μL of PCR Master Mix, 0.5 μL of Q-Solution, 0.4 μL of RNase-free water, 10^{-6} μmols (x 6) of fluorescent labeled primers and 1 μL of DNA, for a total reaction volume of 5 μL. All other steps of the amplification

reaction followed the manufacturer recommendations. Electrophoresis was carried out on an ABI 3700 sequencer (Applied Biosystems, Foster City, CA, USA). Genotyping analyses were carried out using GeneScan 3.7 and Genotyper 3.7 (Applied Biosystems).

Total RNA extracted [59] from xylem tissue samples was treated with RQ1 RNase-free DNase (Promega, Madison, WI, USA), purified in mini spin columns (RNeasy Plant Mini Kit, Qiagen) and had the quality visually checked on agarose gels. Between 150 and 200 mg of lyophilized tissue was used for RNA extractions. Concentrations varied considerably according to tissue quality but absorbance ratios of 260/280 nm and 230/260 nm were generally between 1.8 and 2.2, as measured on a NanoDrop (NanoDrop products, Wilmington, DE, USA). The Two-Color Quick Amp Labeling Kit (Agilent Technologies, Santa Clara, CA, USA) was used to synthesize complementary RNA (cRNA) taking advantage of the low RNA input required. Manufacturer's protocol was followed except for dividing all reagents' volume by two to reduce costs. All samples yielded enough labeled cRNA for hybridization as recommended by the manufacturer. Samples were hybridized at the Interdisciplinary Center for Biotechnology Research (ICBR) of the University of Florida following Agilent's protocol, except for lowering the hybridization temperature to 55 ^{o}C to compensate for the shorter probes. Randomization of all laboratorial steps was a rule whenever possible.

3.4. Microarray experimental design

A loop design [48] was adopted both in the probe screening and genotyping experiments. From the two biological replicates present in the probe screening step, one was labeled with Cy3 dye while the other with Cy5. We opted for this approach aware of the confounding effects of dye and individual to reduce labeling costs. As biological replicates were not available for the full scale genotyping experiment, mRNA from each one of the 68 individuals was labeled with both dyes providing only technical replication.

3.5. Selection of informative SFPs in the probe screening experiment

Log2 transformed, quantile-normalized data for the 20,726 unigenes with expression profiled in 28 biologically replicated individuals (two independent replications) were analyzed by a mixed-model ANOVA with two subsequent steps adapted from Wolfinger *et al.* [50] and Rostoks *et al.* [16]. The first linear model is:

$$y_{gjk} = \mu + A_j + D_k + AD_{jk} + \varepsilon_{gjk}$$

Where, y_{gjk} is the log2, quantile-normalized measurement from the *g*th gene (*g* = 1,…, 20726), *j*th array (*j* = 1,…, 28), and *k*th dye (*k* = 1, 2). Note that the two chambers present on the microarray slide were considered unique arrays because they were independently hybridized and that two distinct repetitions of a genotype were hybridized to the same chamber in a two-color design. Moreover, μ is the overall experimental mean, A is the global main effect for arrays, D is the global main effect for dies, AD is the global interaction effect of arrays and dies, and ε the normal error. Dye was considered as a fixed effect whereas array and its interaction as random effects, except for the mean and error that were always fixed and random effects, respectively. This first model globally normalizes the data and reduces computational time. The residual from this model for each individual on every observation, referred as r_{gipj}, was used as the input for a second gene specific linear model, which was:

$$r_{gipj} = \mu + G_i + P_p + A_j + GP_{ip} + \gamma_{gipj}$$

With G being the effect of genotypes (*i* = 1,…, 28), P being the effect of probes (*p* = varies according to gene from 1 to 10), GP being the interaction effect of genotypes and probes, and γ being the random error. Genotype, probe and their interaction were considered as fixed effects, array as random effect, and mean and error respectively as fixed and random effects. This second model was fit on a gene-by-gene analysis and, thus, all values were indexed at the gene level. In other words, an ANOVA was performed 20,726

times to fit this model on each gene's dataset. The ANOVAs were implemented on SAS 9.1 using the *Proc Mixed* statement.

F tests for the genotype by probe interaction effect of each gene were performed and the probabilities were corrected for multiple testing using a modified false-discovery rate on Q-value 1.0 [60]. Significance of this test indicates genes with putative SFPs. A further analysis was then carried out within genes to identify probes that revealed SFPs by clustering the averaged, log2 transformed, quantile-normalized data for the 28 individuals into two clusters as described below. Finally, a segregation analysis by a chi-square test and modified normal deviate were performed to select distinct clusters corresponding to individual probes that segregated 1:1 or 3:1 using the same stringency described below, resulting in SFP candidates.

3.6. Full scale SFP genotyping and map construction

A total of 96 (28 + 68) F1 interespecific full-sib individuals had their expression profiled for 43,777 probes. The raw median signal intensity for each probe of all 192 hybridizations were log2 transformed and quantile-normalized using the Affy package on R [61]. The averaged normalized data for the 96 individuals was used for the simultaneous identification and genotyping of putative SFP through a *k*-means clustering analysis modified by Drost *et al.* [21] from Luo *et al.* [19]. Shortly, the learning algorithm allocates the signal intensity of each genotype into two distinct clusters on a per-probe basis. [21]. The progeny size (n=96) allowed the identification and distinction of probes segregating 1:1 and 3:1 after a stringent chi-square was used ($\chi^2_{d.f.=1}$<3.84, *P*>0.05) and probes that did not follow this expectation were excluded from the subsequent analysis. Based on the mean and standard deviation calculated for each cluster, the probability that the individual assigned to one cluster is not a member of the other was calculated using the modified normal deviate $z_i = (x_i - m_j)/s_j$, where x_i is the signal intensity of an individual from cluster *i* and m_j and s_j are the mean and standard deviation of the cluster *j* [19]. Individuals with zi ≤ 1.96 have a probability equal to or greater than 5% to

belong to the other cluster and, thus, a greater chance of being ambiguously genotyped. To reduce miscall rate they were scored as missing data and we only kept probes with less than 10% of missing data (*i.e.* at least 86 progenies accurately genotyped considering this stringency). Selected probes segregating 1:1 (pseudo-testcross) and 3:1 (dominant F2) received the acronym of BC and F2 reflecting their segregation configurations.

When a probeset had multiple probes selected through the pipeline described above, an empirical iterative method was developed to select the best possible probe for the probeset as follows: (i) the probe with less missing data was selected; (ii) probes revealing F2 SFPs were preferentially selected over BC; and (iii) probes with the greatest gap between clusters mean were selected. After applying these three criteria only one probe per probeset was selected for mapping. Clustering analyses were implemented on SAS 9.1 using *Proc Fastclus* and filtering steps were implemented on JMP 7.0 (SAS Institute, Cary, NC, USA).

Progeny individuals were coded using a binary coding (lm or ll for 1:1 SFPs or h- or kk for 3:1 SFPs) for mapping according to the assignment to each cluster for each probe. SFP genotyping information for the 96 individuals was consolidated to a 181 microsatellite dataset. JoinMap 3.0 [62] was used for map construction with the following parameters: population type CP; grouping at LOD> 7; recombination fraction ≤ 0.4; ripple value = 1; jump in goodness-of-fit threshold (the normalized difference in goodness-of-fit chisquare before and after adding a locus) = 5 ; Kosambi mapping function. Marker ordering with Joinmap was carried out by simulated annealing, excluding markers that contributed to unstable marker orders in the first two ordering rounds to yield framework maps. The microsatellites anchoring the SFP markers have been previously mapped using both MapMaker and Joinmap and thus provided a reference framework map ordering on which the large number of SFP markers could be mapped.

4. RESULTS

4.1. Analysis of expression and microarray design

To develop an optimum approach for mapping genes in outcrossing species by SFP genotyping, we generated a custom *in-situ* synthesized oligonucleotide array. A microarray was first designed for the probe screening step and only a relatively small set of individuals was hybridized to an array containing the whole unigenes generated during the Genolyptus Project. Following the probe screening step, a new microarray containing only pre-selected probes was used for the full scale genotyping experiment to generate the segregation data for map construction. The screening array had 103,000 high-quality unique probes representing 20,726 unigenes, from now on named genes for simplicity. On average, each gene had five non-overlapping probes designed to randomly interrogate its sequence. For 1,037 genes ten non-overlapping probes were designed to evaluate if a larger probability of revealing SFPs could be arrived to by using a larger number of probes in the probeset.

To correct for global variation between arrays we quantile-normalized the data, Log2 transformed and averaged the expression signal for each repetition, resulting in 103,000 signal variables for each genotype. On this probe screening dataset derived from the hybridization of 28 individuals expressed probes were identified. An individual was considered not to be expressing a particular mRNA when the signal for a particular probe was below 2.297 rfu (relative fluorescence units), which represented 90% of the signals for the 26 negative control probes present on the array, excluding four control probes that were clearly expressed (outliers). When 25 or more individuals (90%) had their signals below that

threshold the probe was considered to be not expressed. This stringency only excluded SFPs if they had severe segregation distortion.

From the initial set of 103,000 probes, approximately half (51,661) were considered to be expressed, an expected low result that reflects the relatively complex tissue and species composition of the EST dataset that was assembled in the unigene set used to design the probes. For example, flower tissue was sampled and ESTs sequenced were represented in the unigene dataset. Transcripts specific to this tissue would most likely not be present in the xylem tissue used in this study. Nevertheless, expressed probes accounted for 78% (16,163) of the genes represented on the array, indicating that the number of expressed probes per gene varied considerably (Figure 1). Interestingly, even though the full transcripts would be theoretically present, for only 3,622 genes all the probes initially designed were consistently expressed, which might be a result of polymorphisms between the unigene probe sequence and the transcripts expressed by the parents (Figure 1). On the other extreme, almost the same amount of genes (3,513) had only one probe expressed and although this might partially be a consequence of the *ad hoc* expression cutoff threshold used that, if changed, could have excluded these probes or included more probes, it could also be the influence of developing microarray from complex EST libraries (Figure 1). For instance, it could be that several probes (four in this case) had a polymorphism between the unigene probe sequence and the transcripts expressed by the parents.

To fit the expressed probes in a 4 x 44K Agilent slide format, we further reduced the 51,661 expressed probes by removing the probes with lowest mean and standard deviation intensity signal, based on the premise that these were likely not to reveal SFPs. The microarray designed following the probe screening experiment thus comprised 43,777 probes interrogating 15,698 distinct genes, with a variable number of probes per probeset. This array containing selected probes was used to genotype the 68 further individuals chosen based on the Selective Mapping approach. This procedure identified the most informative individuals in terms of complementary recombination breakpoints from each parent, optimizing the linkage information to be extracted from these individuals.

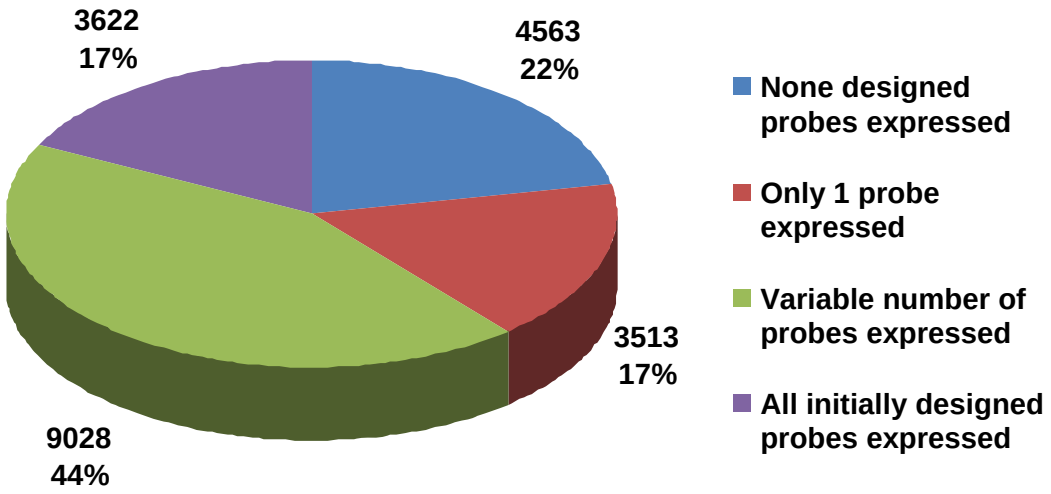


Figure 1: Number of probes expressed per probeset compared to what was initially designed on the screening microarray. A probe-level analysis of expression for the 20,726 unigenes considered probes not expressed when more than 90% of the individuals had signal below a background threshold.

4.2. Simultaneous detection and genotyping of SFPs in progeny data

The averaged, Log2 transformed, quantile-normalized data for the 96 individuals sampled were analyzed together to generate a gene-rich map of *Eucalyptus* based on SFPs anchored to microsatellites. Initially, the methods developed by Drost *et al*. [21] for genotyping SFPs in the highly heterozygous genome of poplar were applied to our dataset. Different from that work our dataset only involved progeny individuals so that the analysis was used to simultaneously identify and genotype SFPs. Working on a per-probe basis, the signal intensity of each individual offspring was assigned to either one of two distinct clusters using the *k*-means clustering learning algorithm. Using a chi-square test, probes showing a 1:1 pseudo-testcross segregation were selected. As SFP segregate as dominant markers probes segregating 3:1 were also selected. Although theoretically such probes could be segregating 1:2:1 (Figure 6D) the separation of the signal intensity into three clusters would be challenging and could result in higher proportions of genotype miscalls due to the expected overlap between signals from these classes. At a $\chi^2_{d.f.=1} < 3.84$ (*P*

> 0.05), 65% (28,304/43,777) of the probes displayed a Mendelian segregation ratio, with 12,148 segregating 1:1 and 16,156 segregating 3:1 (Table 1).

Table 1: Number of probes selected after applying the SFP detection and mapping pipeline to the genotype dataset of 96 F1 individuals of the *E. urophylla* x *E. grandis* pedigree. Unigenes derived from a consensus sequence involving ESTs from different species are called Contigs, while singletons are listed by species. Number of unigenes represented is shown between parentheses.

EST class	Total	Chi squared		zi normal deviate		Grouped		Mapped	
		1:1	3:1	1:1	3:1	1:1	3:1	1:1	3:1
All	43777 (15698)	12148 (7764)	16156 (10364)	2586 (2132)	3063 (2583)	1088	760	457	427
Contig	22598	6668	8139	1407	1587	570	435	252	271
E. urophylla	2071	567	778	132	134	58	40	20	24
E. grandis	10073	2445	3846	496	735	221	134	98	67
E. globulus	3620	1044	1361	216	230	93	62	34	29
E. pellita	2133	550	835	144	146	66	36	24	12
Mixed species	3282	874	1197	191	231	80	53	29	24

The degree of separation between clusters was measured by calculating the probability of individuals assigned to one cluster being a member of the other cluster through a modified normal deviate *zi* (see methods). Individuals with *zi* equal to or smaller than 1.96 ($P \geq 0.05$) are likely to overlap with the other cluster and were assigned as missing data to avoid genotype miscalls. Whenever an excessive number of individuals (more than 10%) were considered as miscall the probe was discarded. This selection step of the SFP detection pipeline is the most effective one as only 5,649 probes, representing altogether 4,300 unigenes were selected. A few genes (255) had more than two probes selected; an important result if one considers that the selection of fewer probes per probeset indicates detection of SFPs rather than GEMs. At this stage, the number of detected SFPs segregating 3:1 was slightly greater than that segregating 1:1, a pattern that was inverted after assigning these markers to linkage groups and stabilized after mapping and ordering them (Table 1).

Although ESTs that grouped into contigs are usually referred to be of higher quality than those that did not group (singletons), we found only a borderline significant difference in the ability of probes derived from these two different types of unigenes to reveal SFPs ($\chi^2_{d.f.=1}$ = 4.94 *P* = 0.0262) (Table 2). Although significant, this result indicates that singleton unigenes are a useful source of probes for SFP discovery.

Table 2: Association between the source of probes (contigs vs. singletons) and the rate of SFP detection.

	Contig	Singleton	Total
SFP Detected	2994	2655	5649
SFP Not detected	19604	18524	38128
Total	22598	21179	43777

$\chi^2_{d.f.=1}$ = 4.94 *P* = 0.0262

Finally, to avoid redundant information the 5,649 probes revealing putative SFPs (2,586 segregating 1:1 and 3,063 segregating 3:1) were filtered to keep only one best SFP per gene for mapping. The selection criteria involved three steps (see methods). The first (pick probes with less missing data) intrinsically gives priority to probes with individuals well assigned to clusters; the second selects probes segregating 3:1 over those segregating 1:1 under the rationale that these markers are more informative because they segregate from both parents; and, the third step uses the gap calculated between clusters mean to minimize genotype miscalls due to overlapping. The resulting 4,300 selected SPFs, being 1,915 and 2,385 segregating 1:1 and 3:1, respectively, were used in the linkage mapping analysis (Table 1).

4.3. Construction of a gene-rich map for *Eucalyptus*

The SFP segregation data resulted in the generation of a genetic linkage map comprising 1,064 makers (Figure 2). The majority of these markers (884

SFPs) represent unique genes while the remaining 180 markers are microsatellites. The microsatellites played an important role to confidently assign the SFPs to the expected 11 linkage groups of *Eucalyptus* formed at a minimum LOD of 7.0. A total of 1,848 SFPs grouped at this LOD threshold (Table 1) could, in principle, be ordered along the linkage groups. However a framework map with high likelihood support was constructed by simulated annealing, excluding markers that contributed to unstable marker orders. As microarray data is inherently noisy, we opted for map quality rather than allocating more genes but increasing the chance of erroneously ordering SFPs.

From the total number of SFPs mapped, 457 segregated 1:1 and 427 segregated 3:1, which received the acronym of BC and F2, respectively. The quality of the map is comparable to that generated by other classes of molecular markers, with markers evenly distributed throughout the linkage group. Except for a few cases on the edge of linkage groups, there were no evidences of major clustering or regions lacking genes (Figure 3), with SFPs spread equally along the intervals of microsatellites (Figure 2). However, some linkage groups had more (*e.g.* 6 and 8) genes mapped than others (*e.g.* 7), possibly resulting from a general greater abundance of genes on those chromosomes, or at least of those expressed in the transcriptome of differentiating xylem tissue (Table 3).

The SFP/microsatellite map had an average density of 1.2 cM with 97.5% of the intermarker distances smaller than 5 cM (Figure 4). For only five of the 1,053 intervals the distance was greater than 10 cM, with a maximum of 12.3 cM (Table 3, Figure 4). Even though the number of mapped markers increased more than five times when compared to the microsatellite only map, the total length of the map did not increase significantly. The total map length estimated at 1275 cM is within the expected range for *Eucalyptus*. The total length and intermarker distances for every linkage group built are consistent with the global results and no linkage group was either too long of had too spread out markers (Table 3).

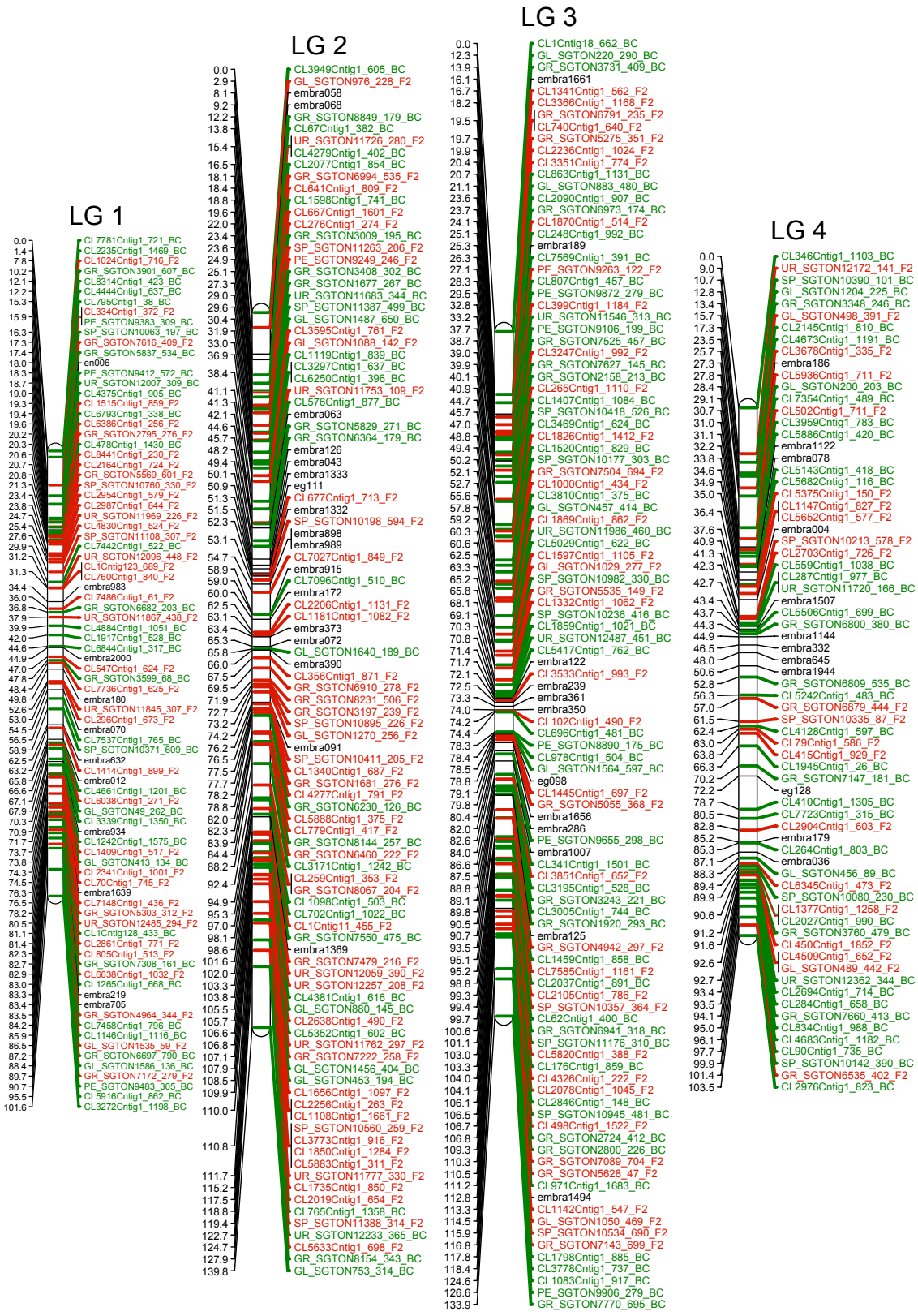
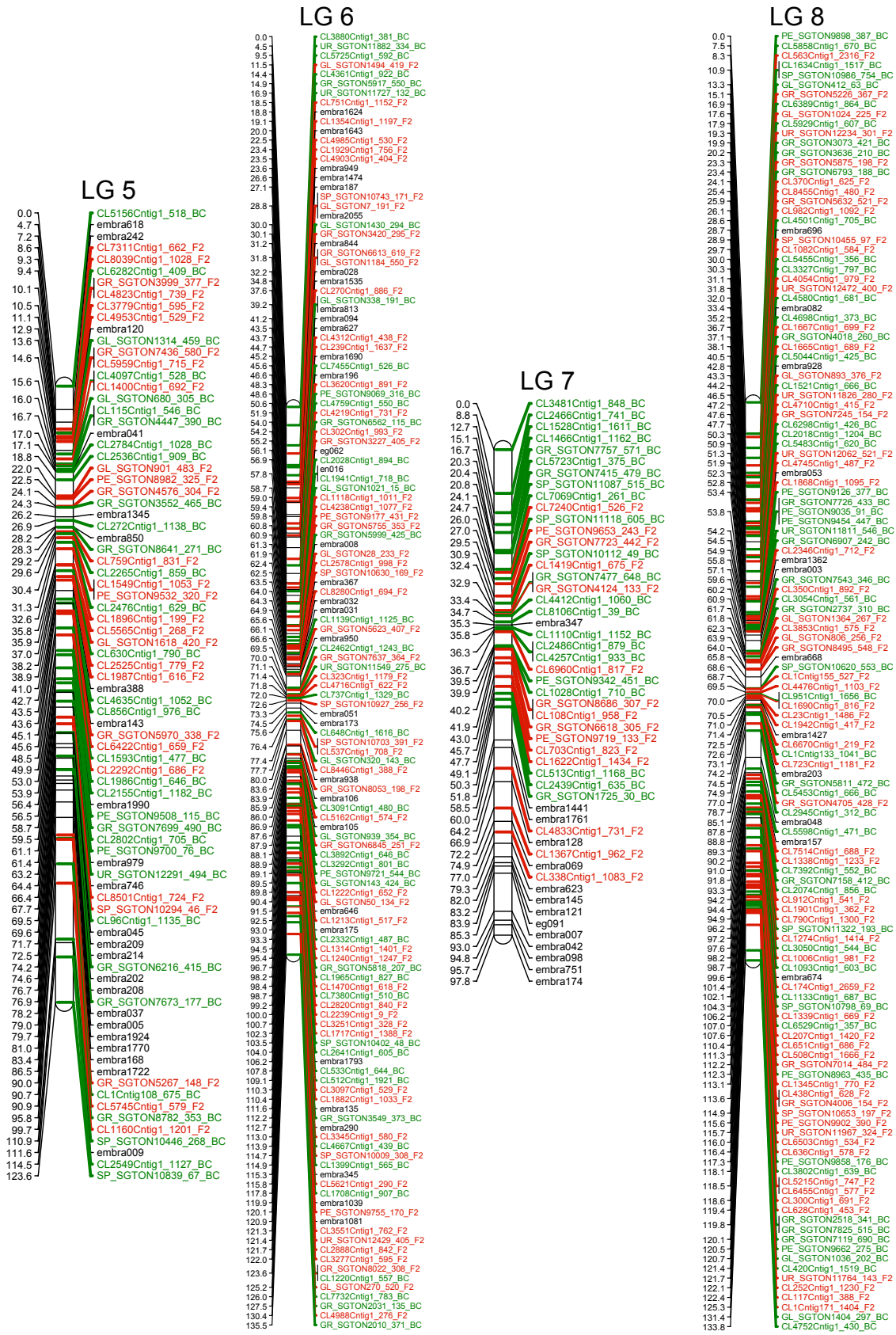
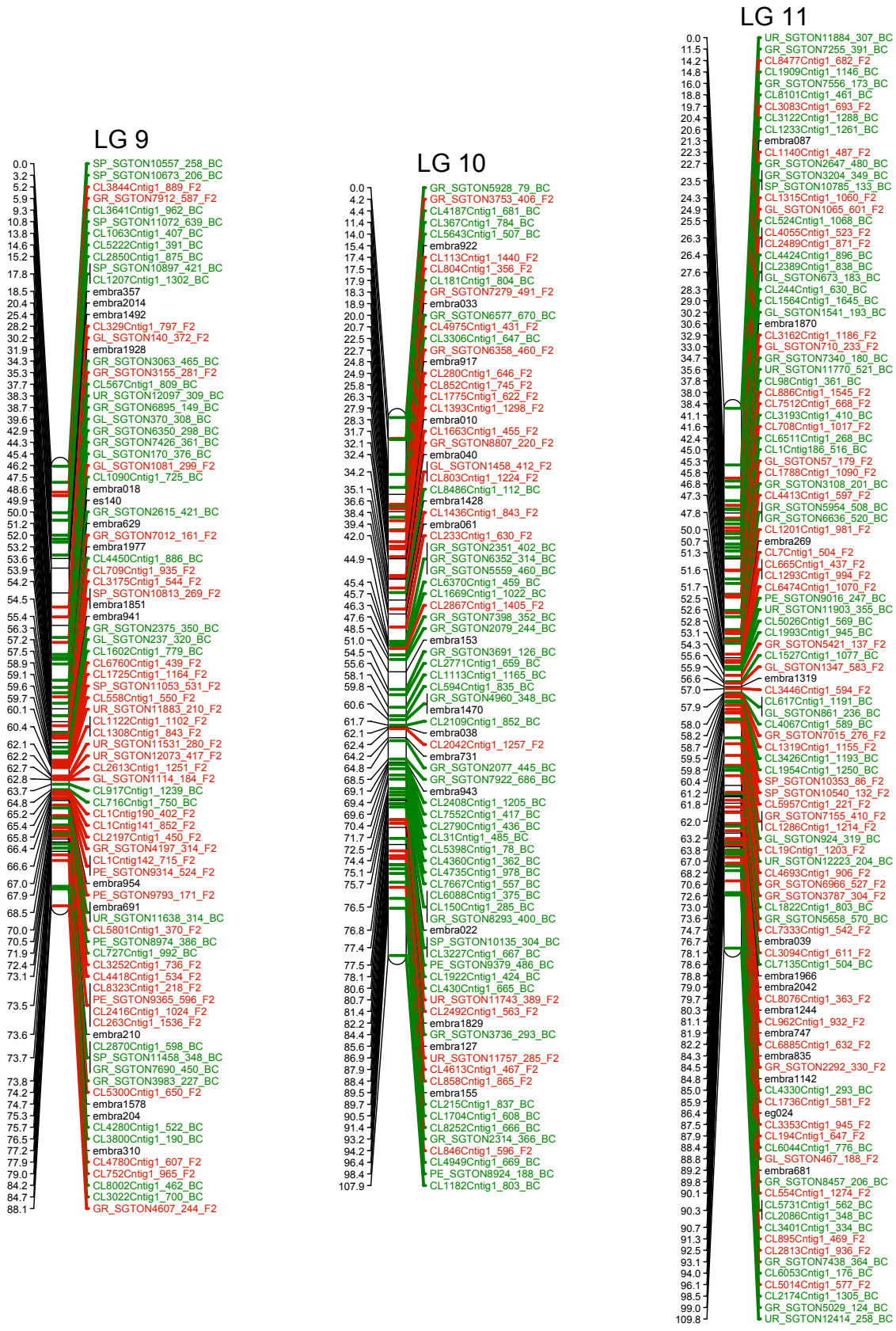


Figure 2: High-density SFP/microsatellite genetic linkage map of *E. urophylla* x *E. grandis*. Linkage groups 1 to 4 are shown. Microsatellites in black, SFPs segregating as F2 (3:1) and pseudo-testcross (1:1) in red and green, respectively. Linkage groups are numbered after Brondani *et al.* [25]. (Continued)


LG 5
LG 6
LG 7
LG 8


LG 9
0.0 SP_SGTON10557_258_BC
3.2 SP_SGTON10673_206_BC
5.2 CL3844Cntig1_889_F2
5.9 GR_SGTON7912_587_F2
9.3 CL3641Cntig1_962_BC
10.8 SP_SGTON11072_639_BC
13.8 CL1063Cntig1_407_BC
14.6 CL5222Cntig1_391_BC
15.2 CL2850Cntig1_875_BC
17.8 SP_SGTON10897_421_BC
CL1207Cntig1_1302_BC
18.5 embra357
20.4 embra2014
25.4 embra1492
28.2 CL329Cntig1_797_F2
30.2 GL_SGTON140_372_F2
31.9 embra1928
34.3 GR_SGTON3063_465_BC
35.3 GR_SGTON3155_281_F2
37.7 CL567Cntig1_809_BC
38.3 UR_SGTON12097_309_BC
38.7 GR_SGTON6895_149_BC
39.6 GL_SGTON370_308_BC
42.9 GR_SGTON6350_298_BC
44.3 GR_SGTON7426_361_BC
45.4 GL_SGTON170_376_BC
46.2 GL_SGTON1081_299_F2
47.5 CL1090Cntig1_725_BC
48.6 embra018
49.9 es140
50.0 GR_SGTON2615_421_BC
51.2 embra629
52.0 GR_SGTON7012_161_F2
53.2 embra1977
53.6 CL4450Cntig1_886_BC
53.9 CL709Cntig1_935_F2
54.2 CL3175Cntig1_544_F2
54.5 SP_SGTON10813_269_F2
embra1851
55.4 embra941
56.3 GR_SGTON2375_350_BC
57.2 GL_SGTON237_320_BC
57.5 CL1602Cntig1_779_BC
58.9 CL6760Cntig1_439_F2
59.1 CL1725Cntig1_1164_F2
59.6 SP_SGTON11053_531_F2
59.7 CL558Cntig1_550_F2
60.1 UR_SGTON11883_210_F2
60.4 CL1122Cntig1_1102_F2
CL1308Cntig1_843_F2
62.1 UR_SGTON11531_280_F2
62.2 UR_SGTON12073_417_F2
62.7 CL2613Cntig1_1251_F2
62.8 GL_SGTON1114_184_F2
63.7 CL917Cntig1_1239_BC
64.8 CL716Cntig1_750_BC
65.2 CL1Cntig190_402_F2
65.4 CL1Cntig141_852_F2
65.8 CL2197Cntig1_450_F2
66.4 GR_SGTON4197_314_F2
CL1Cntig142_715_F2
66.6 PE_SGTON9314_524_F2
67.0 embra954
67.9 PE_SGTON9793_171_F2
68.5 embra691
UR_SGTON11638_314_BC
70.0 CL5801Cntig1_370_F2
70.5 PE_SGTON8974_386_BC
71.9 CL727Cntig1_992_BC
72.4 CL3252Cntig1_736_F2
73.1 CL4418Cntig1_534_F2
CL8323Cntig1_218_F2
73.5 PE_SGTON9365_596_F2
CL2416Cntig1_1024_F2
CL263Cntig1_1536_F2
73.6 embra210
73.7 CL2870Cntig1_598_BC
SP_SGTON11458_348_BC
GR_SGTON7690_450_BC
73.8 GR_SGTON3983_227_BC
74.2 CL5300Cntig1_650_F2
74.7 embra1578
75.3 embra204
75.7 CL4280Cntig1_522_BC
76.5 CL3800Cntig1_190_BC
77.2 embra310
77.9 CL4780Cntig1_607_F2
79.0 CL752Cntig1_965_F2
84.2 CL8002Cntig1_462_BC
84.7 CL3022Cntig1_700_BC
88.1 GR_SGTON4607_244_F2
LG 10
0.0 GR_SGTON5928_79_BC
4.2 GR_SGTON3753_406_F2
4.4 CL4187Cntig1_681_BC
11.4 CL367Cntig1_784_BC
14.0 CL5643Cntig1_507_BC
15.4 embra922
17.4 CL113Cntig1_1440_F2
17.5 CL804Cntig1_356_F2
17.9 CL181Cntig1_804_BC
18.3 GR_SGTON7279_491_F2
18.9 embra033
20.0 GR_SGTON6577_670_BC
20.7 CL4975Cntig1_431_F2
22.5 CL3306Cntig1_647_BC
22.7 GR_SGTON6358_460_F2
24.8 embra917
24.9 CL280Cntig1_646_F2
25.8 CL852Cntig1_745_F2
26.3 CL1775Cntig1_622_F2
27.9 CL1393Cntig1_1298_F2
28.3 embra010
31.7 CL1663Cntig1_455_F2
32.1 GR_SGTON8807_220_F2
32.4 embra040
34.2 GL_SGTON1458_412_F2
35.1 CL803Cntig1_1224_F2
36.6 CL8486Cntig1_112_BC
38.4 embra1428
39.4 CL1436Cntig1_843_F2
42.0 embra061
CL233Cntig1_630_F2
44.9 GR_SGTON2351_402_BC
GR_SGTON6352_314_BC
GR_SGTON5559_460_BC
45.4 CL6370Cntig1_459_BC
45.7 CL1669Cntig1_1022_BC
46.3 CL2867Cntig1_1405_F2
47.6 GR_SGTON7398_352_BC
48.5 GR_SGTON2079_244_BC
51.0 embra153
54.5 GR_SGTON3691_126_BC
55.6 CL2771Cntig1_659_BC
58.1 CL1113Cntig1_1165_BC
59.8 CL594Cntig1_835_BC
GR_SGTON4960_348_BC
60.6 embra1470
61.7 CL2109Cntig1_852_BC
62.1 embra038
62.4 CL2042Cntig1_1257_F2
64.2 embra731
64.8 GR_SGTON2077_445_BC
68.5 GR_SGTON7922_686_BC
69.1 embra943
69.4 CL2408Cntig1_1205_BC
69.6 CL7552Cntig1_417_BC
70.4 CL2790Cntig1_436_BC
71.7 CL31Cntig1_485_BC
72.5 CL5398Cntig1_78_BC
74.4 CL4360Cntig1_362_BC
75.1 CL4735Cntig1_978_BC
75.7 CL7667Cntig1_557_BC
CL6088Cntig1_375_BC
76.5 CL150Cntig1_285_BC
GR_SGTON8293_400_BC
76.8 embra022
77.4 SP_SGTON10135_304_BC
77.5 CL3227Cntig1_667_BC
78.1 PE_SGTON9379_486_BC
80.6 CL1922Cntig1_424_BC
80.7 CL430Cntig1_665_BC
81.4 UR_SGTON11743_389_F2
82.2 CL2492Cntig1_563_F2
84.4 embra1829
85.6 GR_SGTON3736_293_BC
86.9 embra127
87.9 UR_SGTON11757_285_F2
88.4 CL4613Cntig1_467_F2
89.5 CL858Cntig1_865_F2
89.7 embra155
90.5 CL215Cntig1_837_BC
91.4 CL1704Cntig1_608_BC
93.2 CL8252Cntig1_666_BC
94.2 GR_SGTON2314_366_BC
96.4 CL846Cntig1_596_F2
98.4 CL4949Cntig1_669_BC
107.9 PE_SGTON8924_188_BC
CL1182Cntig1_803_BC
LG 11
0.0 UR_SGTON11884_307_BC
11.5 GR_SGTON7255_391_BC
14.2 CL8477Cntig1_682_F2
14.8 CL1909Cntig1_1146_BC
16.0 GR_SGTON7556_173_BC
18.8 CL8101Cntig1_461_BC
19.7 CL3083Cntig1_693_F2
20.4 CL3122Cntig1_1288_BC
20.6 CL1233Cntig1_1261_BC
21.3 embra087
22.3 CL1140Cntig1_487_F2
22.7 GR_SGTON2647_480_BC
23.5 GR_SGTON3204_349_BC
SP_SGTON10785_133_BC
24.3 CL1315Cntig1_1060_F2
24.9 GL_SGTON1065_601_F2
25.5 CL524Cntig1_1068_BC
26.3 CL4055Cntig1_523_F2
26.4 CL2489Cntig1_871_F2
CL4424Cntig1_896_BC
27.6 CL2389Cntig1_838_BC
GL_SGTON673_183_BC
28.3 CL244Cntig1_630_BC
29.0 CL1564Cntig1_1645_BC
30.2 GL_SGTON1541_193_BC
30.6 embra1870
32.9 CL3162Cntig1_1186_F2
33.0 GL_SGTON710_233_F2
34.7 GR_SGTON7340_180_BC
35.6 UR_SGTON11770_521_BC
37.8 CL98Cntig1_361_BC
38.0 CL886Cntig1_1545_F2
38.4 CL7512Cntig1_668_F2
41.1 CL3193Cntig1_410_BC
41.6 CL708Cntig1_1017_F2
42.4 CL6511Cntig1_268_BC
45.0 CL1Cntig186_516_BC
45.3 GL_SGTON57_179_F2
45.8 CL1788Cntig1_1090_F2
46.8 GR_SGTON3108_201_BC
47.3 CL4413Cntig1_597_F2
47.8 GR_SGTON5954_508_BC
GR_SGTON6636_520_BC
50.0 CL1201Cntig1_981_F2
50.7 embra269
51.3 CL7Cntig1_504_F2
51.6 CL665Cntig1_437_F2
51.7 CL1293Cntig1_994_F2
52.5 CL6474Cntig1_1070_F2
52.6 PE_SGTON9016_247_BC
52.8 UR_SGTON11903_355_BC
53.1 CL5026Cntig1_569_BC
54.3 CL1993Cntig1_945_BC
55.6 GR_SGTON5421_137_F2
55.9 CL1527Cntig1_1077_BC
56.6 GL_SGTON1347_583_F2
57.0 embra1319
57.9 CL3446Cntig1_594_F2
CL617Cntig1_1191_BC
58.0 GL_SGTON861_236_BC
58.2 CL4067Cntig1_589_BC
58.7 GR_SGTON7015_276_F2
59.5 CL1319Cntig1_1155_F2
59.8 CL3426Cntig1_1193_BC
60.4 CL1954Cntig1_1250_BC
61.2 SP_SGTON10353_86_F2
61.8 SP_SGTON10540_132_F2
62.0 CL5957Cntig1_221_F2
GR_SGTON7155_410_F2
63.2 CL1286Cntig1_1214_F2
63.8 GL_SGTON924_319_BC
67.0 CL19Cntig1_1203_F2
68.2 UR_SGTON12223_204_BC
70.6 CL4693Cntig1_906_F2
72.6 GR_SGTON6966_527_F2
73.0 GR_SGTON3787_304_F2
73.6 CL1822Cntig1_803_BC
74.7 GR_SGTON5658_570_BC
76.7 CL7333Cntig1_542_F2
78.1 embra039
78.6 CL3094Cntig1_611_F2
78.8 CL7135Cntig1_504_BC
79.0 embra1966
79.7 embra2042
80.3 CL8076Cntig1_363_F2
81.1 embra1244
81.9 CL962Cntig1_932_F2
82.2 embra747
84.3 CL6885Cntig1_632_F2
84.5 embra835
84.8 GR_SGTON2292_330_F2
85.0 embra1142
85.9 CL4330Cntig1_293_BC
86.4 CL1736Cntig1_581_F2
87.5 eg024
87.9 CL3353Cntig1_945_F2
88.4 CL194Cntig1_647_F2
88.8 CL6044Cntig1_776_BC
89.2 GL_SGTON467_188_F2
89.8 embra681
90.1 GR_SGTON8457_206_BC
90.3 CL554Cntig1_1274_F2
CL5731Cntig1_562_BC
CL2086Cntig1_348_BC
90.7 CL3401Cntig1_334_BC
91.3 CL895Cntig1_469_F2
92.5 CL2813Cntig1_936_F2
93.1 GR_SGTON7438_364_BC
94.0 CL6053Cntig1_176_BC
96.1 CL5014Cntig1_577_F2
98.5 CL2174Cntig1_1305_BC
99.0 GR_SGTON5029_124_BC
109.8 UR_SGTON12414_258_BC

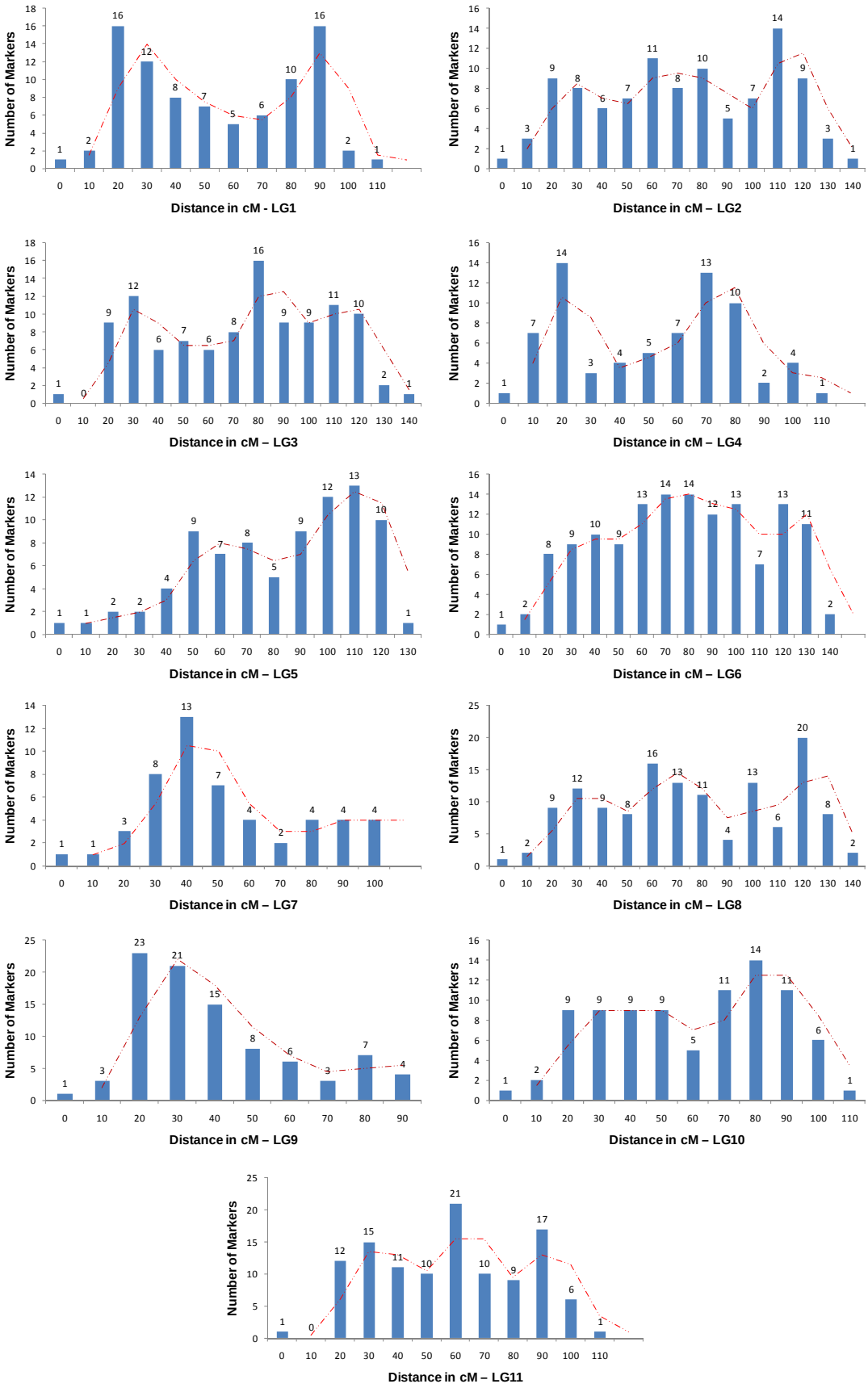


Figure 3: Frequency distribution of SFPs along the extension of the eleven linkage groups indicating variable marker clustering patterns.

Table 3: Descriptive statistics for the 11 linkage groups of the SFP/microsatellite map.

Linkage Group	SSR	Genes	Total	Length (cM)	Intermarker distance (cM)	
					Mean	Std Deviation
LG1	*11*	*75*	*86*	*102*	*1.20*	*1.25*
LG2	*17*	*85*	*102*	*140*	*1.38*	*1.55*
LG3	*12*	*95*	*107*	*134*	*1.26*	*1.63*
LG4	*12*	*59*	*71*	*103*	*1.48*	*1.59*
LG5	*23*	*61*	*84*	*124*	*1.49*	*1.76*
LG6	*34*	*104*	*138*	*136*	*0.99*	*0.97*
LG7	*14*	*37*	*51*	*98*	*1.96*	*1.89*
LG8	*12*	*122*	*134*	*134*	*1.01*	*1.16*
LG9	*16*	*75*	*91*	*88*	*0.98*	*1.06*
LG10	*16*	*71*	*87*	*108*	*1.25*	*1.45*
LG11	*13*	*100*	*113*	*110*	*0.98*	*1.55*
Total markers mapped	***180***	***884***	***1064***	***1275***	***1.21***	***1.44***

Although some linkage groups had more microsatellites than others, this apparently did not influence the amount of SFPs mapped. For example, linkage group 7 has 14 microsatellites and only 37 genes were assigned to this group; whereas linkage group 8, which has 12 microsatellites, was complemented with SFPs for 122 genes (Table 3). This lack of pattern is consistent across the whole map, suggesting that SFPs are behaving independently from the microsatellites and can in fact contribute to map saturation.

To test for the relative contribution that each class of SFP (1:1 and 3:1) contributed to the quality of the final map, a map involving only microsatellites and SFPs segregating 1:1 was built. The same was done with SFPs segregating 3:1. The results indicated that using only SFPs segregating 1:1 severely decreased overall map quality, as illustrated by the expansion of total map length (Table 4). On the other hand, mapping only SFPs segregating 3:1 did not result in a reduction of map quality, based on the generally used premise that a shorter map is the most likely one. For a similar number of markers, the total map length was 1,845 cM for the former case, 1,130 cM for the later situation and 1,275 cM when both kinds of SFPs were used. Thus,

employing an experimental design that allows for the detection of SFPs segregating 3:1 seems to be the best one not only as it allows mapping a larger number of genes but also because the final map quality is substantially increased.

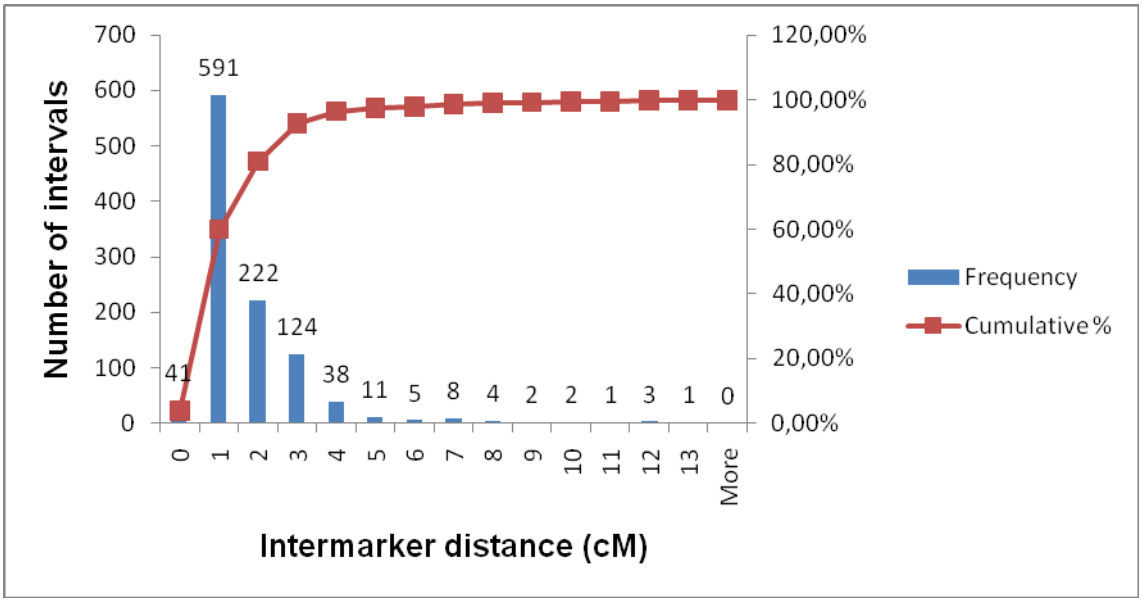


Figure 4: Frequency histogram showing the distribution of intervals between two markers for increasing intermarker distances.

Table 4: Contribution of SFPs segregating 3:1 to map quality. Number of markers mapped and final linkage group length for each linkage group are shown when (i) mapping microsatellites plus SFPs segregating 1:1 (Only 1:1), (ii) microsatellites plus SFPs segregating 3:1 (Only 3:1) and (iii) microsatellites plus SFPs segregating for both patterns (Both).

Linkage Group	Only 1:1		Only 3:1		Both	
	# markers	*Length (cM)*	# markers	*Length (cM)*	# markers	*Length (cM)*
LG 1	45	*146*	66	*92*	86	*102*
LG 2	73	*168*	70	*136*	102	*140*
LG 3	78	*208*	44	*87*	107	*134*
LG 4	59	*151*	40	*67*	71	*103*
LG 5	59	*244*	63	*122*	84	*124*
LG 6	87	*197*	98	*125*	138	*136*
LG 7	39	*98*	42	*109*	51	*98*
LG 8	84	*149*	81	*116*	134	*134*
LG 9	66	*138*	61	*87*	91	*88*
LG 10	73	*182*	48	*88*	87	*108*
LG 11	75	*164*	66	*101*	113	*110*
Total	**738**	***1845***	**679**	***1130***	**1064**	***1275***

4.4. SFP identification using mixed-model analysis of variance

Next, we were interested to know whether it would be possible to recover the mapped SFPs by doing a detection analysis on the data from the screening experiment where only the 28 biologically replicated individuals were used. Wolfinger *et al.* [50] and Rostoks *et al.* [16] previously proposed that fitting microarray data to a mixed-model analysis of variance is an effective way to separate the sources of variation and test the significance of these effects, identifying differentially expressed genes and putative SFPs, respectively. We expanded this principle and hypothesized that the same analysis could be used on our progeny data.

The normalized signal intensity for the 20,726 genes in 56 hybridizations (28 individuals) were first fit to a broad mixed-model intended to correct for global sources of variation not compensated by only quantile-normalizing the data. Subsequently, one analysis of variance was performed for each gene to identify those with significant Genotype x Probe interaction (GP) effect (see

methods). Significance of this source of variation would indicate genes where the signal intensity of one (or more) probe(s) deviates from the probeset's mean in a genotype dependent fashion, which is exactly the definition of a SFP. However, keep in mind that this analysis does not determine which probe of the probeset is behaving as an SFP. After correcting the significance threshold of the F tests for multiple testing, 4,648 genes showed significant GP effect at a false discovery rate < 0.005 (P < 0.0022), represented by a total of 25,600 probes.

This selection using ANOVA recovered 87% (1,603) of the genes that were linkage grouped when the full dataset from the 96 individuals was used (Figure 5A). When considering only genes that were represented by probes ultimately mapped, a fewer proportion was left aside and 811 (92%) genes were common to both selections (Figure 5B). These results indicate that using a mixed-model ANOVA on microarray data is a valid approach to identify genes with the source of variation of interest significant, even when a complex progeny dataset is being used.

As ANOVA does not determine which probe of the probeset is the putative SFP, we searched within each ANOVA-selected probeset for candidate SFPs applying the *k*-means clustering previously described to assign the normalized data of the 28 individuals into two distinct clusters. The chi square analysis of segregation and the modified normal deviate (*zi*) were also calculated to distinguish well separated clusters with 1:1 and 3:1 proportions from non-informative probes.

The set of probes that passed this selection pipeline and were selected reduced to 10,127, representing 4,251 genes. However, as indicated on Figure 5C-D, the majority of the genes that were previously linkage grouped or mapped were still among these 4,251 genes. For genes that were linkage grouped, 85% (1564 genes) were still detected (Figure 5C) and 90% (797 genes) continued to be selected considering only those that were ultimately mapped (Figure 5D).

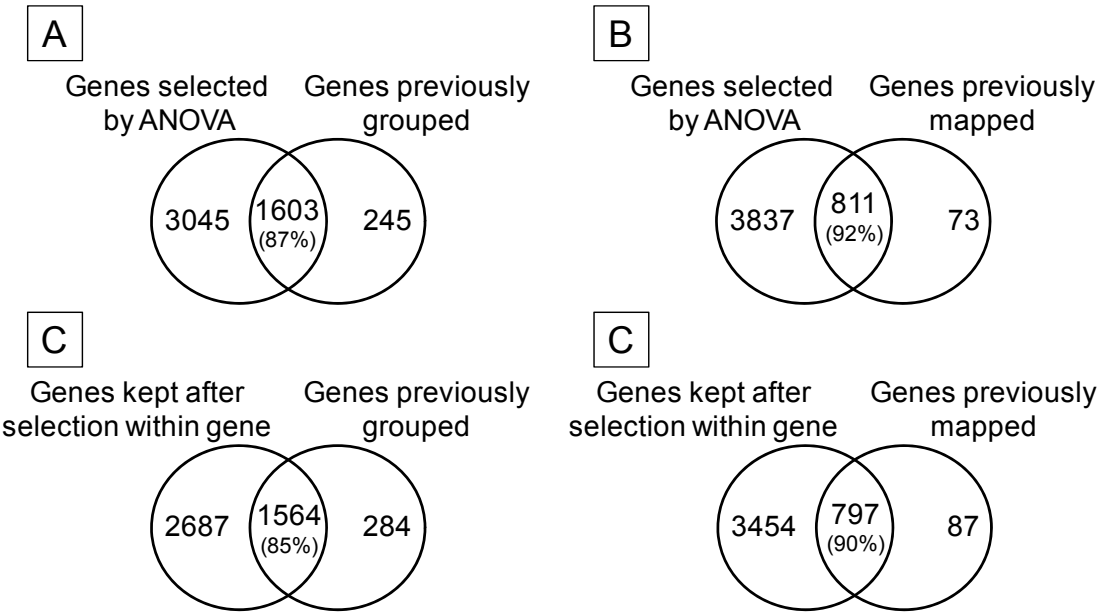


Figure 5: Venn diagrams comparing efficiency of identifying SFPs using mixed-model ANOVA on dataset from 28 individuals. A) Genes with significant GP effect compared to linkage grouped genes from the reference map. B) Genes with significant GP effect compared to mapped genes from the reference map. C) Genes that were still selected after further within gene selection compared to linkage grouped genes from the reference map. D) Genes that were still selected after further within gene selection compared to mapped genes from the reference map. Values between parentheses refer to the percentage of the genes linkage grouped or mapped that were common between analyses.

Assuming that probes which were assigned to linkage groups after our analysis involving 96 individuals are the real SFPs, the number of false positives detected by ANOVA was relatively low (5,917 probes in 2,687 genes). Those probes probably represent spurious segregation that were later removed when the number of individuals increased from 28 to 96. At any rate, it is noticeable that this approach confidentially retrieved the information generated using a much larger number of individuals (28 versus 96) and can be confidentially used during the screening step.

4.5. Number of probes per probeset

During probe design, an average of five unique probes were designed for every unigene on the array, a number that was chosen considering the amount of genes available and the limited funds available for synthesizing the high-

density oligonucleotide array. Nevertheless, because it was not known how many probes would be necessary to have at least one able to discover an SFP into the gene, for a subset of 1,308 genes, corresponding to 6.3% (1,308/20726) of the available unigene set, a probeset with a total of 10 probes was designed (Additional file 2).

Table 5 compares the relative efficiency of SFP discovery when five or 10 probes are used per probeset. A total of 4.6% of the genes with five probes designed had at least one SFP discovered and ultimately mapped on the reference map. In contrast, when 10 probes were used for the SFP screening and discovery experiment, the efficiency was almost doubled, increasing to 7.9%. A chi-squared test of homogeneity confirmed that this difference is highly significant ($\chi^2_{d.f.=1} = 28; P < 0.00001$) clearly indicating that using more probes per unigene does result in a significantly increased probability of discovering a segregating and mappable SFP and ultimately positioning the gene on the linkage map.

Table 5: Association between the number of probes screened for a gene (N) and the number of genes ultimately mapped.

N / Class	5	10	
Not mapped	15787	1205	16992
Mapped	762 (4.6%)	103 (7.9%)	865
	16549	1308	

The percentage of genes mapped when N= 5 and 10 is shown in parenthesis.

Even though we had 2,868 other genes with a variable number between one and four probes designed (Additional file 2), we did not include them in the analysis because this number of probes was not intentionally designed. Rather, they ended up having a smaller number of probes selected per probeset because the remaining probes were not considered high quality according to

Agilent's algorithm. At any rate, out of the 884 genes mapped, only 19 had less than five probes designed per probeset, further suggesting that screening a larger number of probes is probably a key requirement to efficiently discover an SFP for a particular gene and thus being able to map it.

3

5. DISCUSSION

We report that microarray-based detection of SFPs is an effective way for confidently mapping a large number of genes in organisms with unsequenced genomes. The analyses were done on a mapping population derived from an interespecific cross between two elite genotypes of *Eucalyptus urophylla* and *Eucalyptus grandis*. Eucalypts, as most of the other economical and ecological important species, are highly heterozygous as a consequence of a mixed mating system and the genome has not been fully sequenced yet. To overcome the lack of genome information, custom oligonucleotide arrays were designed from a now relatively small EST collection derived from different *Eucalyptus* species and tissue types. Currently, with the availability of next generation sequencing technologies, much larger EST collections can be generated at relatively reduced costs this providing a quick way of generating the unigene set necessary for probe design [63]. Short 25-mer probes were designed to maximize the detection of sequence polymorphisms in the form of SFPs, because transcript hybridization is more likely to be affected by single nucleotide polymorphisms as the probes are shorter. It is worth pointing out that the genotypes used in the present study were not among those sequenced to assemble the EST unigenes, which although on one side it increases the complexity of the probes designed, on the other provides generality that makes the unigene collection more widely applicable.

This SFP mapping study was carried out in an F1 mapping population where up to four alleles might be segregating in the progeny. Furthermore SFPs display a dominant pattern of expression so that homozygous and heterozygous

genotypes cannot be easily discriminated. As a result of this more complex background, SFPs might be polymorphic between the parents but do not segregate in the progeny and vice-versa, be monomorphic between the parents but segregate in a 3:1 ration in the progeny A simple solution to that is to employ a pseudo-testcross marker screening step to optimize the identification of informative SFPs analyzing progeny individuals rather than parents. This approach originally described by Grattapaglia and Sederoff (1994) has been widely used for selecting and mapping several classes of dominant markers such as RAPD and AFLP in hundreds of linkage maps published to date for outcrossing species.

Figure 6A-B represents the experimental design used to date for SPF detection and mapping , which involves the identification of SPFs based solely on the parental data [12-14, 16-22, 54, 55]. This approach makes sense for inbred species as SFPs detected between the homozygous parents will segregate in the F2 or RIL population used for mapping. For simplicity, we assume that there is no differential expression between the parents or allele-specific expression, that only one probe per probeset shows hybridization differences (probe 4) and that the unigene allele always matches the dominant segregating allele in the pedigree. Parental identification would always be successful when genotyping classical mapping populations such as RILs, but it does not apply to outcrossing species because such SFPs would not segregate in the progeny (Figure 6A). Even if segregation does occur (Figure 6B), heterozygous individuals would still have one allele matching the probe. Therefore, the signal intensity of the heterozygous individuals (B) would be, theoretically, two times less than that of the homozygous (A), which in a Log2 transformed data represents a single unit change (*e.g* $\log_2 200 - \log_2 100$).

As a consequence genotyping SFPs segregating 1:1 in outcrossing species is more difficult as there is a bigger overlap between the two groups of individuals, especially when the inherent background noise of the microarray data is taken into account (Figure 6C), and other patterns of segregation are discarded [21]. On the other hand, analyzing progeny data, it is possible to successfully identify and genotype probes that show alleles segregating 3:1 in

the progeny with a much lower chance of overlapping between clustering classes (Figure 6D-E).

The examples given in Figure 6, however, are simplified and require assumptions that are quite unrealistic for *Eucalyptus* and it was even difficult to find experimental data to exemplify the diagrams. Variations in the expression level between individuals might make some of these polymorphisms disappear while others would have a greater gap between classes, especially for the parental data. For example, imagine if the genitors on Figure 6B had differential expression. On the other hand, situations where the probe allele does not follow the assumption (*i.e.* the unigene sequence does not match the dominant segregating allele), the polymorphisms detected behave differentially, creating 1:1 patterns with well separated classes and cases where the four distinct alleles will be detected as 1:1. For example, supposing the SNP on Figure 6D is a T instead of the G, the progeny would segregate in a distinguishable 1:1 pattern. Nevertheless, no matter what combinatorial changes one makes to these sources of polymorphisms, they will always result in patterns of segregation that can be converted to dominant 1:1 and 3:1 segregations in the progeny. Finally, in many cases multiple SFPs were detected per probeset.

All these results together might influence SFP detection depending on the statistical method used. The clustering method, yet relatively simple, efficiently performed this task and a relatively low level of possible clustering errors were visually observed throughout the experiment, with most of these errors occurring for probes segregating 1:1 (data not shown). If parental genotype data were available, maybe it could be used to incorporate a Bayesian probability for each progeny belonging to either parental signal range, similarly to what was recently proposed by Wang *et al*. [20], and *a posteriori* probability of an individual being inappropriately assigned to one cluster could be used to reduce genotyping miscalls. Probes segregating 3:1 could take advantage of this improvement to separate the classes in a 1:2:1 co-dominant segregation.

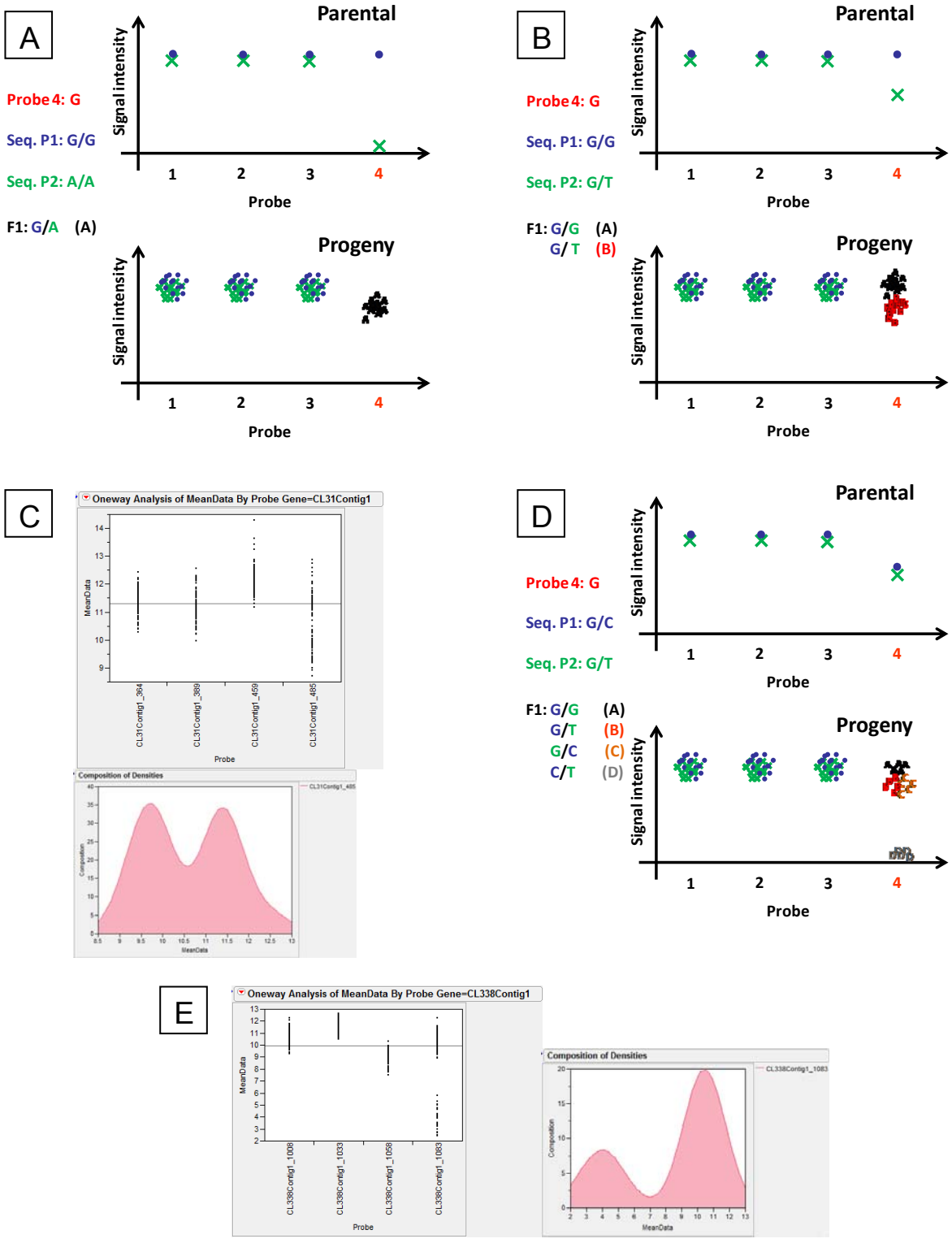


Figure 6: SFP identification in the mapping population. A) Hypothetical example to show situations where the SFP discovery design used for pedigrees derived from inbred lines fails in identifying polymorphisms in outcrossing species. B) Example of a probe segregating 1:1 in the progeny with some degree of overlapping between classes. C) Experimental data exemplifying diagram B. D) Hypothetical example demonstrating advantage of screening for SFPs using progeny individuals rather than parents. E) Experimental data exemplifying diagram D. Polymorphism is always present on probe 4.

Applying a slightly different analysis at the probeset level, Das *et al.* [55] demonstrated that cross-species microarrays can be used to detect SFPs by using Affymetrix soybean genome array on cowpea. However, they also showed that from the 37,500 probesets of the Affy chip, only 7,000 could be used to potentially reveal SFPs in cowpea due to expression and polymorphism differences between both species. Although *Eucalyptus* species are genetically closer to each other, our findings agree with those found with the soybean-cowpea. When we evaluated the background expression level based on the signal intensity of negative control probes, the influence of designing probes from unigene sets derived from complex EST libraries was noticeable. For the genes with probes that had a signal intensity above this minimum background, only 22% (3,622/16,163) had all probes included in this group. The other probesets showed one or more probes with a totally low offscale signal, which we postulate as being a consequence of none of the parents' transcript allele matching the unigene sequence used to design the probe.

These results are of importance if one is willing to use the microarray data to estimate gene expression levels for other downstream analysis such as expression QTLs or to find differentially expressed genes. Kirst *et al.* [44] demonstrated the bias incorporated to gene expression estimates when highly genetic diverse maize inbred lines are hybridized to the array. They also showed that excluding probes detected as putative SFPs was not sufficient to get completely unbiased data. Therefore, in addition to the high degree of polymorphism detected in outcrossing species, the estimates of gene expression from EST-developed microarrays would be somehow biased and the dual application of the microarray data has to be used with caution. Drost *et al.* [21], also working on a highly polymorphic organism, used longer 60-mer probes to mitigate this potential problem, but at a possible cost of a reduced ability to discover SFPs and, consequently mapping genes.

Perhaps an optimum strategy for highly polymorphic outcrossing species, yet to be tested, would be to include both short and long probes on the array for simultaneous genotyping and gene expression analysis. Although there has not been a fully comprehensive genome-wide survey of the number of SNPs in coding regions of *Eucalyptus*, Novaes *et al*. [63] estimated one SNP for every

192 bp on average in high-quality unigenes assembled after 454-based EST sequencing. As the authors discuss, the sequencing depth obtained does not allowed all *E. grandis* haplotypes sampled to be sequenced and their criteria to call a SNP was very stringent, probably excluding less frequent alleles and intentionally not including indels and variants involving more than one nucleotide. A possibly more realistic observation was reported by Poke *et al.* [64] for *E. globulus*. Yet sequencing only two lignin genes, they found 1 SNP for every 48 pb within *CCR* sequence and 1 SNP for every 147 pb in *CAD2* sequence. Therefore, as avoiding sequence polymorphism that affect hybridization will probably be impossible when designinig probes for *Eucalyptus* microarray studies, using longer probes would at least contribute by increasing the length of continuous perfect match, which has been proved to reduce the bias caused by SNPs [65].

Differently from any other SFP study previously reported, we show that the application of the pseudo-testcross screening strategy to simultaneously identify and genotype SFP using a progeny dataset, rather than first identifying SFPs using only the parents, was successful in detecting the two patterns of dominant segregation possible. While using only the two parents for SFP screening is obviously less expensive, the number of genes mapped and the overall map quality was substantially increased by screening using a progeny dataset that allowed the discovery of markers segregating 3:1 that would not be selected if only the parents would have been used (Table 4). There are two possible explanations for this observation. Firstly, for probes segregating 3:1 fewer individuals would likely not be correctly assigned to the two clusters and, thus, the chance of genotype miscalling is minimized. A high rate of miscalls generates false recombinants complicating map construction. Secondly, this class of SFPs segregating from both parents significantly improves the process of linkage phase estimation adding quality to the final map. This is possibly due to the fact that dominant SFP markers, even though segregating in a less informative 3:1 configuration, as they can occur as tightly linked in repulsion phase provide the same information content as a single co-dominant marker as originally proposed by Williams *et al*. [66] and later formalized by Plomion *et al.*

[67]. As a consequence, our final map had total length within that reported for *Eucalyptus* developed using other traditional molecular markers [5, 23, 25].

Although the clustering method clearly makes no attempt to differentiate SFP from GEM, our results suggest that most of the polymorphisms of signal intensity detected are likely revealing SFPs. Wang *et al.* [20] suggested that detecting few significant probes per probeset is an indication that the method outperforms in detecting SFPs and, even though this measure is confused with the effect of the original sequence wherefore the probes were derived, a very small percentage, only 66 genes out of the 884 mapped, had more than two probes declared as putative SFPs. Also, among genes for which several probes were detected as SFPs, different segregation patterns were identified (*i.e.* 1:1 and 3:1), which contradicts the definition of GEMs. Furthermore, even if the same segregation pattern was identified, correlation between the probes was sometimes different than 1 (considering the assigned cluster as the main variable), suggesting that the source of polymorphism affecting such probes are different (data not shown). Finally, similar to our observations, West *el al.* [18] noticed that using the RIL distribution to detect GEMs resulted in fewer markers than using parental distributions, a result they attributed to the effects that influence gene expression patterns in segregating populations, such as transgressive segregation, epistasis and genotype x environment interactions.

The relative position and ordering of the genes mapped by the SFPs could not be fully confirmed at the physical sequence level as a reference genome is not yet available for *Eucalyptus*. However, a preliminary analysis on linkage group 6, the most saturated one, against the scaffolds assembled from the 4X genome sequence, indicates that SFPs position is accurate as those that co-localized on the genetic map were located on the same scaffold. Also, the largest scaffold (2 Mbp) assembled on this linkage group suggests preliminary agreement of the genetic position of SFPs with their physical position within the scaffold (data not shown). Previous SPF mapping studies have consistently reported that the genetic position and order of such markers follow the order

expected from the annotation of the sequenced genome, with only few exceptions [18, 19, 21].

A major problem associated with using SFPs for gene mapping is that only those transcripts present in the tissue analyzed can be screened and ultimately mapped. We hybridized cRNA from differentiating xylem with a two-fold objective. First and most important was to reduce the genome complexity to allow the generation of data on the microarray by using only the low copy, high complexity transcribed portions of the genome. Secondly to focus the experiment on expressed genome regions so as to enrich the existing genetic maps with genes that coud be then proposed as positional candidates for QTLs. Given the results of this study the maximization of the number of genes mapped could be achieved using more than one segregating family and tissue type for SFP discovery.

For practical applications our findings show that screening for SFPs using a relatively small subset of the mapping population allows the discovery of a large number of high quality mappable SFPs. Applying mixed-model analysis of variance to this dataset it was possible to retrieve most of the mapped SFPs. Since this analysis isolates the effects of gene expression, we first hypothesized that the few genes not selected by this method could have been GEMs. Nevertheless, there was no direct correlation between these genes and having a larger number of probes designed per probeset. On the other hand, in the 73 mapped but undetected genes significantly more BC markers (57) than F2 ones (16) ($\chi^2_{d.f.=1} = 2; P < 0.00001$) were observed, which also displayed overlapping clusters with relatively small gaps between clusters' mean ($\mu = 1.67; \sigma = 0.7$), probably being this the reason for not being detected via ANOVA.

Hybridizing an even smaller number of individuals it may be possible to detect genes with putative SFPs using ANOVA, but as the number of individuals is decreased, reduced is the ability to identify probes segregating 3:1. If we had performed this analysis before designing the genotyping array, selection using only ANOVA would still have kept 25,600 probes for the full scale segregation experiment. Applying subsequent analysis within the ANOVA-selected probesets barely lost power in detecting genes that mapped and the number of

probes that would have been kept lowered to 10,127. Therefore, spending a little more to initially hybridize more individuals will certainly pay off by enriching a smaller microarray format with probes with a much higher probability of detecting high quality SFPs with mendelian segregation that will ultimately allow mapping their corresponding genes.

Our results also suggest that increasing the number of screened probes per gene significantly increases the probability of discovering SFP and ultimately the possibility of mapping genes. Assuming that this increment would be linear and experiment-wide, if we had designed the array with ten probes for the 17,857 (16,549+1308) genes where they were available, theoretically this would have resulted in 1,411 (7.9% of 17,857) genes mapped. This represents an increment of almost 60% over the 884 genes we mapped. In addition to the demonstration that the best candidate probes can be immediately detected using ANOVA, starting with more probes is an approach that will ultimately result in a reduction of costs per datapoint.

The reduction of costs and the availability of increasingly higher microarray density formats facilitate the use of a larger number of probes. Furthermore generating an EST resource from which to derive these probes is currently a relatively inexpensive and fast process for any species. For example, Novaes *et al.* [63] used 454 pyrosequencing technology to generate a much larger unigene set (71,384) than what we used and showed that such approach even outperforms Sanger sequencing on sampling genes, although the contig length was considerable smaller. At an approximate cost of $10,000 USD per run for the GS-FLX Titanium chemistry, which produces longer reads of 400-600 bp, EST resources are no longer a limitation even for complex genomes.

For our experimental design the cost of array purchase, labeling, hybridization and data collection was approximately $300 USD per genotype. It has to be considered, however, that sampling more genes and increasing probe density will require some compensation in terms of array size. The recently released 3G Agilent microarray supports configurations up to 1M and more

adequate formats of 2 x 400K and 4 x 180K will probably be available for custom expression designs later this year or first next year at an estimate cost of $1,030 and $1,200 USD, respectively (Matt Angel, Agilent sales representative, personal communication). Given such variety of formats a three-step experimental design can be imagined where, (i) genitors plus very few individuals could be initially hybridized only to exclude clearly not expressed probesets; (ii) followed by an expanded screening with more individuals that will allow selection of putative probes by ANOVA and clustering analyses; and, (iii) a final step where these probes will be included in a much smaller format to genotype the other selected individuals of the mapping population.

Although microarray based markers have been mostly employed in inbred, model organisms for which genome information is generally available, our study shows that a combination of a relatively limited EST resource together with SFP discovery and mapping on oligonucleotide arrays is a powerful approach to quickly localize several hundred or even thousands of genes to a reference map. Finally, the large number of genes mapped by SFP detection provide markers for several different applications. Such a gene-rich map represents a very useful resource for gene discovery when used in combination with QTL and association mapping and should be especially valuable for uncharacterized genomes of plant and animal species. In a molecular breeding scenario, for example, the co-localization of the genes mapped by SFPs with previously detected QTLs could result in positional candidate genes to have SNPs designed and genotyped for association genetics studies. Conversely, genome wide selection has been proposed to assist breeding programs for multiple complex quantitative traits [30] and the SFP mapping information could be used for increasing the relative weight of genomic regions in predictive models.

6. CONCLUSIONS

The results of our SFP experiment on *Eucalyptus* may be summarized in the following main conclusions:

i) The use of inexpensive EST resources has shown to be a valid source of sequence information to design oligonucleotide arrays for large-scale gene mapping via SFP in unsequenced genomes;
ii) The pseudo-testcross strategy using a subset of the mapping population allowed dominant SFPs to be detected segregating both 1:1 and 3:1;
iii) Simultaneous detection and genotyping using clustering analysis of a subset of the mapping population selected by Selective Mapping approach was sufficient to map SFPs in our genetic pedigree;
iv) A highly saturated genetic map was constructed positioning 884 unique genes from SFP markers;
v) Inclusion of SFPs segregating 3:1 has been demonstrated to substantially increase overall final map quality;
vi) Selection of putative SFP on a screening array by mixed-model ANOVA associated to clustering analysis within probeset recovered most of the mapped genes and are indicated for such purpose;
vii) And, increasing the number of probes sampling the unigene sequence resulted in more genes being ultimately mapped via SFP detection.

7. REFERENCES

1. Morgan TH, Sturtevant AH, Muller HJ, Bridges CB: **The Mechanism of Mendelian Heredity**. New York: Henry Holt and Company; 1915.
2. Botstein D, White RL, Skolnick M, Davis RW: **Construction of a genetic linkage map in man using restriction fragment length polymorphisms**. *Am J Hum Genet* 1980, **32**(3):314-331.
3. Grattapaglia D, Kirst M: **Eucalyptus applied genomics: from gene sequences to breeding tools**. *New Phytol* 2008, **179**(4):911-929.
4. Gion JM, Rech P, Grima-Pettenati J, Verhaegen D, Plomion C: **Mapping candidate genes in Eucalyptus with emphasis on lignification genes**. *Molecular Breeding* 2000, **6**(5):441-449.
5. Thamarus KA, Groom K, Murrell J, Byrne M, Moran GF: **A genetic linkage map for Eucalyptus globulus with candidate loci for wood, fibre, and floral traits**. *Theor Appl Genet* 2002, **104**(2-3):379-387.
6. Komulainen P, Brown GR, Mikkonen M, Karhu A, Garcia-Gil MR, O'Malley D, Lee B, Neale DB, Savolainen O: **Comparing EST-based genetic maps between Pinus sylvestris and Pinus taeda**. *Theor Appl Genet* 2003, **107**(4):667-678.
7. Eckert AJ, Pande B, Ersoz ES, Wright MH, Rashbrook VK, Nicolet CM, Neale DB: **High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (Pinus taeda L.)**. *Tree Genetics & Genomes* 2009, **5**(1):225-234.
8. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.
9. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ: **Genome-wide expression monitoring in Saccharomyces cerevisiae**. *Nat Biotechnol* 1997, **15**(13):1359-1367.
10. Shiu SH, Borevitz JO: **The next generation of microarray research: applications in evolutionary and ecological genomics**. *Heredity* 2008, **100**(2):141-149.
11. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ *et al*: **Direct allelic variation scanning of the yeast genome**. *Science* 1998, **281**(5380):1194-1197.
12. Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J: **Large-scale identification of single-feature polymorphisms in complex genomes**. *Genome Res* 2003, **13**(3):513-523.
13. Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, Briggs SP: **A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization**. *PLoS Genet* 2006, **2**(9):e144.
14. Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D, Nguyen HT: **Single feature polymorphism discovery in rice**. *PLoS ONE* 2007, **2**(3):e284.
15. Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L: **Simultaneous genotyping, gene-expression measurement, and**

detection of allele-specific expression with oligonucleotide arrays. *Genome Res* 2005, **15**(2):284-291.
16. Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S, Morris J, Cardle L, Marshall DF, Waugh R: **Single-feature polymorphism discovery in the barley transcriptome**. *Genome Biol* 2005, **6**(6):R54.
17. Cui X, Xu J, Asghar R, Condamine P, Svensson JT, Wanamaker S, Stein N, Roose M, Close TJ: **Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit**. *Bioinformatics* 2005, **21**(20):3852-3858.
18. West MA, van Leeuwen H, Kozik A, Kliebenstein DJ, Doerge RW, St Clair DA, Michelmore RW: **High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis**. *Genome Res* 2006, **16**(6):787-795.
19. Luo ZW, Potokina E, Druka A, Wise R, Waugh R, Kearsey MJ: **SFP genotyping from affymetrix arrays is robust but largely detects cis-acting expression regulators**. *Genetics* 2007, **176**(2):789-800.
20. Wang M, Hu X, Li G, Leach LJ, Potokina E, Druka A, Waugh R, Kearsey MJ, Luo Z: **Robust detection and genotyping of single feature polymorphisms from gene expression data**. *PLoS Comput Biol* 2009, **5**(3):e1000317.
21. Drost DR, Novaes E, Boaventura-Novaes C, Benedict CI, Brown RS, Yin T, Tuskan GA, Kirst M: **A microarray-based genotyping and genetic mapping approach for highly heterozygous outcrossing species enables localization of a large fraction of the unassembled Populus trichocarpa genome sequence**. *Plant J* 2009.
22. Bernardo AN, Bradbury PJ, Ma H, Hu S, Bowden RL, Buckler ES, Bai G: **Discovery and mapping of single feature polymorphisms in wheat using Affymetrix arrays**. *BMC Genomics* 2009, **10**(1):251.
23. Grattapaglia D, Sederoff R: **Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers**. *Genetics* 1994, **137**(4):1121-1137.
24. Ashburner M, Bergman CM: **Drosophila melanogaster: a case study of a model genomic sequence and its consequences**. *Genome Res* 2005, **15**(12):1661-1667.
25. Brondani RP, Williams ER, 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 Biol* 2006, **6**:20.
26. Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK: **An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts**. *Euphytica* 2005, **142**(1-2):169-196.
27. Price AH: **Believe it or not, QTLs are accurate!** *Trends Plant Sci* 2006, **11**(5):213-216.
28. Grattapaglia D: **Aplicações Operacionais de Marcadores Moleculares**. In: *Biotecnologia Florestal.* Edited by Borém A. Viçosa; 2007: 175-200.

29. Bernardo R: **Molecular markers and selection for complex traits in plants: Learning from the last 20 years**. *Crop Science* 2008, **48**(5):1649-1664.
30. Goddard ME, Hayes BJ: **Mapping genes for complex traits in domestic animals and their use in breeding programmes**. *Nat Rev Genet* 2009, **10**(6):381-391.
31. 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.
32. 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**(1):175-180.
33. Grattapaglia D, Bertolucci FLG, 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**(3):1205-1214.
34. Verhaegen D, Plomion C, Gion JM, Poitel M, Costa P, Kremer A: **Quantitative trait dissection analysis in Eucalyptus using RAPD markers .1. Detection of QTL in interspecific hybrid progeny, stability of QTL expression across different ages**. *Theoretical and Applied Genetics* 1997, **95**(4):597-608.
35. Brondani RPV, Brondani C, Tarchini R, Grattapaglia D: **Development, characterization and mapping of microsatellite markers in Eucalyptus grandis and E-urophylla**. *Theoretical and Applied Genetics* 1998, **97**(5-6):816-827.
36. Fraser LG, Tsang GK, Datson PM, De Silva HN, Harvey CF, Gill GP, Crowhurst RN, McNeilage MA: **A gene-rich linkage map in the dioecious species Actinidia chinensis (kiwifruit) reveals putative X/Y sex-determining chromosomes**. *BMC Genomics* 2009, **10**:102.
37. Neves LG, Faria DAd, Pappas GJ, Jr., Pasquali G, Grattapaglia D: **Diversidade nucleotídica e utilização de SNPs para o mapeamento de genes candidatos em Eucalyptus spp.** *EMBRAPA Comunicado Técnico 180* 2008.
38. Gupta PK, Rustgi S, Mir RR: **Array-based high-throughput DNA markers for crop improvement**. *Heredity* 2008, **101**(1):5-18.
39. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*: **Life with 6000 genes**. *Science* 1996, **274**(5287):546, 563-547.
40. Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY *et al*: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**(6814):796-815.
41. Pena-Castillo L, Hughes TR: **Why are there still over 1000 uncharacterized yeast genes?** *Genetics* 2007, **176**(1):7-14.
42. **A quick guide to sequenced genomes** [http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_p1.shtml]
43. Kirst M, Myburg AA, De Leon JP, Kirst ME, Scott J, Sederoff R: **Coordinated genetic regulation of growth and lignin revealed by**

quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiol* 2004, **135**(4):2368-2378.
44. Kirst M, Caldo R, Casati P, Tanimoto G, Walbot V, Wise RP, Buckler ES: **Genetic diversity contribution to errors in short oligonucleotide microarray analysis**. *Plant Biotechnol J* 2006, **4**(5):489-498.
45. Kerr MK, Churchill GA: **Statistical design and the analysis of gene expression microarray data**. *Genet Res* 2001, **77**(2):123-128.
46. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays**. *Nat Genet* 1999, **21**(1 Suppl):33-37.
47. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G: **The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster**. *Nat Genet* 2001, **29**(4):389-395.
48. Kerr MK, Churchill GA: **Experimental design for gene expression microarrays**. *Biostatistics* 2001, **2**(2):183-201.
49. Yang H, Harrington CA, Vartanian K, Coldren CD, Hall R, Churchill GA: **Randomization in laboratory procedure is key to obtaining reproducible microarray results**. *PLoS ONE* 2008, **3**(11):e3724.
50. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS: **Assessing gene significance from cDNA microarray expression data via mixed models**. *J Comput Biol* 2001, **8**(6):625-637.
51. Ayroles JF, Gibson G: **Analysis of variance of microarray data**. *Methods Enzymol* 2006, **411**:214-233.
52. Churchill GA: **Fundamentals of experimental design for cDNA microarrays**. *Nat Genet* 2002, **32 Suppl**:490-495.
53. Knapen D, Vergauwen L, Laukens K, Blust R: **Best practices for hybridization design in two-colour microarray analysis**. *Trends Biotechnol* 2009.
54. Bischoff SR, Tsai S, Hardison NE, York AM, Freking BA, Nonneman D, Rohrer G, Piedrahita JA: **Identification of SNPs and INDELS in swine transcribed sequences using short oligonucleotide microarrays**. *BMC Genomics* 2008, **9**:252.
55. Das S, Bhat PR, Sudhakar C, Ehlers JD, Wanamaker S, Roberts PA, Cui X, Close TJ: **Detection and validation of single feature polymorphisms in cowpea (Vigna unguiculata L. Walp) using a soybean genome array**. *BMC Genomics* 2008, **9**:107.
56. Vision TJ, Brown DG, Shmoys DB, Durrett RT, Tanksley SD: **Selective mapping: a strategy for optimizing the construction of high-density linkage maps**. *Genetics* 2000, **155**(1):407-420.
57. Grattapaglia D: **Integrating genomics into Eucalyptus breeding**. *Genet Mol Res* 2004, **3**(3):369-379.
58. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *Science* 2006, **313**(5793):1596-1604.
59. Chang S, Puryear J, Cairney J: **A simple and efficient method for isolating RNA from pine trees** *Plant Molecular Biology Reporter* 1993, **11**(2):113-116.

60. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9440-9445.
61. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
62. Van Ooijen JW, Voorrips RE: **JoinMap® 3.0, Software for the calculation of genetic linkage maps.** In. Edited by International PR, 3.0 edn. Wageningen, the Netherlands.
; 2001.
63. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Jr., Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome**. *BMC Genomics* 2008, **9**:312.
64. Poke FS, Vaillancourt RE, Elliott RC, Reid JB: **Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase 2 (CAD2)**. *Molecular Breeding* 2003, **12**(2):107-118.
65. Rennie C, Noyes HA, Kemp SJ, Hulme H, Brass A, Hoyle DC: **Strong position-dependent effects of sequence mismatches on signal ratios measured using long oligonucleotide microarrays**. *BMC Genomics* 2008, **9**:317.
66. Williams JGK, Hanafey MK, Rafalski JA, Tingey SV: **GENETIC-ANALYSIS USING RANDOM AMPLIFIED POLYMORPHIC DNA MARKERS**. *Method Enzymol* 1993, **218**:704-740.
67. Plomion C, Liu BH, Omalley DM: **Genetic analysis using trans-dominant linked markers in an F-2 family**. *Theoretical and Applied Genetics* 1996, **93**(7):1083-1089.

8. ADDITIONAL FILES

Additional file 1: Microsatellite panel used for parentage and identity confirmation of the offspring from the U15 x G38 cross. f and r represents forward and reverse primers, respectively.

SSR Loci	Fluorescence	Expected aleles (bp) U15	Expected aleles (bp) G38	Primer sequence
EMBRA 646	6-FAM	142/150	146/148	f: AAAGCGTTACGTGCGACTCT r: GGTACAGAAGAGGGCGTCAA
EMBRA 310	6-FAM	280/292	282/288	f: CTCCGTCTTCTCCATCCGTG r: GGCATAGCAAGTGATCAAGC
Eg 096	NED	283/285	277/279	f: CCAGGGAAAACAATTCAAGC r: GAGCGACAAACCCAAGTTTC
EMBRA 4	NED	88/102	80/94	f: ATACAATGATTTGAAAGGGG r: GAGTTGTTTGTTTTGTCGAA
EMBRA 101	HEX	124/126	122/146	f: TGATAGAGAGGTACATGGAGC r: TAAGACTCATGTGAACTAATTGG
EMBRA 746	HEX	173/195	177/203	f: GCCAGTAGTGTTTTCCTCGG r: TTGCCCTCCTCATGGTATTC

Additional file 2: Summary of the number of probes selected for the unigenes represented in the microarray used for polymorphic probe screening.

Probes per probeset	Number of unigenes	% of the total
1	721	3
2	722	3
3	698	3
4	727	4
5	16549	80
8	1	< 1
10	1308	6
Total	20726	

Additional file 3: Exposed differentiating xylem tissue after removing the bark. Tissue was collected immediately with a stainless steel blade to minimize oxidation.