UNIVERSIDADE ESTADUAL DE CAMPINAS

Rodrigo Tristan Lourenço

"ESTRUTURA GENÔMICA DE TRÊS MEGABASES DE DNA GENÔMICO (SHOTGUN) DE *Eucalyptus*: CONTEÚDO NUCLEOTÍDICO, SEQÜÊNCIAS REPETITIVAS E GENES"

Tese apresentada ao Instituto de Biologia para obtenção do Título de Mestre em Genética e Biologia Molecular na área de Genética de Microrganismos.

Orientador: Prof. Dr. Gonçalo Amarante Guimarães Pereira
Co-orientador: Prof. Dr. Dario Grattapaglia

Campinas
2004

"Diante da vastidão do tempo e da imensidão do universo, é um imenso prazer para mim dividir um planeta e uma época com você."

(Carl Sagan)

“So no one told you life was gonna be this way. Your job's a joke, you're broke, your love life's D.O.A. It's like you're always stuck in second gear. When it hasn't been your day, your week, your month, or even your year. But, I'll be there for you (When the rain starts to pour). I'll be there for you (Like I've been there before (…) No one could ever know me. No one could ever see me. Sometime the only one who knows what its like to be me. Someone to face the day with, make it through all the mess with. Someone I'll always laugh with, even under the worst I'm best with you. It's like you're always stuck in second gear. When it hasn't been your day, your week, your month, or even your year. I'll be there for you (…) ('Cause you're there for me too)”

(*Friends Theme Song*, Rembrants)

Aos meus pais, Ricardo e Marlene, e irmãs, Marisa e Isabela,

Pela educação, cuidados, carinho e apoio durante toda minha vida,

A Debora, que financiou parte deste trabalho com amor, paciência e companheirismo

Dedico

Agradecimentos

Aos meus pais, Ricardo e Marlene (Coês), e irmãs, Marisa (Má) e Isabela (Bebéla), minha família amada, pelo amor, apoio, amizade e compreensão durante toda minha vida e principalmente durante esta tese. Obrigado por serem muito mais do que uma família.

A Debora (Môdeu), que mesmo nos momentos mais difíceis foi minha maior amiga e incentivadora. Obrigado por ter sido minha companheira inseparável, mesmo estando, às vezes, a 1000 km de distância.

A Ana Carolina (Fubequinha), minha filha de coração, por me fazer cada vez mais feliz com atos e palavras puros de criança.

Ao Dario, pelo convite para fazer parte de sua equipe e pela oportunidade de trabalhar em seu laboratório na Embrapa Recursos Genéticos e Biotecnologia, e pelos ensinamentos, críticas e discussões.

Ao Gonçalo (Vamos lá, galera!), pela oportunidade de ser seu orientado e fazer parte de sua equipe no Laboratório de Genômica e Expressão, onde pude aprender não só a pesquisar, mas também a cultivar amizades e a me divertir com meu trabalho.

Ao Georgios Pappas, pelo grande e imprescindível apoio durante as análises dos dados gerados neste trabalho.

A minha família, principalmente meus avôs, Ilaura (Voóli), Cassemiro (Tudo bobeira!) e Yaya (Rainha-mãe), que durante minha estada em Campinas me ajudaram a manter meu peso e minha saúde física e mental, nas minhas visitas de fim-de-semana a São Paulo.

Aos meus amigos e irmãos da vida, Alexandre (Bixcate!), Antônio e Rodrigo (Batman e Robin), pela grande amizade e pelas discussões holísticas durante os intervalos de trabalho. Também agradeço aos amigos, Octávio (Oc), Ete, Saulo, Carol (Pitbinha) e Paulinha, pelo apoio, amizade.

A todos do Laboratório de Genômica e Expressão, principalmente Anders (Marsupial), Anderson (Homem-suor), Andréa, Fernando Tsukumo (Buda), Carlinha, Eliane, Marcão (Ó, grande mestre!), Marcos Renato (Maraujo), Naiara (O Pinto), Raquel e Victor Genú (Não me chamem de Vitão!), que me apoiaram e auxiliaram diretamente neste trabalho.

A "minha" estagiária Adriana (Bob-esponja), agradeço por ter me ajudado na otimização dos microssatélites e por ter sido minha primeira estagiária-cobaia.

A todos amigos do Laboratório de Genética de Plantas, principalmente à equipe Genolyptus, Alexandre Povoa, Clarissa (Clarissão), Eva, Evandro (Fala, Negadinha!), Guilherme, Isabela (sangue do meu sangue), Marília, Nathalia e Suzana, pelos momentos de discussão e distração e por terem dividido o laboratório comigo.

Aos amigos da Unicamp pelos momentos de amizade e por terem sido minha família durante os longos dias na universidade, Ana Flávia, Ana Paula, Anita, Babi, Bárbara, Cene, Diana, Fernandinha, Johana, Karla, Marcelo (Jean Reno), Odalys, Prianda, Vitão, e todos os outros que me ensinaram que amigos são para sempre, e que devemos tratar bem os estudantes de outras cidades.

Aos amigos da agremiação "Garotos Perdidos Bats Society" e a turma do cinema, Aluana, Cínthia Juzinha, Horácio, Kubota, Márcio, Mário, Miúdo, Rafael (estilo russo de bats), Renato, Patrícia e Véio, pelos grandes momentos esportivos e hilários nos domingos ensolarados de Campinas, regados a cerveja e tererê.

Aos amigos Felipe, Flávio e Lucas (Chewbacca), pela amizade, companheirismo, divisão de bens alimentícios e de teto, pelos momentos de descontração durante minha estadia em Campinas e por ter cedido um canto nas várias viagens que fiz até Campinas (matrícula, assinatura de termo compromisso, qualificação, reuniões, defesa, etc).

A todos amigos do Laboratório de Biotecnologia da Universidade Católica de Brasília, principalmente Alessandra e Lélia.

A empresa International Paper S.A., pelo fornecimento de material biológico.

Aos professores Doutores Giancarlo Pasquali, Rosana Brondani e Sergio Brommonschenkel, pelas sugestões e correções.

Aos Professores Doutores Gonçalo Amarante Guimarães Pereira, Michel Georges Alberts Vincentz e Sergio Hermínio Brommonshenkel, pela análise crítica e discussões na defesa desta tese.

A Universidade Estadual de Campinas e ao Departamento de Genética e Evolução, pela oportunidade de estudar mais uma vez numa grande universidade.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, pela bolsa de estudos.

A Finep/MCT pelo apoio financeiro para o desenvolvimento deste trabalho.

RESUMO

Com o intuito de obter uma visão da estrutura e composição do genoma de *Eucalyptus*, sequenciou-se aleatoriamente cerca de 10.000 fragmentos de DNA genômico de *Eucalyptus grandis* obtidos por meio de seqüenciamento por fragmentação randômica de DNA (*shotgun*) de uma biblioteca genômica, totalizando mais de 3,0 Mb válidos (phred >=20), isto é, cerca de 0,5% do genoma (640 Mpb). Depois de selecionadas quanto ao tamanho e qualidade, estas seqüências foram analisadas em termos do seu conteúdo nucleotídico, presença de regiões repetitivas e número de genes. Para análise do conteúdo de bases guanidílicas e citidílicas (GC) e do conteúdo de seqüências repetitivas utilizou-se o programa *RepeatMasker*, o qual indicou que as 10 mil seqüências continham, em média, 40,15% de GC. Aproximadamente 1,4% das bases pertenciam a seqüências transponíveis, distribuídas em 310 elementos repetitivos interespersados, dentre os quais 299 eram retroelementos, principalmente LTRs ("Long Terminal Repeats") e apenas 11 eram transposons. Também foram identificados 986 microssatélites e 1.636 seqüências de baixa complexidade. No total, cerca de 5,8% do genoma de *Eucalyptus* é composto por seqüências repetitivas. Para a identificação de genes putativos presentes, utilizou-se uma estratégia alternativa baseada na comparação deste banco genômico com bancos de ESTs ("Expressed Sequence Tags") de *Eucalyptus* utilizando o programa *GenESTate*, nomeando os genes identificados de acordo com o resultado do "BLAST" ("Basic Local Alignment Search Tool") encontrado para as ESTs. Também comparou-se todas as seqüências genômicas com o banco de dados não-redundante de proteínas do NCBI ("National Center for Biotechnology Information") com o intuito de identificar outros genes. Aproximadamente 44 seqüências similares a ESTs foram identificadas, contabilizando 2% do total de pares de bases analisado. É importante ressaltar a identificação de íntrons e éxons, além de regiões promotoras, a partir desta comparação, visto que estas estruturas não podem ser identificadas em ESTs. Cerca de 166 genes foram identificados a partir da comparação de todas as seqüências com o banco de dados de proteínas do NCBI por meio do protocolo "blastx-nr". Também foram identificados genes putativos para 16 tRNAs utilizando o programa *tRNAscan-SE*. Este banco de dados genômicos poderá ser utilizado no âmbito do Projeto Genolytpus para guiar o processo de ancoragem do mapa genético com o mapa físico, no desenvolvimento de novos marcadores do tipo microssatélites e na identificação de regiões promotoras.

ABSTRACT

In this work we intended to obtain an overview of the structure and composition of the Eucalyptus genome by sample sequencing 10.000 genomic DNA fragments obtained from a shotgun genomic library from *E. grandis*, that represents 3,0 Mbp of the *E. grandis* genome. The reads were filtered by their quality and length (phred value >=20; length >=150) and analyzed for their nucleotide content, repetitive patterns, repetitive elements and gene content. The program RepeatMasker was used to analyze the %GC content and repetitive patterns and elements. The results indicate that on average the *Eucalyptus* genome is composed of 40.15% of GC. From the total of the bases sequenced approximately 1.4% were located in transposons, distributed in 310 interespersed repetitive genetic elements, among which 299 classified as retroelements, mainly LTRs. We also identified 986 microsatellites and 1636 low complexity sequences. 5.8% of the sequenced bases were located on repetitive sequences. We used an alternative approach to identify putative genes by comparing the genomic sequences with a *Eucalyptus* ESTs database using the GenESTate software. We attributed putative functions using a pipeline were the éxons of each gene were put togheter and compared with protein domains data banks. This procedure avoids the misleading results obtained when comparing DNA sequences with sequences deposited in GenBank. The sequences were clustered using the CAP3 software, resulting in 766 agrupamentos contíguos and 5428 singletos, the former showing an average of 1200 bp. These 766 agrupamentos contíguos were compared with more than 5,000 *E. grandis* ESTs from mature leaf tissue and 6,000 *E. urophylla* ESTs from xylem. From the 766 agrupamentos contíguos we found 44 that showed high similarity to some ESTs. The coding portion of the sequences accounted for around 2% of the total sequences. It is important to highlight that by this approach it was possible to identify íntrons and éxons, beside core promoter regions, which can't be identified in the ESTs. Other 166 possible genes were identified among the genomic sequences by using blastx-nr in NCBI. We also identified putative genes responsible for 16 tRNAs using the tRNAscan-SE software. These sequences are being used in the Genolyptus Project for the development of novel randomly distributed microsatellites markers, for the identification of promoter regions and will be used to assist in the development of overgo-probes to be applied in the anchoring of the genetic map to the physical map.

1 INTRODUÇÃO

Árvores de *Eucalyptus* são utilizadas em todo o mundo com diversas finalidades, desde a produção de papel até óleos essenciais. Embora existam mais de 700 espécies do gênero, o melhoramento e a produção florestal são concentrados em apenas algumas espécies, principalmente *E. grandis*, *E. globulus*, *E. urophylla*, *E. calmadulensis*, *E. saligna* e *E. tereticornis*, por possuírem maior e melhor aproveitamento industrial e comercial.

Originário da Oceania, o gênero *Eucalyptus* passou a ser amplamente utilizado e distribuído geograficamente apenas a partir do século XX, quando métodos para a extração de celulose utilizando madeira desta espécie foram otimizados. Antes desta data, a principal utilidade de madeira de eucalipto era para carvão em ferrovias e usinas siderúrgicas.

No Brasil, o plantio de florestas de *Eucalyptus* teve início no século passado e, atualmente, fornece a matéria-prima para indústrias de papel, celulose e aço que contribuem significativamente para a balança comercial brasileira, criando centenas de milhares de empregos diretos e indiretos. As indústrias de papel e siderúrgica são as principais consumidoras das árvores de *Eucalyptus* produzidas no Brasil, sendo o país o maior produtor mundial de pasta celulósica e de papel derivados de fibra curta de eucalipto.

Apenas a partir da década de 60 do século passado, o gênero *Eucalyptus* passou a ser amplamente utilizado para produção de celulose e, desde então, pesquisadores vêm trabalhando em silvicultura, genética e melhoramento e, mais

recentemente, biologia molecular e genômica, com o intuito de aumentar a produtividade e a qualidade da madeira para diversos fins industriais. A preocupação central é de aumentar a produção de celulose por hectare de floresta plantada. Para isso, além do aumento da produtividade volumétrica de madeira por hectare, tem se tornado cada vez mais evidente a necessidade de reduzir o consumo específico de madeira, ou seja, melhorar a qualidade da madeira visando usar cada vez menos madeira para produzir a mesma tonelada de celulose, evitando-se, assim, a necessidade de incrementar a área plantada e reduzindo-se a geração de efluentes provenientes da produção de pasta celulósica.

Neste sentido, programas de pesquisa em genética, melhoramento, tecnologia da madeira e silvicultura têm sido conduzidos intensamente, tanto na indústria como no setor público, gerando informações cruciais para estes setores. As espécies mais utilizadas em plantações com fins industriais, *E. grandis*, *E. globulus* e *E. urophylla* têm sido também as mais estudadas geneticamente, seja do ponto de vista da genética clássica, seja do ponto de vista das tecnologias genômicas. Por exemplo, mapas genéticos já foram desenvolvidos para estas espécies e o tamanho total do mapa genético tem sido estimado entre 1.150 cM e 1.600 cM (Grattapaglia e Sederoff, 1994). Com base nos mapas genéticos disponíveis alguns estudos de mapeamento de QTL ("Quantitative Trait Loci") foram realizados gerando uma perspectiva de utilização da informação genômica no melhoramento destas espécies (Grattapaglia, 2001).

Apesar da sua crescente importância no cenário florestal mundial, estudos genômicos envolvendo espécies do gênero *Eucalyptus* ainda são incipientes. A maior parte dos esforços vêm concentrando-se no seqüenciamento de ESTs (Forests, 2001; Grattapaglia, 2003). Além do seu valor C, 0,58 pg, e do tamanho do genoma estimado

em 640 Mbp, para *E. grandis* (Grattapaglia e Bradshaw, 1994), muito pouco se sabe sobre a estrutura e a composição do genoma de espécies de *Eucalyptus*. Não sabemos, por exemplo, a freqüência e a estrutura de seqüências microssatélites, transposons e retrotransposons, genes, conteúdo nucleotídico do genoma em geral e destas estruturas. Esta informação será fundamental para o empreendimento de estudos e experimentação genômica envolvendo mapeamento genético, físico e clonagem posicional de genes.

Uma estratégia interessante e de grande poder informativo para gerar rapidamente um panorama geral do genoma de um organismo de interesse, mas ainda muito pouco utilizada em projetos genoma de plantas e animais, é a de *sample sequencing* (seqüenciamento por amostragem), também chamada de *random sequencing* (seqüenciamento aleatório), aliada a uma estratégia de clonagem às cegas (*shotgun*) para a obtenção de clones para seqüenciamento. Nesta estratégia, todo o DNA genômico, ou alguns longos trechos deste, são sonicados, nebulizados ou cortados com enzimas de restrição, e os fragmentos resultantes (em torno de 1,5 a 2,0 kb) são clonados e seqüenciados. A seqüência de nucleotídeos de cada fragmento é, então, analisada, com o intuito de descrever a estrutura e a composição do genoma por meio de amostragem. Esta estratégia contrasta com a versão clássica de seqüenciamento parcial baseada exclusivamente em genes expressos (ESTs), pois não busca a montagem de um catálogo de genes dos organismos, mas sim uma visão da estrutura do genoma como um todo.

Por eliminar a necessidade de mapas físicos para a montagem do genoma, esta estratégia foi utilizada inicialmente no estudo de genomas de microrganismos que não possuíam estes mapas (Karlyshev *et al.*, 1999; Wong *et al.*, 1999; Bell *et al.*, 2002), e

constituiu a base da abordagem, hoje comum, de seqüenciamento por fragmentação randômica de DNA (*shotgun*) para a montagem completa de genomas de alguns microrganismos, utilizada de forma pioneira para *Haemophilus* há cerca de 10 anos (Fleischmann *et al.*, 1995).

A estratégia de *sample sequencing* foi utilizada com o intuito de obter uma visão geral do conteúdo genômico de alguns organismos, inclusive de plantas, a partir de análises de menos de 1% até 10% do genoma total. Dentre as informações esperadas provenientes do seqüenciamento aleatório, pode-se citar aquelas relativas à riqueza e distribuição de microssatélites, elementos genéticos transponíveis e retrotransponíveis, genes, pseudogenes, regiões promotoras, conteúdo nucleotídico, tRNAs, dentre outras. Em *Trypanosoma cruzi*, diversos parâmetros foram analisados, dentre eles o conteúdo nucleotídico, riqueza de elementos genéticos transponíveis e retrotransponíveis e microssatélites, além de seu conteúdo de genes (Aguero *et al.*, 2000). O genoma do parasita unicelular *Paramecium* também foi analisado através de *sample sequencing*, mas a ênfase do trabalho foi sobre o conteúdo de genes (Sperling *et al.*, 2002). Em *Zea mays*, as análises concentraram-se sobre a riqueza de elementos genéticos e quantificação das famílias mais comuns, e algumas observações em relação a conteúdo de bases guanidílicas e citidílicas entre regiões codificantes e não-codificantes (Meyers *et al.*, 2001). De maneira geral, portanto, esta estratégia de geração de conhecimento da estrutura de um genoma de um determinado organismo tem sido ainda muito pouco utilizada.

No caso do *Eucalyptus*, a partir da comparação com o genoma de *Arabidopsis thaliana*, alguns genes homólogos envolvidos no controle de florescimento foram identificados e clonados (Southerton *et al.*, 1998), mas poucos foram estudados

profundamente. Informações como número de genes, características de éxons e íntrons, regiões promotoras, tamanho médio de genes, conteúdo GC, dentre outras, são raras ou nunca foram obtidas. Mesmo locos de microssatélites, já isolados em grande número e amplamente utilizados para a geração de mapas genéticos, foram pouco estudados em toda sua composição, e pouco se sabe sobre a riqueza e variação destes no genoma de *Eucalyptus*. A identificação e caracterização das principais classes de microssatélites, juntamente com a quantificação daquelas mais numerosas, e a avaliação da sua distribuição, são tarefas de fundamental importância para o desenvolvimento de novos marcadores que permitam a saturação e o refinamento do mapa genético.

A quantificação das famílias de elementos genéticos transponíveis e retrotransponíveis em *Eucalyptus* também não foi realizada ainda. Não se sabe quanto do genoma de *Eucalyptus* é ocupado por este tipo de seqüência, nem mesmo quais são as principais famílias e como estes elementos estão distribuídos. A identificação das famílias mais numerosas e de novos elementos genéticos repetitivos em *Eucalyptus* também nunca foi feita e será de grande valia para a condução de estudos genômicos integrados envolvendo, por exemplo, o mapeamento físico, a ancoragem de mapa genético com mapa físico e a busca de genes por abordagens posicionais.

O objetivo central desse trabalho foi, portanto, a obtenção de informações sobre a estrutura e a composição do genoma do *Eucalyptus* do ponto de vista das principais classes de DNA repetitivo, microssatélites, promotores e DNA codificante. Para isso, foi construída uma biblioteca genômica de insertos pequenos via seqüenciamento por fragmentação randômica de DNA (*shotgun*) e, em seguida, cerca de 3 Mpb do genoma de eucalipto foram seqüenciadas com base na estratégia de *sample sequencing*.

2 REVISÃO BIBLIOGRÁFICA

O gênero *Eucalyptus* L´Herit pertence à família *Myrtaceae*. Possui mais de 700 espécies, variedades e híbridos, distribuídos em 8 subgêneros, sendo o principal deles o S*ymphyomyrtus*, com mais de 300 espécies, dentre as quais estão as espécies mais plantadas para fins comerciais, *E. grandis*, *E. globulus*, *E. urophylla*, *E. calmadulensis*, *E. saligna* e *E. tereticornis* (FAO, 1981; Eldridge, 1994; Mora e Garcia, 2000). O gênero é originário da Austrália, com exceção de *E. urophylla* e *E. deglupta*, que são naturais do Timor e Papua Nova Guiné, respectivamente (Pryor, 1985).

Dentre as espécies de *Eucalyptus* há arbustivas e arbóreas, com algumas espécies alcançando até 100 metros de altura, e todas possuem troncos lenhosos. Em geral, são espécies perenes, com algumas poucas que perdem suas folhas durante as monções em seu habitat. Suas flores são agrupadas em corimbos, panículas ou mais freqüentemente, em umbelas. Elas são monóicas, hermafroditas e protândricas, e a planta é preferencialmente alógama, com alguma autogamia, sendo que os principais polinizadores são abelhas e formigas. Uma característica marcante do gênero é a presença de opérculo. Os frutos são capsulares e deiscentes, liberando sementes de cor preta a amarelada e de tamanho variável, de menos de 1 mm até 2 cm (Cavalcanti, 1963). As folhas variam de acordo com a maturidade da planta, mas normalmente são coriáceas, lanceoladas, com elevada quantidade de cutina e ricas em esclerênquima. Normalmente as folhas são alternadas, com algumas espécies mostrando folhas opostas. Os troncos apresentam pouca ramificação e normalmente são lisos e com casca descídua. Possuem, em geral, alburno delgado e claro, e cerne de amarelo a avermelhado (FAO, 1981).

O genoma de *Eucalyptus* tem entre 500 e 650 Mbp, distribuídos em 22 cromossomos pequenos na maior parte das espécies, e o seu valor C é igual a 0,58 pg (Grattapaglia e Bradshaw*,* 1994a). As espécies *E. grandis*, *E. globulus* e *E. urophylla* estão entre as mais cultivadas e também entre as mais estudadas, possuindo diversos mapas genéticos produzidos a partir da utilização de diversas metodologias. O tamanho total estimado do mapa genético tem entre 1.150 cM e 1.600 cM (Grattapaglia e Sederoff, 1994).

A cultura do *Eucalyptus* para exploração comercial da madeira teve início no Brasil por volta do início do século passado, por iniciativa da Companhia Paulista de Estradas de Ferro (Andrade, 1939). O interesse pelo eucalipto, na época, surgiu da necessidade de carvão para abastecer as locomotivas e de dormentes para ferrovias.

Atualmente, no Brasil, a cultura de eucalipto gera milhares de empregos, divisas e impostos, e florestas industriais de *Eucalyptus* estão amplamente espalhadas pelo território nacional, principalmente no Sul e Sudeste. Em 2000, quase 3 milhões de hectares estavam cobertos com esta cultura, a maior parte desta área nos estados de Minas Gerais e São Paulo (SBS, 2002), e as finalidades da madeira de *Eucalyptus* vão muito além de carvão para locomotivas e dormentes. O setor florestal brasileiro, principalmente aquele ligado à produção de *Eucalyptus* e seus derivados, tem elevada importância econômica e social, gerando 500 mil empregos diretos e 2 milhões indiretos, e foi responsável, em 1998, por US$ 13 bilhões (4%) do Produto Interno Bruto (PIB) brasileiro. Destes, US$ 5,4 bilhões provêm de exportações, e mais de US$ 1,5 bilhões são arrecadados em impostos. Os principais produtos obtidos a partir da madeira do *Eucalyptus* são a celulose e o carvão vegetal para a indústria siderúrgica, além de outros derivados da madeira, para móveis e construção, papelão, óleos e

outros (Mora e Garcia, 2000). Para 2003, analisando-se apenas o setor de papel e celulose, prevê-se um crescimento em produção e exportação, com uma balança favorável em cerca de US$ 2,5 bilhões e aumento da produção de papel e celulose de 5% e 13%, respectivamente (Bracelpa, 2003).

A maior parte dos 3 milhões de hectares cobertos por *Eucalyptus* pertence a empresas produtoras de papel e celulose, cultivada, em sua maioria, com clones ou híbridos interespecíficos superiores provenientes de programas de melhoramento genético. Graças a estes programas e ao manejo florestal, nas últimas 3 décadas a produtividade desta espécie foi elevada, a área de plantação foi triplicada e a qualidade da madeira foi melhorada de forma a atender à demanda das indústrias de papel e de celulose.

Entretanto, a cultura do *Eucalyptus* ainda não supre completamente esta demanda, tanto em qualidade de madeira como em produção, apesar do Brasil possuir alta competitividade. A produtividade e as áreas plantadas anualmente são consideradas insuficientes para suprir a demanda, o que ameaça as florestas nativas brasileiras (Scharf, 2003). Assim, a melhor estratégia para aumentar a produção e a produtividade das florestas de *Eucalyptus* seria através do melhoramento genético.

As empresas privadas, juntamente com algumas instituições públicas, como a Empresa Brasileira de Agropecuária (Embrapa) e algumas Universidades, possuem, no total, bancos de germoplasma contendo dezenas de acessos, híbridos e clones, com variabilidade genética considerável, e possuem também em seus quadros de funcionários alguns dos melhores melhoristas, fitotécnicos e geneticistas especialistas em *Eucalyptus* do mundo. Apesar disto, a evolução da indústria de papel e sua demanda por novos produtos, juntamente com o incremento em tecnologia que ocorre

em outros países produtores de *Eucalyptus*, fazem com que os esforços para o melhoramento da espécie, a fim de aumentar, ou pelo menos manter a competitividade do Brasil no mercado, devam ser cada vez mais intensos e objetivos.

Desta forma, novas metodologias e estratégias têm sido utilizadas com o intuito de elevar ainda mais a produção de *Eucalyptus* no Brasil, principalmente nos fatores produtividade, resistência a pragas e doenças e qualidade da madeira, mais especificamente quanto às características ligadas à produção industrial de derivados da madeira, de forma a incrementar a produção e reduzir a poluição proveniente dos variados processos pelo qual a madeira passa. Para se ter uma idéia, a redução do teor de lignina em 1% representaria uma economia de US$ 1 milhão para cada 300.000 ton de celulose produzidas. Essa economia corresponderia, em 2002, a mais de US$ 25 milhões, excluindo-se os benefícios indiretos associados devido à redução da poluição que o processo de retirada desta substância acarreta.

Dentre as novas tecnologias utilizadas para o melhoramento do *Eucalyptus*, bem como para o melhoramento de quase todas as culturas, destacam-se as pesquisas utilizando informações moleculares e genômicas, principalmente os marcadores moleculares, a transgenia e o conhecimento que provém de estudos genômicos.

Em termos de conhecimento e tecnologia na área molecular para o melhoramento de *Eucalyptus*, o Brasil vem se destacando já há alguns anos. Pesquisadores brasileiros já possuem bibliotecas bastante numerosas de *primers* para amplificação de regiões microssatélites para o gênero, as quais podem ser utilizadas em estudos de variabilidade genética, em seleção assistida por marcadores e em análises de certificação e proteção de cultivares. Diversos QTLs, principalmente ligados à qualidade da madeira, já foram detectados para estas espécies, e a procura por QTLs

de interesse direto para utilização no melhoramento é um dos principais objetivos das pesquisas envolvendo esta espécie.

Também já há no Brasil metodologia preliminar e pessoal treinado para transformação genética de tecidos de *Eucalyptus*, embora ainda existam problemas de dependência genotípica e de regeneração de plantas após a transformação (Brasileiro, ACM, com. pess.). Em geral, a transformação de espécies arbóreas é complicada, principalmente para *Eucalyptus* (González *et al*., 2002). Já existem testes de campo de eucalipto transgênico, mas ainda não há plantas transgênicas de *Eucalyptus* sendo cultivadas comercialmente, mais por obstáculos de regulamentação jurídica do que por capacitação. A utilização de transformação genética de *Eucalyptus*, como de várias outras espécies, é uma tecnologia promissora e aguarda apenas regulamentação e desenvolvimento de procedimentos eficientes de transformação e regeneração.

A alternativa de aumentar e melhorar a produção de *Eucalyptus* utilizando técnicas que acessam diretamente informações genômicas foi beneficiada pela redução dos custos e da elevação da eficiência das tecnologias de seqüenciamento de DNA. Diversos programas de seqüenciamento de genomas, das mais variadas espécies de procariotos e eucariotos, têm sido executados com sucesso em diversos laboratórios do mundo. Para o eucalipto, existem dois projetos brasileiros: o Projeto Forests, que envolve essencialmente o seqüenciamento de 100.000 ESTs, e o Projeto Genolyptus, que se fundamenta em uma estratégia integrada de melhoramento molecular (*molecular breeding*) envolvendo mapeamento genético e físico e seqüenciamento de ESTs, ações estas desenvolvidas sobre uma linha central de forte experimentação de campo e de investigação das propriedades tecnológicas da madeira (Genolyptus, 2003). A intenção deste projeto, portanto, é a integração da genômica com o

melhoramento convencional na busca de processos de seleção e desenvolvimento de clones de *Eucalyptus* com propriedades da madeira mais adequadas para a indústria. Isto representa o grande desafio que o Brasil tem pela frente, inclusive com várias culturas agrícolas, a partir do momento que os procedimentos de seqüenciamento já estão suficientemente dominados.

Dentre várias informações genômicas provenientes desses projetos, podemos citar como uma das mais importantes e prioritárias, de início, aquelas relativas à estrutura e à composição do genoma e, atualmente, ainda há conhecimento limitado sobre estes aspectos em genomas de eucariotos. Apenas aquelas espécies de plantas cujos genomas já foram totalmente seqüenciados estão bem representadas neste aspecto, como *Arabidopsis thaliana* (AGI, 2000) e *Oryza sativa* (Goff *et al.*, 2002; Yu *et al.*, 2002). Informações mais detalhadas para uma espécie perene arbórea são praticamente inexistentes, embora, em princípio, não exista uma expectativa de ocorrência de diferenças muito significativas.

No âmbito do Projeto Genolyptus, diversas ações estão previstas que demandam um conhecimento mais detalhado da estrutura e organização do genoma de *Eucalyptus*. Apesar de ter alguns mapas genéticos publicados, espécies do gênero *Eucalyptus* ainda carecem de dados referentes à estrutura e à composição do seu genoma. Uma análise mais detalhada do seu genoma faz-se necessária não apenas visando o aspecto descritivo, mas também a utilização dos dados em experimentos de mapeamento genético e físico. Entre outros, é interessante, por exemplo, estimar a abundância e distribuição de microssatélites, retrotransposons e transposons, genes putativos, segmentos de genes, pseudogenes e sua distribuição e conteúdo de nucleotídeos, principalmente conteúdo GC e sua correlação com estas estruturas.

2.1 Microssatélites

Microssatélites, ou *Simple Sequence Repeats* (SSR), são repetições em *tandem* de pequenos motivos de seqüência com 1 a 6 nucleotídeos, sendo encontrados amplamente distribuídos pelo genoma da maior parte dos eucariotos, embora também presente em procariotos (Litt e Luty, 1989). Acredita-se que o principal mecanismo por trás do surgimento e amplificação destas seqüências nos genomas seja o deslize (*slippage*) ou o mal-pareamento de motivos microssatélites durante a replicação. Durante a replicação de uma região repetitiva, as fitas de DNA separam-se e se reassociam de forma incorreta, o que geraria cópias de trechos de DNA (alelos) com diferentes tamanhos ou números de repetições de um determinado motivo no próximo ciclo de replicação por meio da inserção ou deleção de uma unidade de repetição (Schlotterer Tautz, 1992).

Essas seqüências apresentam uma elevada taxa de mutação, da ordem de 10^{-3}, resultando em uma ampla variação no número de unidades repetidas, o que faz com que marcadores baseados em microssatélites sejam altamente informativos e amplamente utilizados em programas de melhoramento de plantas, em mapeamento genético e em identificação de indivíduos (Byrne *et al.*, 1996; Goldstein e Schlotterer, 1999).

Goldstein e Schlotterer (1999), citando Freimer e Slatikin (1996), afirmam que apesar de serem bastante numerosas, regiões microssatélites são pouco caracterizadas em nível de seqüência, já que a maior parte dos locos microssatélites estudados até o momento foram isolados de bibliotecas genômicas utilizando-se sondas oligonucleotídicas. Desta forma, torna-se complicado avaliar a variação de

tamanho destas regiões em populações. Some-se a isto a impossibilidade de avaliar, por meio desta metodologia de isolamento de seqüências microssatélites, qual é a variabilidade e quantidade de tipos de seqüências no genoma. Também devemos citar que microssatélites contendo motivos formados por nucleotídeos que possam formar estruturas secundárias, e.g. (AT)n/(TA)n, não são passíveis de serem identificados por meio de bibliotecas enriquecidas. Há que se citar também a dificuldade em construir bibliotecas para tri-, tetra- e pentanucleotídicos sem saber suas freqüências, já que as possibilidades de composição de motivos aumentam em progressão geométrica de acordo com o aumento do tamanho do motivo.

Em *Eucalyptus,* marcadores baseados em microssatélites foram desenvolvidos a partir de bibliotecas genômicas enriquecidas (Brondani *et al.*, 1998). Embora se saiba, por exemplo, que seqüências repetitivas de (AG)n são mais freqüentes do que (AC)n (Brondani e Grattapaglia, 1997), o procedimento utilizado para o seu isolamento não permitiu estimar a sua abundância no genoma e muito menos a freqüência de ocorrência de tri e tetranucleotídicos. Uma busca mais detalhada por estas seqüências e o desenvolvimento dos respectivos marcadores são necessários, tanto do ponto de vista descritivo para a espécie como para sua utilização em programas de melhoramento, em certificação e em proteção de variedades. Vale destacar que *Eucalyptus* é até agora a única espécie vegetal cujo registro de proteção de cultivares no Ministério da Agricultura prevê a utilização de marcadores moleculares.

Regiões microssatélites são altamente conservadas entre espécies e, em alguns casos, até mesmo entre gêneros, tanto em animais como em plantas. Em *Eucalyptus,* microssatélites mostram elevada conservação entre *pedigrees*, entre populações, e até mesmo entre espécies de um mesmo subgênero, principalmente no subgênero

Symphyomyrtus, dentro do qual estão as espécies mais plantadas comercialmente, mas a conservação de locos, ou transferibilidade, entre subgêneros não é tão notável (Brondani *et al.*, 1998). O desenvolvimento de *primers* microssatélites deve levar em consideração a maximização de sua utilização, ou seja, quanto maior a conservação de locos microssatélite menor será seu custo fixo relativo. Em *Eucalyptus,* locos microssatélite conservados entre espécies, pelo menos entre as mais plantadas, são essenciais, já que programas de melhoramento normalmente lançam mão de várias espécies e híbridos (Byrne *et al.*, 1996).

Apesar de presentes em todos os eucariotos, de estar ubiqüamente presente nos seus genomas e de apresentarem elevadas taxas de mutação, nem todos locos microssatélites identificados são aproveitáveis para o mapeamento genético. A seleção dos locos informativos é baseada nas características destes locos e nos resultados que estes microssatélites permitirão obter na identificação de segregação nas progênies resultantes de cruzamentos controlados. Dentre os principais parâmetros utilizados nesta seleção estão: o tipo de repetição (di-, tri-, tetra- ou pentanucleotídicos), o número de repetições, o número de um mesmo loco no genoma, o número de alelos ao loco, a heterozigosidade, dentre outros (Brondani *et al.*, 1998; Edwards *et al.*, 1998; Garner, 2002)

Locos contendo repetições de um único nucleotídeo não têm praticamente utilidade, a priori, em genotipagem, pois a definição dos alelos torna-se bastante trabalhosa, levando freqüentemente a erro, e o efeito denominado *stutter*, que é caracterizado pelo surgimento de “sombras” de bandas com uma repetição a mais ou a menos que o alelo, torna-se ainda mais pronunciado do que em dinucleotídicos, onde *stutter* é um dos principais complicadores da análise. O número de locos no genoma e

o número de repetições total no microssatélite parecem mostrar uma relação decrescente com o tipo de motivo, sendo mais comum encontrarmos microssatélites dinucleotídicos do que trinucleotídicos, e assim por diante (Edwards *et al.*, 1998; Elgar *et al.*, 1999; Aguero *et al.*, 2000). Microssatélites tri-, tetra- e pentanucleotídicos apresentam menor *stutter* ou, pelo menos, é mais fácil distinguir alelos de *stutter* nestes e, por conseqüência, seus alelos parecem ter uma melhor definição, facilitando a análise, aumentando a eficiência na distinção de indivíduos e reduzindo os erros de genotipagem.

Notou-se que microssatélites trinucleotídicos são menos freqüentes que dinucleotídicos no genoma total, mas em regiões codificantes parece ocorrer o inverso, ou seja, seqüências microssatélites compostas por motivos trinucleotídicos são mais comuns do que qualquer outro motivo. É possível que isto resulte de seleção negativa contra mutações que alterem a fase de leitura em regiões codificantes (Morgante *et al.*, 2002).

Em espécies poliplóides e em espécies que passaram por eventos de duplicação de segmentos, ou em diplóides originadas de poliplóides como *Arabidopsis*, por exemplo (AGI, 2000; Bancroft, 2001), locos microssatélites podem estar presentes em múltiplas cópias, o que é indesejável do ponto de vista de sua utilização em genotipagem. Para esta finalidade, é preferível usar locos microssatélites de cópia única, cuja informação alélica seja facilmente visualizada, permitindo uma rápida análise.

Aparentemente, há uma correlação entre o número de repetições presentes num loco microssatélite e, portanto, o tamanho final da seqüência, com seu número de alelos e heterozigosidade (Weber, 1990; Byrne *et al.*, 1996; Cho *et al.*, 2000). Desta

forma, regiões microssatélites maiores seriam mais úteis para análise genética, possibilitando a geração de maior informação por loco, principalmente em estudos de paternidade e genética de populações. Byrne *et al.* (1996) descrevem, também, que microssatélites compostos, aqueles que são formados por repetições contíguas ou adjacentes de diferentes motivos, parecem seguir este padrão e podem, em alguns casos, mostrar maior polimorfismo por apresentarem mais de um tipo de motivo e uma região microssatélite maior. Entretanto, estas correlações não foram confirmadas em *Eucalyptus* (Brondani *et al.*, 1998).

Outras características estruturais podem estar correlacionadas com o nível de polimorfismo de locos microssatélites. Correlações negativas foram observadas entre o conteúdo GC de regiões flanqueadoras de locos microssatélites e seu nível de polimorfismo (Glenn *et al.*, 1996), enquanto outros trabalhos não conseguiram confirmar este padrão (Balloux *et al.*, 1998).

Alguns estudos mostraram que regiões não-codificantes continham maior associação com microssatélites em humanos, *Caenorhabditis* sp. e *Saccharomyces* sp (Hancock, 1995) ou com regiões repetitivas, mais especificamente retrotransposons, em *Hordeum vulgare* (Ramsay *et al.*, 1999). Novos estudos trouxeram ainda melhores perspectivas na utilização de microssatélites como marcadores genéticos. Morgante *et al.* (2002) avaliaram 5 diferentes espécies de plantas e encontraram uma associação significativa entre locos microssatélites e regiões codificantes ou ricas em genes. Em regiões codificantes foram identificadas diferentes pressões de seleção influenciando a presença de microssatélites, sendo que a menor pressão de seleção presente nas regiões 5' u*ntranslated region* (região não-traduzida) de ESTs parece ter relação com a maior freqüência de microssatélites polimórficos observada, enquanto que regiões

traduzidas apresentaram menor freqüência. Esta maior freqüência de microssatélites próximos a genes seria, do ponto de vista da aplicação genética destes marcadores, bastante interessante. Resultados corroborando a relação entre regiões microssatélites e regiões codificantes foram encontrados em outras espécies, como *Arabidopsis*, artrópodes, vertebrados e outros (Toth *et al.*, 2000).

A relação diretamente proporcional entre o tamanho do genoma e o número de microssatélites descrita por Hancock (1995) em diversos organismos não parece ser verdadeira em plantas. Na verdade, o número de seqüências microssatélites mostra aparentemente uma relação diretamente proporcional com a fração de DNA cópia-única (Morgante *et al.*, 2002). Desta forma, genomas de plantas mais compactos, mas tão complexos em conteúdo gênico quanto genomas mais extensos, permitiriam uma maior aproximação de locos microssatélites a genes, já que haveria um mesmo número de genes e microssatélites, mas em um menor espaço.

2.2 Elementos Genéticos Repetitivos

Elementos genéticos repetitivos têm importância fundamental no Reino *Plantae*, principalmente no que tange à composição e estrutura de genomas, pois são bastante numerosos na maior parte das espécies de eucariotos, principalmente naquelas com maior genoma.

Inicialmente descritos em milho como elementos controladores (McClintock, 1950), este tipo de estrutura tem sido bastante estudado em diversos organismos nas últimas décadas, e muitas funções e particularidades estruturais ainda estão por ser descritas.

Um dos principais tipos de transposons encontrados em plantas são os chamados retrotransposons. O mecanismo básico de movimentação e multiplicação desses elementos é baseado num modelo onde é transcrito um RNA intermediário, o qual é convertido em DNA extra-cromossomal pela enzima transcritase reversa. Este DNA é então inserido no genoma. Elementos de DNA transponíveis utilizam uma estratégia diferente de transposição, por meio de mecanismos de excisão e reparo, e não aumentam muito o tamanho do genoma de plantas pois, em geral, os próprios elementos genéticos transponíveis removidos são inseridos, ao contrário de retrotransposons, onde uma cópia do elemento genético retrotransponível é inserido em um outro loco (Kumar e Bennetzen, 1999).

Em *Z. mays* (milho), elementos geneticamente transponíveis correspondem a, aproximadamente, 50% de todo o genoma. Nesta espécie, cerca de 25% do genoma é composto por apenas 5 famílias diferentes de retrotransposons (Meyers *et al.*, 2001). Já em *Arabidopsis*, apenas 10% do genoma corresponde a este tipo de estrutura, com elementos genéticos transponíveis bem balanceados entre os diversos tipos (AGI, 2000) .

Além de sua importância na constituição de genomas de plantas, transposons também têm importância em estudos de indução de mutação, podendo ser utilizados em identificação de genes em análises funcionais, e como marcadores moleculares em estudos filogenéticos e de biodiversidade. Quando inseridos dentro ou próximo a genes, transposons e retrotransposons normalmente alteram a função deste gene (Watson, 1992; Kumar e Bennetzen, 1999), permitindo estudos funcionais através de silenciamento gênico.

Os retrotransposons normalmente são subdivididos em duas classes principais, os *Long Terminal* Repeat (LTR) e os não-LTR. Em plantas, ambos parecem estar bem representados, mas elementos retrotransposons LTR encontram-se em muito maior número, sendo que em milho nenhum retrotransposon não-LTR encontra-se presente em grande quantidade. Em humanos ocorre o inverso: elementos retrotransposons não-LTR são mais numerosos (Meyers *et al.*, 2001).

Retrotransposons LTR são subdivididos em duas outras classes, Ty1-*copia* e Ty3-*gypsy,* os quais diferem apenas na ordem em que seus genes estão dispostos e em algumas partes de suas seqüências. Retrotransposons não-LTR são representados principalmente por elementos *Long Interespersed Repetitive Element* (LINE) e *Short Interespersed Repetitive Element* (SINE).

Tanto retrotransposons do tipo Ty1-*copia* como Ty3-*gypsy*, bem como LINE e SINE, encontram-se distribuídos de forma aleatória no genoma de eucariotos. Há variação nesta distribuição e no número de elementos, de acordo com o táxon analisado e, em algumas espécies, em locais específicos do genoma foram identificadas regiões cuja concentração destes elementos era maior. Entretanto, estes elementos também parecem ter inserção preferencial por alguns sítios. Elementos Ty1-*copia* têm preferência de inserção em eucromatina, enquanto que Ty3-*gypsy* prefere inserção em heterocromatina. LINEs preferem inserção em regiões teloméricas ou próximo a outros elementos LINE. Não se sabe exatamente se esta preferência por sítios ocorre por uma real afinidade destes elementos pelo sítio em questão ou se devido a uma deficiência ou menor pressão de seleção na remoção destes elementos nestes locais (Kumar e Bennetzen, 1999).

Cada tipo de retrotransposon (Ty1-*copia*, Ty3-*gypsy*, SINE e LINE) evoluiu de forma distinta e, por isto, são subdivididos em diversas famílias, as quais possuem em comum a organização dos genes que caracteriza cada tipo de retrotransposon, mas apresentam diferenças adquiridas com a evolução. Em plantas, observou-se uma menor taxa de *turnover*, ou seja, há mais retrotransposons sendo inseridos do que sendo retirados, o que faz com que este tipo de elemento repetitivo constitua a maior parte do genoma de plantas e seja um dos principais responsáveis pela grande diferença de tamanho do genoma no reino vegetal.

A presença de um maior número de retrotransposons, por sua vez, faz com que haja maior número de sítios de mutação, gerando novas famílias. Por isto há, em plantas, maior variedade de famílias e quantidade de retrotransposons. Também influi na heterogeneidade de famílias uma característica reprodutiva inerente às plantas. Devido a formação de órgãos reprodutivos a partir de tecidos meristemáticos, mutações de retrotransposons que ocorreram na planta podem ser passadas diretamente, e de forma independente, para sua prole. Isto faz com que diferentes espécies, e até mesmo populações, apresentem certa heterogeneidade de elementos retrotransponíveis.

Além de influenciar bastante no tamanho do genoma, retrotransposons também exercem função no rearranjo genômico. Quando inseridos em genes e após uma possível aquisição de partes destes, a translocação de um retrotransposons poderia multiplicar o número de cópias destes genes ou fragmentos de genes. Além disto, um evento raro, mas não impossível, como a ação da enzima de transcrição reversa proveniente de retrotransposons sobre mRNAs normais de uma célula, poderia levar a inserção de pseudogenes sem íntrons que, futuramente, poderão sofrer alterações e passar a funcionar como genes (White *et al.*, 1994). Durante a recombinação,

retrotransposons podem também agir na reorganização do genoma, já que seu elevado número de cópias pode originar pontos de recombinação, gerando deleções, duplicações, inversões e translocações (Kumar e Bennetzen, 1999).

A quantificação de elementos transponíveis foi beneficiada, a partir do início das análises genômicas no final da década de 90, pelo grande número de seqüências de DNA de diversos organismos começou a ser depositado em bancos de dados, gerando um grande volume de informação para análise. Como elementos transponíveis ativos e geneticamente identificáveis são minoria, a quantificação destes antes da era genômica era sempre subestimada. Além disto, diversas famílias de transposons não haviam sido identificadas.

Não há ainda estudos de quantificação e identificação das principais famílias de elementos genéticos transponíveis na maior parte de espécies de plantas, inclusive em *Eucalyptus*, e este tipo de análise mostra-se importante para futuros estudos de filogenia, aplicação de elementos genéticos repetitivos como marcadores moleculares, organização do genoma e, como um exemplo prático e de imediata aplicação, auxiliando a montagem do mapa físico de *Eucalyptus*. O mapa físico de *Eucalyptus* tem início a partir da obtenção de bibliotecas de insertos longos (*Bacterial Artificial Chromosomes,* BACs) e seqüenciamento de suas extremidades. A partir da comparação de todas as extremidades de BACs é possível identificar BACs ligados entre si, e a presença de seqüências repetitivas em grande quantidade torna este procedimento de montagem ainda mais complicado, gerando resultados espúrios e errôneos. Desta forma, a identificação das principais famílias de elementos genéticos repetitivos permitirá o mascaramento eficiente destas estruturas transponíveis, permitindo maiores agilidade e acurácia na montagem dos contíguos de BACs.

2.3 Genes, Fragmentos de Genes e Pseudogenes

A anotação genômica, extração de significado biológico de seqüências genômicas anônimas, é tarefa bastante complicada, e diversos esforços foram feitos para sua automação (Aubourg e Rouzé, 2001).

A identificação de genes a partir da comparação de seqüências procurando por regiões de similaridade já é uma ferramenta amplamente utilizada. Para isto, em geral, pesquisadores utilizam o algoritmo *Basic Local Alignment Sequence Tool* (BLAST) (Altschul *et al.*, 1997). A partir do alinhamento de seqüências, identificação das regiões alinhadas e pontuação destas a partir de uma matriz padrão, pode-se descobrir seqüências similares e com diversos graus de identidade, permitindo qualificar seqüências não-identificadas e descrever possíveis funções (Pertsemlidis e Fondon, 2001).

Outra estratégia bastante útil quando se tem em mãos as ferramentas e informações necessárias, é a comparação das seqüências anônimas com ESTs, de preferência da mesma espécie, indicando com precisão genes putativos nas seqüências anônimas. A função destes genes seria atribuída diretamente de acordo com a função putativa da EST semelhante.

Bibliotecas de ESTs são obtidas por meio de extração de mRNA de um organismo, os quais são tratados posteriormente para gerar cDNAs, que são mais estáveis e permitem seqüenciamento e outras manipulações impossíveis de serem feitas com mRNA. Entretanto, informação estrutural é perdida quando se utiliza esta estratégia. Devido ao processamento do mRNA no núcleo, os íntrons são retirados e não estão presentes nos ESTs. Também não é possível identificar regiões promotoras,

que se situam a montante do gene e não fazem parte, nem mesmo, do transcrito primário.

Desta forma, estratégias alternativas ao seqüenciamento de ESTs são essenciais para se obter informações gerais sobre o genoma como seqüências de íntrons, regiões promotoras, bordas éxons/íntrons e a reconstrução de seqüências UTR.

2.4 Conteúdo Nucleotídico

O conteúdo de nucleotídeos das seqüências obtidas a partir de seqüenciamento aleatório refere-se a estudos que procuram tendências de determinados nucleotídeos em algumas regiões do genoma, procurando por desvios que possam indicar padrões e características intrínsecas de cada genoma. Regiões ricas em GC podem apresentar propriedades biológicas diferentes daquelas pobres em GC, e este tipo de correlação pode fornecer mais subsídios para uma identificação mais rápida e eficiente de determinadas estruturas.

A composição GC difere notavelmente entre diversos organismos e dentro do próprio genoma. Entre genomas, a composição de GC varia de 26 a 65%. Como exemplo, podemos citar *A. thaliana*, com 40,5% de GC, *Lycopersicum* com 36,3%, *Z. mays* com 55,3% e *Oryza sativa* com 51,0% de GC, o que nos permite observar que monocotiledôneas tendem a possuir maior porcentagem de GC do que dicotiledôneas. Este padrão também é observado quando comparamos éxons e íntrons de monocotiledôneas com éxons e íntrons de dicotiledôneas (Sato *et al*., 1998). Outro fato marcante é a diferença entre conteúdo GC dentro de cada espécie. Quando observado

o conteúdo GC entre íntrons e éxons de *Arabidopsis*, encontramos que em íntrons há menos nucleotídeos G e C, 32%, do que em éxons, 43%, e mesmo dentro de éxons há uma queda deste conteúdo a partir do início do ponto de tradução na direção 3' do éxon (Mizuno e Kanehisa, 1994).

Algumas correlações entre o conteúdo GC e genes têm sido demonstradas em alguns estudos (Comings, 1978; Mizuno e Kanehisa, 1994). Acredita-se que regiões ricas em genes normalmente possuem um maior conteúdo GC. Em humanos, a presença de *isochores*, grandes trechos de DNA que se diferenciam em quantidade de genes e conteúdo GC, e a correlação positiva entre presença de genes e elevado teor GC está relacionada com uma maior atividade de transcrição e necessidade de uma capacidade de adaptação de conformação física da molécula de DNA para transcrição e regulação de genes (Vinogradov, 2003). Em monocotiledôneas, acredita-se que possa haver ainda seleção por regiões ricas em GC em regiões gênicas devido à termoestabilidade maior que os nucleotídeos GC proporcionam.

Também em relação à composição de nucleotídeos, íntrons e éxons parecem diferir significativamente em conteúdo GC, e em monocotiledôneas este padrão é bastante significativo. Na verdade, quanto mais extenso o genoma, e quanto mais rico em AT, mais evidente se torna este padrão. Regiões para reconhecimento de íntrons no processamento de mRNA diferem bastante entre diferentes organismos. Vertebrados e leveduras, por exemplo, possuem padrões nestas regiões bastante diferentes dos padrões em plantas dicotiledôneas, as quais são caracteristicamente ricas em nucleotídeos A e U (Goodall e Filipowicz, 1991). Desta forma, a caracterização da composição GC de íntrons de *Eucalyptus* seria bastante interessante para prover

futuros projetos de identificação de regiões codificantes, principalmente na identificação de íntrons e éxons.

Em microssatélites, o conteúdo GC pode estar ligado ao nível de polimorfismo do loco, o que poderia ser interessante durante a seleção de locos microssatélites para desenho de *primers*. Segundo Glenn *et al.* (1996), em crocodilianos notou-se uma correlação entre o conteúdo GC das regiões flanqueadoras dos microssatélites e o nível de polimorfismo, sendo que quanto menor o conteúdo GC, maior foi o polimorfismo apresentado.

Algumas abordagens têm sido utilizadas para o estudo da estrutura de genomas eucariotos incluindo o isolamento de frações com diferentes composições de bases GC (Bernardi, 1989) ou a utilização de hibridização fluorescente *in situ* (FISH) para a investigação da distribuição de seqüências repetitivas no genoma de plantas (Pearce *et al.*, 1996). Obviamente que a metodologia mais informativa para estudos de estrutura de genoma é a análise de seqüência de DNA. Entretanto, embora o seqüenciamento tenha se tornado significativamente mais rápido e econômico, o seqüenciamento de genomas completos de eucariotos ainda se constitui em tarefa complexa, demorada e muito cara com a atual tecnologia. Uma opção estratégica interessante e muito informativa para o estudo de estrutura de genomas tem sido o seqüenciamento por amostragem (*sample sequencing*), (Elgar *et al.*, 1999) de uma biblioteca genômica tipo *shotgun*. Esta estratégia tem sido aplicada com sucesso, por exemplo, na investigação da estrutura do genoma de milho (Meyers *et al.*, 2001).

2.5 Genes de RNAs transportadores

Outra estrutura interessante e bastante numerosa em genomas eucariotos são genes codificantes de tRNAs. Os genes de tRNAs compõem a família gênica mais numerosa em um genoma eucariótico típico, onde centenas de tRNAs são encontrados. A identificação dos genes de tRNA de um determinado organismo auxilia na caracterização de seu *codon usage*, ou seja, quais codons são preferencialmente utilizados para cada aminoácido (Lowe e Eddy, 1997). Mais de 4 mil moléculas de tRNA ou seqüências de genes codificando tRNAs já foram identificadas em diversos organismos e muito mais ainda devem ser identificadas.

O seqüenciamento de genomas completos e a identificação de tRNAs permitiram um melhor entendimento de sua organização e estrutura molecular específicos para cada organismo, auxiliando e esclarecendo com maior precisão seu funcionamento. Diferenças significativas foram identificadas entre tRNAs dos 3 principais domínios da vida: Eucarya, Bacteria e Archaea. Dentro de cada domínio, as variações mostraram-se proporcionalmente menores conforme o nível de agrupamento (Marck e Grosjean, 2002).

Para *Eucalyptus* não há ainda nenhum estudo onde tRNAs foram identificados, muito menos comparando-se a variação entre este gênero com outras espécies de plantas cujo seqüenciamento do genoma já se encontra concluído ou próximo da conclusão, como *Arabidopsis* sp e *Oryza* spp.

3 OBJETIVO

O objetivo deste trabalho de mestrado é a geração e a caracterização de um banco de seqüências genômicas de *Eucalyptus grandis* utilizando a estratégia de seqüenciamento por amostragem (*sample sequencing*) de uma biblioteca para seqüenciamento por fragmentação randômica de DNA (*shotgun*). A partir da construção deste banco, serão investigados alguns aspectos específicos da estrutura e organização do genoma do eucalipto conforme detalhado nos objetivos listados abaixo.

3.1 Objetivos Específicos

- Construção de uma biblioteca genômica para seqüenciamento por fragmentação randômica de DNA (*shotgun*) de *E. grandis* com tamanho médio de inserto de 1,5 kb.
- Seqüenciamento aleatório de um mínimo de 5000 e um máximo de 10.000 seqüências genômicas, tendo como critério para aceitação um mínimo de 150 pb a *phred* 20, cobrindo, portanto, entre 1,5 e 3,0 Mpb, de DNA amostrando, aproximadamente 0,25 a 0,5 % do genoma de *Eucalyptus*.
- Análise detalhada das seqüências geradas e caracterização das mesmas quanto à composição de DNA codificante versus não-codificante.

- Identificação de seqüências microssatélites, com especial atenção a tri e tetranucleotídicos, para quantificar a riqueza e variedade de motivos microssatélites presentes no genoma de *Eucalyptus*.
- Desenho, caracterização e otimização de pares de *primers* para geração de marcadores hipervariáveis para mapeamento genético e identificação individual.
- Identificação de características estruturais de regiões contendo microssatélites e de correlações destas características com o nível de polimorfismo apresentado pelo loco.
- Caracterização de transposons, e identificação da quantidade e a qualidade destes, agrupando-os em famílias, procurando descrever as principais classes e sua distribuição.
- Identificação de genes, íntrons e éxons a partir da comparação das seqüências genômicas com o banco de dados de ESTs construído a partir de mRNA de *E. grandis* e *E. urophylla* no âmbito do Projeto Genolyptus.
- Identificação de genes putativos a partir de comparação das seqüências obtidas com bancos de dados públicos (principalmente *Arabidopsis* e *Oryza)* com ênfase especial para a identificação de genes candidatos a funções importantes na formação da madeira (lignina, celulose, hemicelulose) e resistência a doenças.
- Identificação de regiões promotoras a partir de genes completos e éxons iniciais identificados nas seqüências genômicas.
- Identificação de genes putativos para tRNAs.

- Análise comparativa do banco genômico gerado com um banco de ESTs de *Eucalyptus* quanto ao conteúdo GC.

4 MATERIAIS E MÉTODOS

4.1 Material Biológico

Para a obtenção de DNA de *Eucalyptus* foram utilizadas folhas de uma árvore adulta de *E. grandis* com cerca de 55 m pertencente à *International Paper* do Brasil, com sede em Mogi Guaçu, SP. Esta árvore possuía cerca de 30 anos e tem origem em uma semente proveniente de *Coff`s Harbor*, Austrália. Na época do corte, este indivíduo encontrava-se em perfeito estado fitossanitário.

O DNA foi extraído segundo protocolo modificado de Doyle & Doyle (Doyle e Doyle, 1986; Ferreira e Grattapaglia, 1998). Após a extração, o DNA foi solubilizado em 200 μl de tampão TE (10 mM Tris-HCl pH 8,0; 1 mM EDTA) contendo RNAse (10 μg/ml; Amersham Biosciences) e sua concentração aproximada foi estimada em cerca de 500 ng/μl utilizando gel de agarose 1%.

4.2 Contrução da Biblioteca Genômica

4.2.1 Sonicação

Cerca de 8 alíquotas de 10 μl de DNA a 500 ng/μl foram colocadas em tubos de 0,2 ml. Estes tubos foram colocados sobre gelo em um adaptador onde uma fonte de

ultra-som causou a ruptura do DNA em tamanhos variáveis. Metade dos tubos foram sonicados por 25 segundos, enquanto a outra metade foi sonicada por 30 segundos, sendo que a temperatura durante a sonicação foi mantida entre 0 ºC e 4 ºC para que a quebra do DNA fornecesse fragmentos de todos os tamanhos sem qualquer tipo de viés gerado por aumento de temperatura (Bankier *et al.*, 1986).

Os volumes dos 4 tubos sonicados por 25 segundos foram reunidos em um só, da mesma forma que os 4 tubos sonicados por 30 segundos, após confirmação em gel de agarose 1% de que havia DNA sonicado de boa qualidade nestes tubos.

4.2.2 Tratamento das extremidades dos fragmentos

Após a sonicação foram obtidos fragmentos de DNA com pontas não-coesivas, o que inviabiliza sua ligação ao vetor utilizado (pUC 18). Diante disto, os fragmentos foram tratados com T4 DNA polimerase (Gibco) e klenow (Invitrogen). Nesta reação, as enzimas completam as extremidades dos fragmentos a partir da incorporação de nucleotídeos na direção 5'-3' utilizando a ponta fita-simples como molde, sendo que a atividade exonucleotídica 3'-5' da T4 DNA polimerase retira as fitas simples que resultam da incorporação anterior, tornando as extremidades coesivas para posterior ligação com o vetor pUC 18 (Amershan Biosciences).

Reação de tratamento dos fragmentos:

Reagentes	Volume (1 reação)
DNA sonicado (500 ng/μl)	39,0 μl
Tampão T4 (10x; Gibco)	5,0 μl
T4 DNA pol (5 u/μl;Gibco)	2,0 μl
dNTP 1,25 mM (Amersham Biosciences)	2,0 μl

Reação incubada a 37ºC por 36 minutos;

Adição de 1,7 μl de klenow (6 u/μl; Invitrogen);

Reação incubada à temperatura ambiente por 20 minutos;

Reação colocada em gelo por 5 minutos.

4.2.3 Separação dos fragmentos

O DNA fragmentado e tratado enzimaticamente foi submetido à eletroforese em gel *low-melting* 1% por 3 horas a 50 V. Sob luz ultravioleta os fragmentos compreendidos entre 0,5 a 2 kb, para o DNA sonicado por 30 segundos, e de 2 a 4 kb, para o DNA sonicado por 25 segundos, foram selecionados e removidos do gel.

Os fragmentos de DNA foram colocados em tubos de microcentrífuga de 1,5 ml e eluídos do gel de agarose segundo protocolo de extração de DNA de agarose com fenol, descrito em Ausubel (2002). Após secagem do precipitado, este foi solubilizado em 20 μl de água.

4.2.4 Ligação dos fragmentos com o vetor

Para clonagem, foi utilizado o kit *SureClone* pUC 18 (Amersham Biosciences).

Reagentes	Volume (1 reação)
H_2O:q.s.	12,0 μl
Vetor (pUC18 10 ng/μl)	1,5 μl
TampãoT4 DNA ligase (2x)	6,0 μl
T4 DNA ligase (40 u/μl)	1,5 μl
DNA tratado (500 ng/μl)	3,0 μl

A ligação dos fragmentos compreendidos entre 0,5 a 2 kb ocorreu à 16 ºC *overnight*, enquanto que a ligação de fragmentos entre 2 a 4 kb ocorreu a 4 ºC *overnight*.

4.2.5 Transformação

As reações de ligação foram dialisadas visado a retirada de excesso de sal antes de efetuar a transformação. Todo o volume da reação de ligação foi colocado sobre filtro (Millipore) de 0,025 μM e este foi colocado sobre água milli-Q autoclavada durante 20 minutos.

Após a diálise a reação de ligação foi transferida para um novo tubo de microcentrífuga de 0,2 ml. A ligação, juntamente com as cuvetas de transformação estéreis, as células competentes e 1 ml de meio de cultura LB em tubo de microcentrífuga de 1,5 ml foram mantidos em gelo até o momento da transformação. Uma alíquota de 3 μl da reação de ligação foi adicionada a 40 μl de células competentes de *E. coli*, linhagem DH10β, previamente aliquotada em uma cuveta. A eletroporação foi realizada utilizando eletroporador MicroPulse (BioRad) sob as condições de 2,5 kV, 25 μF e 400 ohm. Imediatamente após o pulso elétrico, a solução

contendo as células transformadas foi ressuspensa em 1 ml de meio de cultura LB. O volume total da ressuspensão foi transferido para um novo tubo de microcentrífuga de 1,5 ml estéril e incubado em estufa à 37 ºC, por 1 hora.

Alíquotas de 80 μl de células transformadas foram plaqueadas em meio LB sólido contendo 100 ul de X-gal (20 mg/ml dimetilformamida; Amersham Biosciences), IPTG (200 mg/ml; Amersham Biosciences) e ampicilina (50 mg/l; Amersham Biosciences). As placas foram incubadas em estufa a 37 º C por 12 horas.

4.2.6 Seleção de transformantes e cultura permanente

Para validação do banco genômico apenas 96 clones foram previamente selecionados. Colônias transformantes foram transferidas individualmente para uma placa 96-*well* (Costar) contendo, em cada poço, 110 μl de meio LB-glicerol 60%. As colônias foram incubadas por 12 horas em estufa à 37 ºC e posteriormente congelada a –80 ºC. Algumas colônias foram aleatoriamente selecionadas para serem amplificadas via Reação em Cadeia de Polimerase (PCR, *Polimerase Chain Reaction*) utilizando o *primer* universal M13, com a finalidade de confirmar a presença dos fragmentos clonados dentro da faixa de tamanho esperada.

Após confirmação, outros 10.000 clones transformados foram aleatoriamente selecionados, e incubados por 12 horas em estufa a 37°C. Após o crescimento das colônias, as placas foram armazenadas a –80 °C.

4.3 Extração do Plasmídio (Microprep)

O DNA plasmidial contendo os fragmentos inseridos foi extraído utilizando o procedimento de lise alcalina (Ausubel, 2002). Após secagem, o DNA plasmidial foi solubilizado em 30 μl de água Milli-Q autoclavada e amostras aleatórias de cada placa tiveram sua concentração estimada em aproximadamente 200 ng/μl utilizando gel de agarose 1%.

4.4 Seqüenciamento

As reações de seqüenciamento foram feitas utilizando-se a tecnologia *Big Dye Terminator* versões 2.0 e 3.0 (Applied Biosystems). Inicialmente, aproximadamente 6700 clones foram seqüenciadas utilizando o *primer* M13 *forward*. Destes clones, aproximadamente 2400 foram novamente seqüenciados utilizando o *primer* M13 *reverse*.

Reagentes	Volume (1 reação)
H2O Milli-Q autoclavada	3,2 μl
Tampão 5x	3 μl
DNA (100-500 ng/μl)	2,0 μl
Primer F ou R (2 μM)	0,8 μl
Big Dye v.3.0 (Applied Biosystems)	1,0 μl

Obs.: o tampão de seqüenciamento 5x é composto por Tris-HCl 0,1 M pH 9,4 e $MgCl_2$ 10 mM.

Os seqüenciamentos foram realizados em plataformas de seqüenciamento automático ABI 377, ABI 3100 e ABI 3700 (Applied Biosystems).

4.5 Análise das Seqüências

As seqüências obtidas foram imediatamente depositadas no banco de seqüências do Projeto Genolyptus, localizados no Laboratório de Genômica e Expressão da Universidade Estadual de Campinas, Unicamp, e no Laboratório de Bioinformática, da Universidade Católica de Brasília. Todas as seqüências foram analisadas quanto a sua qualidade utilizando o programa *Phred* (Ewing e Green, 1998; Ewing *et al.*, 1998). Os nucleotídeos correspondentes a seqüências de vetor foram mascarados. As seqüências com pelo menos 150 nucleotídeos, excluindo-se o vetor, com qualidade *Phred* maior ou igual a 20 foram aceitas.

Para as análises posteriores as seqüências foram verificadas utilizando o programa *Lucy* (Chou e Holmes, 2001) e os nucleotídeos com qualidade inferior à descrita acima foram removidos da seqüência. Aproximadamente 7.395 seqüências com mais de 150 nucleotídeos foram aceitas.

Utilizando o programa *CAP3* (Huang e Madan, 1999) e os procedimentos de agrupamento de ESTs utilizando o programa *ESTate* (Slater, 1999) as 7.395 seqüências obtidas foram analisadas e agrupadas. O agrupamento de seqüências utilizando *CAP3* permitiu a obtenção de 766 agrupamentos contíguos e 5.429 singletos, totalizando 6195 seqüências únicas. O agrupamento utilizando o programa *ESTate* gerou uma estimativa da redundância da biblioteca genômica *shotgun*, já que este programa é mais estringente que o *CAP3*.

4.5.1 Seqüências repetitivas

Para quantificação da riqueza e distribuição de seqüências repetitivas ao longo das seqüências, foi utilizado o programa *RepeatMasker* (Smith e Green, 2001), desenvolvido com o intuito de "limpar" as seqüências de elementos repetitivos e seqüências de baixa complexidade, visando reduzir o número de resultados espúrios em comparações de seqüências com banco de dados na procura por genes. Atualmente, o programa envolve uma análise detalhada destes tipos de seqüências, identificando elementos repetitivos e de baixa complexidade.

Outra abordagem também foi utilizada para identificar elementos genéticos transponíveis além daqueles identificados pelo *RepeatMasker*, onde muitos elementos genéticos não estudados até o momento, ou presentes apenas em *Eucalyptus,* não foram identificados. Foi utilizada uma estratégia simples à procura de genes pertencentes a elementos genéticos transponíveis pela comparação com seqüências destes genes e elementos por meio de *Basic Local Alignment Search Tool*), contra seqüências nucleotídicas traduzidas ou proteínas (blastx-nr)(Morgante, M., comunicação pessoal).

Duas estratégias diferentes foram utilizadas para identificar microssatélites. A busca por seqüências contendo microssatélites em DNA genômico foi realizada utilizando os algoritmos *RepeatMasker* e *Tandem Repeat Occurrence Locator* (TROLL) (Castelo *et al*., 2002), sendo que cada análise foi feita com objetivos distintos. A análise com o programa *RepeatMasker* teve como objetivo quantificar a riqueza de microssatélites de uma forma geral no genoma de *Eucalyptus,* onde regiões microssatélites com pelo menos 18 nucleotídeos simples, compostos e imperfeitos,

foram quantificados. A análise utilizando o programa *TROLL* objetivou a identificação de regiões microssatélites completas e passíveis de serem utilizadas para desenvolvimento de marcadores moleculares, já que este programa é mais estringente e identifica apenas microssatélites perfeitos.

O programa *TROLL* é baseado no algoritmo de *Aho Corasick*. Este algoritmo utiliza uma estratégia denominada *dictionary approach*, onde as seqüências repetitivas são conhecidas a priori, isto é, procura-se por regiões microssatélites utilizando-se um dicionário de motivos. O programa *TROLL* utiliza o algoritmo *Aho Corasick* para procurar padrões pré-selecionados em uma seqüência-texto e, acrescentado a isto, também há procura por estas repetições em *tandem*, tornando possível a identificação de regiões microssatélites (Castelo *et al*., 2002)

4.5.1.1 Microssatélites

Na identificação de seqüências simples repetitivas, ou microssatélites, o programa *RepeatMasker* localiza quase todas as seqüências assim denominadas, perfeitas ou não, e que tenham um comprimento igual ou superior a 18 pb. Este programa foi utilizado com o intuito de acessar o conteúdo total de microssatélites presente no genoma de *Eucalyptus*. O algoritmo do programa *RepeatMasker* também procura por microssatélites pelo *dictionary approach*, o que o torna tão rápido nesta função quanto o *TROLL*.

A análise utilizando o programa *TROLL* foi feita com o objetivo de quantificar e separar microssatélites que poderiam ser utilizados em mapeamento genético. O programa *Primer3* (Rozen e Skaletsky, 2000) foi acrescentado ao *pipeline* do software *TROLL* visando o desenvolvimento dos pares de *primers* para os microssatélites

considerados ótimos diretamente em plataforma *web*. Uma outra diferença bastante significativa do programa *TROLL* é que ele não permite identificação de motivos uni- e hexanucleotídicos.

Além de analisar e quantificar a presença de microssatélites nas seqüências genômicas, um conjunto de seqüências do banco de ESTs do Projeto Genolyptus também foi analisado utilizando os mesmos programas, com o intuito de comparar a presença destas estruturas próximas ou dentro de genes. A análise de ESTs com relação à presença de microssatélites permitiu comparar este tipo de seqüência com seqüências genômicas, no que diz respeito ao número de microssatélites passíveis de serem utilizados como marcadores moleculares. Como há milhares de ESTs já seqüenciados no Projeto Genolyptus, a identificação de microssatélites em abundância nestas estruturas proverá o projeto com um número ainda maior de marcadores moleculares hipervariáveis.

A identificação de seqüências microssatélites e possíveis genes em um mesmo fragmento seqüenciado, juntamente com a utilização de microssatélites presentes em ESTs, permitirão o desenvolvimento de marcadores contidos ou intimamente ligados a genes, possibilitando o mapeamento do gene no mapa genético de *Eucalyptus* e a posterior análise de co-localização de genes candidatos e QTLs.

As regiões contendo microssatélites perfeitos foram analisadas também de acordo com o motivo e com o número de repetições. Após a quantificação e qualificação destas regiões, foram selecionadas aquelas passíveis de serem utilizadas em mapeamento genético, baseado em características como tamanho da região microssatélite e qualidade das regiões flanqueadoras. Desta forma, apenas aquelas

regiões microssatélites perfeitas e com regiões flanqueadoreas com qualidade suficiente para desenho de *primers* foram selecionadas para esta finalidade.

Pares de *primers* específicos para a amplificação destas seqüências repetitivas foram desenhados utilizando o *software Primer3* e testados em parentais distintos para a detecção de polimorfismo. O *software Primer3* identifica seqüências de oligonucleotídeos flanqueando as seqüências microssatélites identificadas pelo *TROLL* obedecendo alguns parâmetros, como: tamanho dos oligonucleotídeos variando entre 18 a 23 nucleotídeos, temperatura de anelamento variando entre 56 ºC e 64 ºC, diferença de temperatura de anelamento máxima entre os pares de *primers* menor ou igual a 1 ºC, conteúdo GC variando entre 40 e 60 %, composição da extremidade 3', complementariedade entre os pares de *primers* e presença de *hairpins* e dímeros, entre outras.

Para a otimização das condições de amplificação dos locos foram utilizados 12 indivíduos de espécies diferentes ou híbridos de *Eucalyptus*. A reação em cadeia de polimerase (PCR) para amplificação das regiões microssatélites foi feita em termocicladores GeneAmp System 9700 (Applied Biosystems), com 95 ºC por 5 minutos seguidos por 30 ciclos compostos por 3 fases: 95ºC por 1 minuto, temperatura de anelamento (Tm) por 1 minuto e 72 ºC por 1 minuto, e uma fase final de elongação, após os 30 ciclos, a 72 ºC por 30 minutos.

Reagentes	Volume (1 reação)
dd H_2O:q.s.	5,00 μl
Tampão PCR (10x ;Phoneutria)	1,25 μl
dNTP(2.5 mM;Amersham Biosciences)	1,00 μl
$MgCl_2$ (50 mM; Gibco)	0,39 μl
Primer (10 mM; Operon Technologies)	1,25 μl
Taq DNA Polimerase (5 u/μl; Phoneutria)	0,125 μl
DNA (2 ng/μl)	2,00 μl

Os produtos de PCR foram inicialmente analisados em gel de agarose 1% corados com brometo de etídeo, visualizados sobre luz ultravioleta e documentados em equipamento *EagleEye* (Stratagene). Com a confirmação do funcionamento correto da reação, estes produtos de PCR foram posteriormente analisados em gel de agarose 3%, corados com brometo de etídeo. Alguns locos foram testados também em gel de poliacrilamida corado com nitrato de prata, de acordo com protocolo descrito por Creste *et al.* (2001). O tamanho dos alelos foi estimado por comparação com DNA padrão de 50 pb (Gibco), e a presença de polimorfismo para cada par de *primer* foi anotada. O nível de polimorfismo foi avaliado atribuindo notas a cada loco microssatélite amplificado, sendo que a nota mínima e máxima possíveis foram de 1 e 24, respectivamente, de acordo com o número de alelos identificáveis e diferentes.

Após a distribuição das notas para cada loco, foi feita uma busca para encontrar correlações entre características estruturais das seqüências microssatélites, como tamanho total da seqüência microssatélite, tamanho do motivo, conteúdo GC das regiões flanqueadoras e número de repetições de motivos microssatélites, com o nível

de polimorfismo do loco. Além destas, também foi procurada correlação entre o conteúdo nucleotídico do motivo com o nível de polimorfismo. Para isto, foi atribuída uma nota (*score*) a cada loco, de acordo com seus conteúdos GC e AT.

4.5.1.2 Elementos Genéticos Repetitivos

As seqüências similares a transposons e retrotransposons foram identificadas utilizando dois protocolos diferentes. A partir do programa *RepeatMasker*, foram analisados todos os singletos e agrupamentos contíguos em busca de elementos genéticos repetitivos catalogados no banco de dados *Repbase*, o qual contém todas seqüências deste tipo, descritas em todos os organismos estudados.

Como a identificação destes elementos pelo programa *RepeatMasker* é limitada apenas a elementos já identificados em outros organismos, muitos elementos genéticos não identificados ou presentes apenas em *Eucalyptus* não seriam passíveis de identificação. Para resolver este problema, foram identificadas seqüências genômicas de *Eucalyptus* similares a genes pertencentes a elementos genéticos transponíveis, como Capsídeo (CP), Trancritase reversa (RT), Integrase (INT) e outros, depositados no NCBI (Morgante, M., com. pess.).

4.5.2 Genes, Fragmentos de Genes e Pseudogenes

4.5.2.1 Comparação com ESTs

Possíveis genes contidos nas seqüências da biblioteca genômica desenvolvida neste trabalho foram, primeiramente, identificados a partir da comparação dos fragmentos de DNA genômico seqüenciados com bancos de ESTs de *Eucalyptus*,

desenvolvidos no âmbito do Projeto Genolyptus. Este tipo de comparação fornece melhores evidências na determinação de regiões gênicas, pelo menos no que diz respeito a sua estrutura.

Inicialmente, os 766 agrupamentos contíguos formados no agrupamento (*clustering*) das seqüências com *CAP3* (Huang e Madan, 1999) foram comparados com mais de 11.000 ESTs obtidos a partir de duas bibliotecas de cDNA, uma de xilema de *E. urophylla* e outra de folha madura de *E. grandis*, utilizando o programa *GenSeqer* (Schlueter *et al.*, 2003). Aproximadamente 44 seqüências apresentaram similaridades significativas e foram consideradas como codificantes. Nestas seqüências, foi possível identificar 2 éxons e 23 íntrons. Pela comparação dos ESTs correspondentes às seqüências genômicas identificadas como codificantes com o banco de dados não-redundante do NCBI utilizando o protocolo *blastx*, foi possível identificar as funções putativas destes genes.

4.5.2.2 Análise utilizando o *Gene Projects*

Uma segunda abordagem na identificação de possíveis genes foi feita a partir da utilização do *pipeline* denominado *Gene Projects*, desenvolvido no Laboratório de Genômica e Expressão, do Departamento de Genética e Evolução da Unicamp. Todas as seqüências brutas, sem serem agrupadas, foram comparadas com o banco de dados não-redundante de seqüências de proteínas do NCBI. As seqüências com valor *e* igual ou superior a 10^{-20} e escore superior a 100 foram consideradas como sendo significativas para a análise de presença de regiões codificantes dentro do banco de DNA genômico.

4.5.2.3 Identificação de genes para RNAs transportadores

Possíveis genes codificantes de tRNAs foram identificados pelo programa *tRNAscan-SE* (Lowe e Eddy, 1997). Neste programa, todas as seqüências são primeiramente analisadas por dois algoritmos distintos. As seqüências onde genes relativos a tRNAs são identificados passam por um algoritmo mais estringente baseado em um modelo de covariância. Como genes para tRNAs são relativamente pequenos, a busca por este tipo de estrutura foi feita tanto nas seqüências genômicas agrupadas como nos singletos.

4.5.2.4 Identificação de regiões promotoras

Para identificação de padrões de seqüências possivelmente pertencentes a regiões promotoras, todas as seqüências que apresentaram similaridade a genes completos e a seqüências iniciais de fragmentos de genes foram alinhadas utilizando os programa *BioEdit* versão 5.0.9 (Hall, 2001) e *Clustal X* (Thompson *et al.*, 1997), e seqüências conservadas na região a montante aos genes identificados pelo *pipeline Gene Projects* foram identificadas.

4.5.3 Composição nucleotídica

4.5.3.1 Conteúdo GC

Por meio do programa *GeeCee* (Bruskiewich, 1999), do pacote *European Molecular Biology Open Software Suite* (Emboss), e do programa *RepeatMasker,* foi calculado o conteúdo médio de nucleotídeos C e G tanto nas seqüências genômicas, como em ESTs. Adicionalmente, foi calculado o conteúdo para íntrons e éxons, com o

intuito de descrever possíveis diferenças no conteúdo GC entre seqüências codificantes e não-codificantes.

O programa *GeeCee* foi utilijzado para quantificar o conteúdo GC nas regiões flanqueadoras de microssatélites, visando avaliar o nível de correlação entre o conteúdo de GC e o nível de polimorfismo do loco SSR.

5 RESULTADOS

5.1 Seqüências Repetitivas

Por meio das análises dos fragmentos de DNA genômico de *Eucalyptus* seqüenciados aleatoriamente, foi possível obter uma visão geral e informativa sobre o genoma deste gênero. Diversos microssatélites e elementos genéticos transponíveis foram encontrados, além de terem sido identificadas regiões de baixa complexidade bastante numerosas. No total, cerca de 176.000 pares de bases foram localizados em seqüências repetitivas, o que equivale a 5,8% do total seqüenciado.

Foram identificadas 2.933 seqüências repetitivas dentre as 7.395 seqüências de DNA genômico de *Eucalyptus* analisadas, dentre as quais 1.636 são seqüências de baixa complexidade, 986 são regiões microssatélites (807, quando excluímos os uninucleotídeos) e 310 são parte de elementos genéticos transponíveis e retrotransponíveis (Tabela 1).

Tabela 1: Resumo dos resultados obtidos na análise do genoma de *Eucalyptus* através de "sample sequencing".

Elementos Genéticos Repetitivos		
Transposons	310	
Microssatélites	986	
Baixa complexidade	1636	
Total		2933
Genes		
Comparação com ESTs	44	
Gene Projects	166	
Genes tRNA	16	
Total		226
Seqüências pouco informativas		4547
Total de fragmentos seqüenciados		**7706**

Dentre as seqüências de baixa complexidade, foi observado que regiões ricas em AT são muito mais numerosas do que qualquer outro tipo de seqüência repetitiva, correspondendo a mais de 30% do total de 3922 seqüências, seguidas por seqüências ricas em A e ricas em T. Do total de nucleotídeos seqüenciados, cerca de 74532 (1,65%) encontram-se em regiões de baixa complexidade. Quando foi analisada a presença do elemento de baixa complexidade em termos de número absoluto de seqüências, observou-se um total de 1628 fragmentos contendo um ou mais destes elementos em 7395 fragmentos, o que equivale a aproximadamente 1 seqüência de baixa complexidade a cada 5 fragmentos de DNA genômico seqüenciados, um número bastante elevado que demonstra que este tipo de estrutura está ubiqüamente presente no genoma de *Eucalyptus* (Tabela 2).

Tabela 2: número total de seqüências de baixa complexidade identificadas na biblioteca de DNA genômico de E*ucalyptus*.

Seqüência	nº ocorrências
Ricas em AT	1195
Ricas em T	125
Ricas em A	108
Ricas em GC	83
Ricas em CT	50
Ricas em GA	29
Ricas em C	22
Ricas em G	16
Polipurínicas	4
Polipirimidínicas	4
TOTAL	1636

5.1.1 Microssatélites

5.1.1.1 Identificação de regiões microssatélites em seqüências genômicas

As duas análises distintas feitas sobre os fragmentos de DNA genômico em busca de regiões microssatélites revelaram que há um elevado número destas no genoma de *Eucalyptus*, permitindo identificar tipos e motivos mais numerosos. Além disto, implicações práticas imediatas foram tomadas com base neste resultado, como o desenvolvimento, caracterização e aplicação de pares de *primers* para amplificação destas regiões.

5.1.1.1.1 Análise de microssatélites ou Simle Sequence Repeats (SSRs) em seqüências genômicas pelo programa RepeatMasker

A análise da riqueza e quantidade de microssatélites em *Eucalyptus* utilizando o programa *RepeatMasker* identificou 986 regiões microssatélites dentre 7.395 fragmentos de DNA genômico analisados, somando 38.653 pares de bases, ou 1,3 % de todas as bases seqüenciadas. Em média, foi encontrada uma região microssatélite em cada oito fragmentos de DNA genômico seqüenciados. Na análise com o programa *RepeatMasker,* para fins de caracterização da estrutura do genoma, microssatélites de uninucleotídeos seriam computados na classe de seqüências simples repetitivas. Com relação aos objetivos práticos de utilização visando o desenvolvimento de marcadores, na análise com o programa *TROLL*, foram excluídas desta análise as seqüências contendo repetições em *tandem* de um único nucleotídeo.

Microssatélites compostos por motivos uninucleotídicos compõem 18,2 % (Figura 1) dentre todos os tipos de motivos identificados, sendo que o motivo mais numeroso foi (A)n (Tabela 3). Dinucleotídicos compõem cerca de 39 % de todas regiões microssatélites identificadas e os motivos mais freqüentes foram (TC)n e (TA)n (Tabela 4), representando aproximadamente 37 % de todos microssatélites identificados (Figura 2), e 94 % dentre os dinucleotídicos. Dentre os microssatélites trinucleotídicos, foi encontrada uma melhor distribuição entre os diversos motivos (10 diferentes motivos), sendo que os motivos mais freqüentes (TAA)n, (CCG)n e (GAA)n, correspondendo a 13 % de todos os microssatélites identificados e a 78 % de todos os trinucleotídicos (Tabela 5).

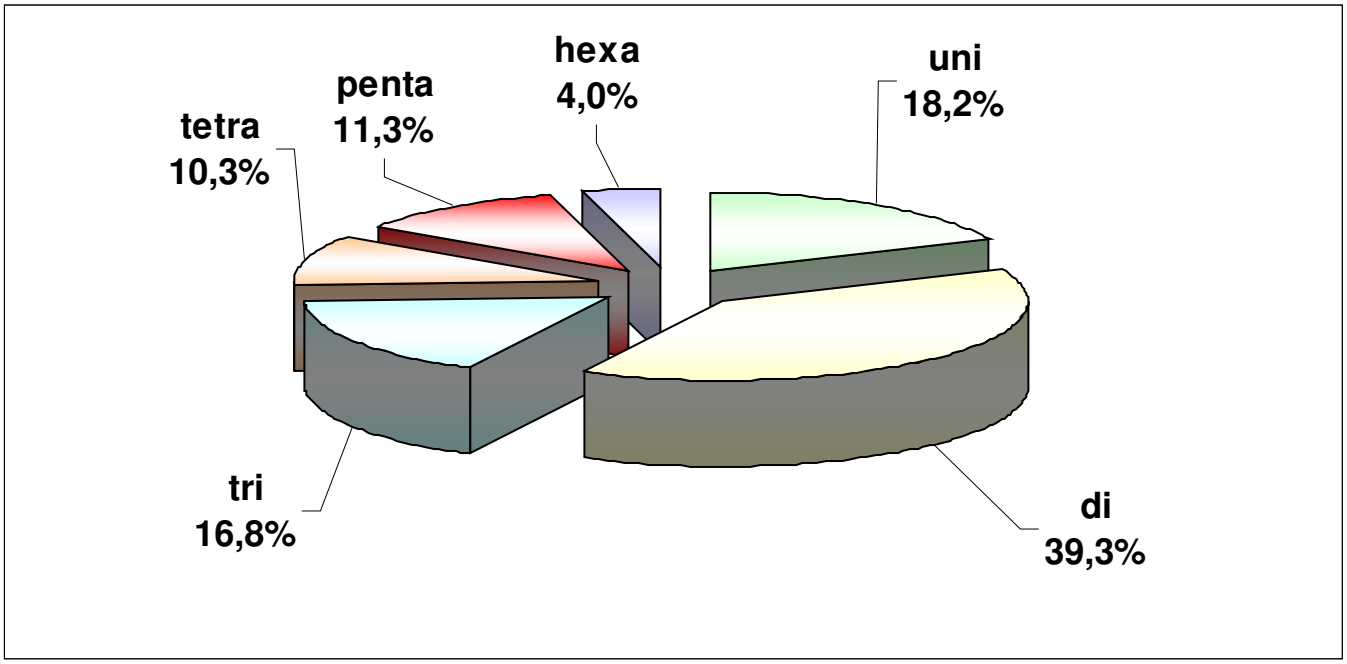

Figura 1: Proporções de microssatélites identificados pelo programa *RepeatMasker*, em DNA genômico, de acordo com o tamanho do motivo.

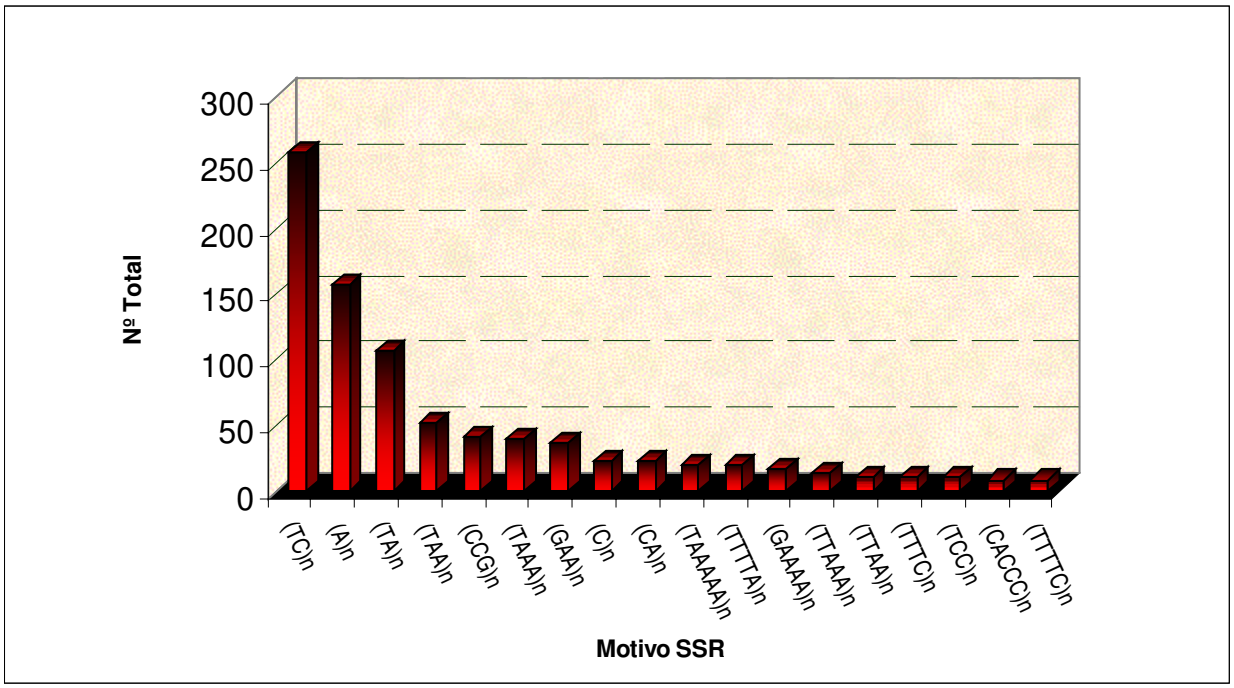

Figura 2: Motivos microssatélites mais freqüentes identificados na bilbioteca de DNA genômico de *Eucalyptus*.

Os tetranucleotídeos mostraram uma distribuição bem maior de regiões microssatélites dentre as diversas variedades de motivos (16 motivos diferentes) (Tabela 6). Os motivos mais freqüentes foram (TAAA)n, (TTTC)n e (TTAA)n, os quais corresponderam a 6,3 % de todos os microssatélites identificados e 62 % dentre os tetranucleotídicos. Os microssatélites pentanucleotídicos estavam distribuídos em 25 motivos diferentes (Tabela 7), com uma concentração da maior parte dos microssatélites em apenas alguns motivos, sendo que (GAAAA)n, (TTTTA)n e (TTAAA)n correspondem a 6,2 % de todas regiões microssatélites identificadas e a 55 % de todos os pentanucleotídeos. Foram encontrados 40 microssatélites hexanucleotídicos (Tabela 8), distribuídos em uma gama de motivos bastante reduzida, dentre os quais destacou-se o motivo (TAAAAA)n, correspondendo a 2,0 % dos microssatélites identificados e a 50 % de todos hexanucleotídeos.

Tabela 3: Total de microssatélites uninucleotídicos identificados em seqüências genômicas de *Eucalyptus.*

(A)n	157
(C)n	23
TOTAL	180

Tabela 4: Total de microssatélites dinucleotídicos identificados em seqüências genômicas de *Eucalyptus.*

(TC)n	142
(GA)n	116
(TA)n	106
(CA)n	13
(TG)n	10
(CG)n	1
TOTAL	388

Tabela 5: Total de microssatélites trinucleotídicos identificados em seqüências genômicas de *Eucalyptus.*

(TAA)n	36
(CCG)n	27
(GAA)n	20
(TTC)n	17
(TTA)n	16
(CGG)n	14
(TCC)n	10
(ATG)n	6
(TCG)n	5
(CAA)n	3
(GGA)n	3
(TGG)n	3
(CAG)n	2
(TTG)n	2
(CCA)n	1
(CGA)n	1
TOTAL	166

Tabela 6: Total de microssatélites tetranucleotídicos identificados em seqüências genômicas de *Eucalyptus.*

(TAAA)n	22
(TTTA)n	18
(TTAA)n	11
(TTTC)n	11
(TATG)n	5
(CAAT)n	4
(CCAA)n	4
(TCCC)n	4
(ATTG)n	3
(CAAA)n	2
(CAGA)n	2
(CATA)n	2
(CTAA)n	2
(GAAA)n	2
(GGAA)n	2
(TCTA)n	2
(CATG)n	1
(CCCA)n	1
(CGAG)n	1
(TCCA)n	1
(TTAG)n	1
(TTCC)n	1
TOTAL	102

Tabela 7: Total de microssatélites pentanucleotídicos identificados em seqüências genômicas de *Eucalyptus.*

(GAAAA)n	16
(TTAAA)n	14
(TTTTA)n	12
(CACCC)n	8
(TTTTC)n	8
(TAAAA)n	7
(TTTAA)n	4
(ATTAG)n	3
(CACAC)n	3
(CTAAT)n	3
(CAAAA)n	2
(CACAG)n	2
(CCGAG)n	2
(CCGGG)n	2
(CTAAA)n	2
(CTTTA)n	2
(GGATG)n	2
(TCGGG)n	2
(TTTTG)n	2
(AACTG)n	1
(ATATG)n	1
(ATTTG)n	1
(CAACC)n	1
(CACAA)n	1
(CACGA)n	1
(CCCCG)n	1
(CCCGG)n	1
(CCTAA)n	1
(CTTAA)n	1
(TATAA)n	1
(TCGTG)n	1
(TTCTC)n	1
(TTTAG)n	1
TOTAL	110

Tabela 8: Total de microssatélites hexanucleotídicos identificados em seqüências genômicas de *Eucalyptus.*	
(TAAAAA)n	13
(TTTTTA)n	7
(CAGAGA)n	4
(TCTCCC)n	4
(CAAAAA)n	3
(CCCCCA)n	3
(TTTTTG)n	2
(ATGGTG)n	1
(CCCGAA)n	1
(CGGGGG)n	1
(TATATG)n	1
TOTAL	40

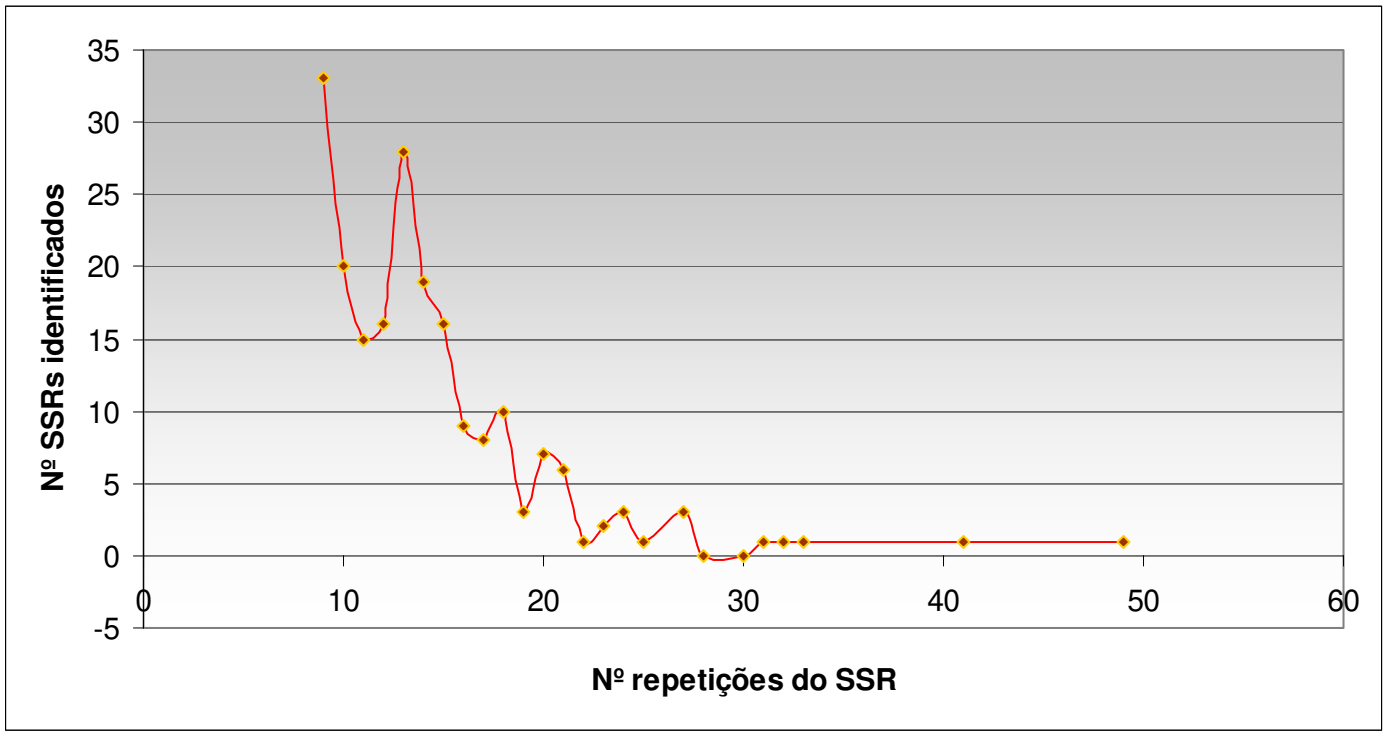

Figura 3: Relação entre o número de seqüências microssatélites identificadas e o número de motivos repetidos em *tandem* em *Eucalyptus*.

Em geral, todos os tamanhos de motivos apresentaram um mesmo padrão com relação ao número total de regiões microssatélites identificadas e ao número de

repetições, mostrando que locos microssatélites com maior número de repetições são mais raros (Figura 3).

Na Tabela 9 podemos observar o número total de regiões microssatélites identificadas pelo programa *RepeatMasker* dentre as 7.395 seqüências genômicas de *Eucalyptus* analisadas, além de conter os tamanhos médios dos fragmentos, de acordo com o tamanho do motivo microssatélite.

Tabela 9: Número total de microssatélites identificados em *Eucalyptus* pelo programa RepeatMasker por tamanho de motivo, com o respectivo tamanho médio das regiões microssatélites.

	n	tamanho médio
uninucleotídeos	180	24,8
dinucleotídeos	388	33,8
trinucleotídeos	167	40,8
tetranucleotídeos	102	36,9
pentanucleotídeos	110	48,2
hexanucleotídeos	40	31,5
Total/Média geral	987	35,2

5.1.1.1.2 Análise de SSRs em seqüências genômicas pelo programa TROLL

Por meio da outra estratégia de identificação de microssatélites, utilizando o programa *TROLL*, um menor número destes elementos foi encontrado, devido aos parâmetros mais conservadores utilizados. Além disto, não consideramos nesta análise os microssatélites uninucleotídicos, já que estes não têm aplicação prática como marcadores moleculares, nem hexanucleotídicos, devido a restrições do programa. Deve-se considerar esta análise como uma subamostragem da análise utilizando o programa *RepeatMasker*, desconsiderando uni- e hexanucleotídeos. Nesta análise, foram encontrados 319 microssatélites perfeitos e completos.

Microssatélites com motivos dinucleotídicos compunham mais da metade daqueles identificados (64,1 %), seguidos por trinucleotídicos (24,1 %),

tetranucleotídicos (8,4 %) e pentanucleotídicos (3,1 %) (Figura 4). Na Tabela 10 estão detalhados o número de regiões microssatélites, o tamanho médio das regiões e o número médio de repetições, de acordo com o tamanho do motivo microssatélite.

Tabela 10: Número de microssatélites di-, tri-, tetra- e pentanucleotídicos identificados em *Eucalyptus* pelo programa TROLL, com tamanho médio e o número médio de repetições.

	n	tamanho médio	média de repetições
dinucleotídeos	205	27,2	13,6
trinucleotídeos	77	23,7	7,9
tetranucleotídeos	27	26,0	6,5
pentanucleotídeos	10	23,0	4,6
Total/ Médias	319	26,1	11,3

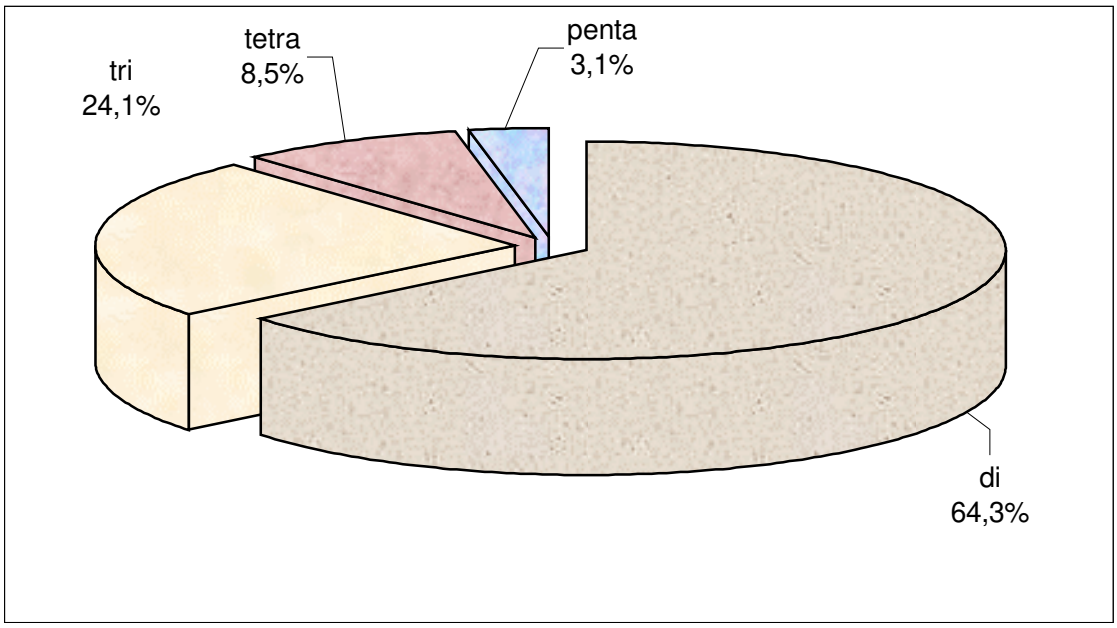

Figura 4: Proporções de microssatélites identificados nas seqüências de DNA genômico de *Eucalyptus* utilizando o programa *TROLL*.

Dentre os microssatélites dinucleotídicos, notou-se que aqueles mais numerosos nesta análise com o probrama *TROLL* foram também os mais numerosos na análise com *RepeatMasker*, (TC)n e (TA)n. Em trinucleotídeos a distribuição de freqüências foi mais tênue que na análise com *RepeatMasker*, sendo que os motivos mais freqüentes foram (TAA)n, (CCG)n e (GAA)n. Os microssatélites tetranucleotídicos também

apresentaram uma distribuição de freqüência mais tênue que aquela observada na análise com *RepeatMasker*, com os motivos mais freqüentes (TAAA)n, (CTCC)n e (GAAA)n. Poucas classes de pentanucleotídeos foram identificadas, sem nenhum destaque para qualquer motivo.

Com relação ao número médio de repetições, percebemos que microssatélites com motivos menores possuem maior número de repetições (Figura 5). Entretanto, quando foi analisado o tamanho médio do fragmento, a diferença entre os tipos de motivos é bem menos pronunciada. Uma ligeira tendência à redução no tamanho dos fragmentos é notada com o aumento do tamanho do motivo, sendo que motivos pares (di- e tetranucleotídeos) têm, em média, fragmentos maiores que motivos ímpares (tri- e pentanucleotídeos) entre os microssatélites identificados pelo programa *RepeatMasker* (teste t, bicaudal; $p=7{,}26 \times 10^{-7}$). Entretanto, entre os microssatélites analisados pelo programa *TROLL*, esta diferença não foi significativa (teste t, bicaudal; p=0,05). É possível que esta seja resultado de uma diferença na proporção de microssatélites imperfeitos, com tri- e pentanucleotídeos possuindo um maior número destes na análise pelo *RepeatMasker*, já que na análise com o programa *TROLL* este padrão não foi significativo.

Dentre as 319 regiões microssatélites identificadas foram selecionadas 156, procurando escolher regiões microssatélites de acordo com a proporção de cada tipo encontrado, sendo que 102 foram dinucleotídeos, 31 trinucleotídeos, 19 tetranucleotídeos e 4 pentanucleotídeos.

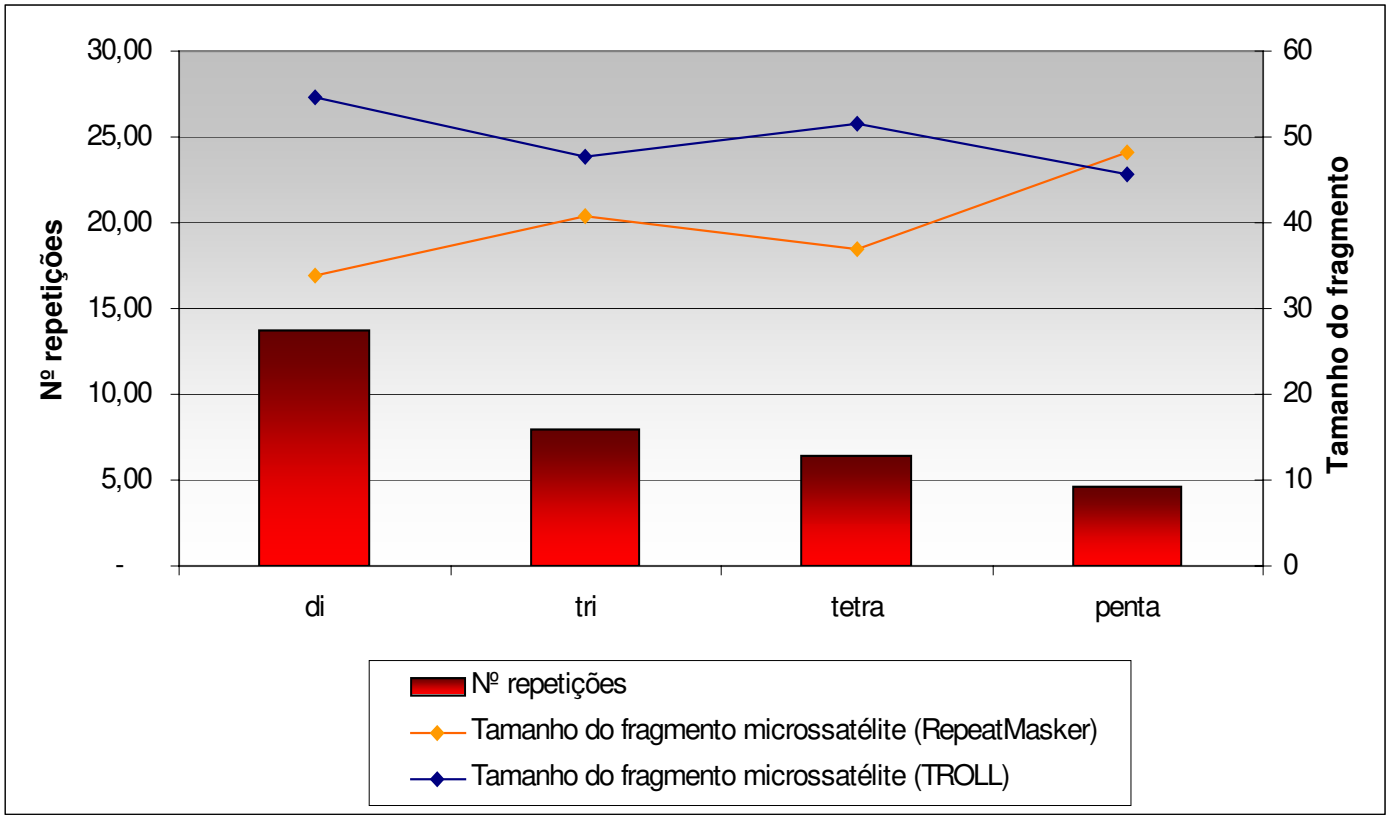

Figura 5: Relação entre o tamanho do motivo microssatélite, número de repetições médio e tamanho total médio da seqüência microssatélite. As barras vermelhas mostram o número médio de repetições em cada tipo de microssatélite, segundo o programa *Troll*.

5.1.1.2 Identificação de regiões microssatélites em ESTs de *Eucalyptus*

5.1.1.2.1 Análise de SSRs em ESTs pelo programa RepeatMasker

A análise em paralelo de ESTs de *Eucalyptus* mostrou resultados interessantes. Dentre as cerca de 4.451 seqüências analisadas com o programa *RepeatMasker*, em apenas 1 biblioteca (ESTs de folhas maduras de *E. grandis*), foram identificadas 1.529 regiões microssatélites, equivalentes a 81.683 pb seqüenciadas de um total de 3.520.884, o que corresponde a 2,3 % de todas as bases seqüenciadas analisadas. Em média, foi encontrada uma região microssatélite a cada três ESTs, uma freqüência significativamente maior (valor p = 3,36 x 10^{-12} por teste exato de Fisher bilateral) do

que no banco de seqüências de DNA genômico, distribuídos numa ampla gama de motivos (Figura 6).

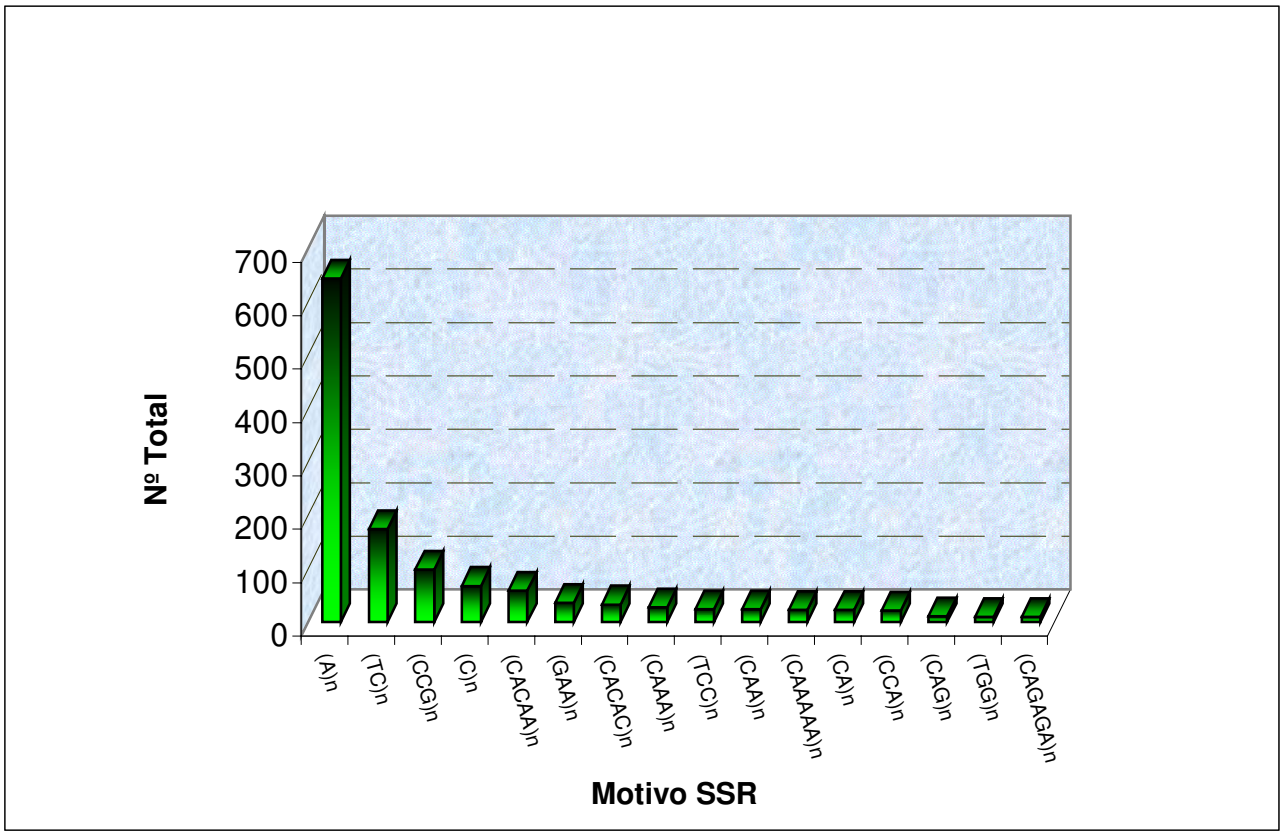

Figura 6: Motivos microssatélites mais freqüentes identificados em ESTs de *Eucalyptus* pelo programa *RepeatMasker*.

Com relação à proporção de microssatélites de acordo com o tamanho do motivo, notou-se uma considerável diferença com relação a microssatélites de regiões genômicas, com microssatélites compostos por motivos uninucleotídeos compondo quase metade de todos microssatélites identificados (49,2%), e microssatélites compostos por motivos (A)n sendo os mais numerosos (Tabela 11). Microssatélites trinucleotídicos ocorreram em maior número (17,4%) que dinucleotídicos (13,2%) (Figura 7).

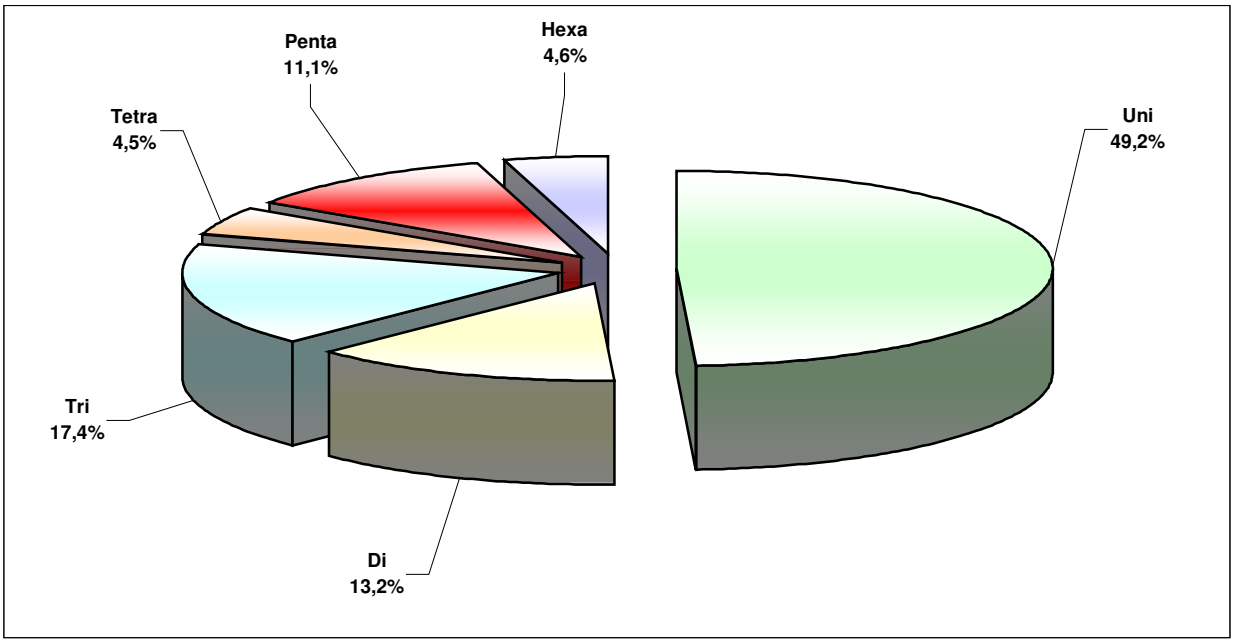

Figura 7: Proporções de microssatélites identificados pelo programa *RepeatMasker*, em ESTs, de acordo com o tamanho do motivo.

Dentre os 9 motivos trinucleotídicos diferentes identificados, as classes mais comuns foram (CCG)n, (GAA)n e (TCC)n (Tabela 12), e dentre os 4 motivos dinucleotídicos, destacou-se o motivo (TC)n, correspondendo a 86 % de todos dinucleotídicos (Tabela 13).

Foram identificados tetranucleotídicos distribuídos em 13 classes diferentes, sendo a mais numerosa (CAAA)n (Tabela 14), correspondendo a 40% de todos tetranucleotídicos encontrados.

Regiões microssatélites contendo repetições pentanucleotídicas compunham 11,1% de todos microssatélites identificados pelo programa *RepeatMasker* em ESTs, distribuídas em 32 diferentes motivos, com destaque para os motivos (CACAA)n e

(CACAC)n que, juntos, corresponderam a 56,2 % de todos os pentanucleotídicos identificados em ESTs (Tabela 15).

Microssatélites hexanucleotídicos dividiram-se em 15 classes de motivos diferentes, e corresponderam a 4,6% de todos microssatélites identificados em ESTs de *Eucalyptus* (Tabela 16), com destaque para o motivo (CAAAAA)n.

Tabela 11: Total de microssatélites uninucleotídicos identificados em ESTs de *Eucalyptus.*

(A)n	643
(C)n	110
TOTAL	753

Tabela 12: Total de microssatélites dinucleotídicos identificados em ESTs de *Eucalyptus.*

(TC)n	174
(CA)n	23
(TA)n	4
(TG)n	1
TOTAL	202

Tabela 13: Total de microssatélites trinucleotídicos identificados em ESTs de *Eucalyptus.*

(CCG)n	97
(GAA)n	43
(CCA)n	32
(TCC)n	31
(CAA)n	25
(CAG)n	15
(TCG)n	13
(CAT)n	8
(TAA)n	2
TOTAL	266

Tabela 14: Total de microssatélites tetranucleotídicos identificados em ESTs de *Eucalyptus.*

(CAAA)n	28
(CAGA)n	6
(CCCA)n	6
(CTCG)n	4
(GAAA)n	8
(CCTG)n	3
(GGGA)n	5
(CCTA)n	2
(CGAT)n	2
(TAAA)n	2
(CAAG)n	1
(CCCG)n	1
(TTCG)n	1
TOTAL	69

Tabela 15: Total de microssatélites pentanucleotídicos identificados em ESTs de *Eucalyptus.*

(CACAA)n	61
(CACAC)n	34
(CAAAA)n	8
(CACCC)n	10
(CAAGA)n	7
(CAGAA)n	5
(GAAAA)n	8
(TTATA)n	5
(CGAGA)n	3
(CAGAG)n	2
(CCGAG)n	2
(CCGCT)n	2
(GGAGA)n	2
(TCCGG)n	2
(ATCTG)n	1
(CAAAC)n	1
(CAACT)n	1
(CAATA)n	1
(CACTG)n	1
(CATAA)n	1
(CATGC)n	1
(CCCCG)n	1
(CCCGA)n	1
(CGAAT)n	1
(CGCGG)n	1
(CGGAA)n	1
(CGGGG)n	1
(CTGTG)n	1
(GATTG)n	1
(TTCGG)n	1
(TTTCG)n	1
(TTTTA)n	1
TOTAL	169

Tabela 16: Total de microssatélites hexanucleotídicos identificados em ESTs de *Eucalyptus.*

(CAAAAA)n	23
(CAGAGA)n	10
(CCCCCA)n	8
(CCCGAA)n	7
(TCTCCC)n	7
(AGGGGG)n	2
(CACCAT)n	2
(GGAGAA)n	2
(GGGAGA)n	2
(TTCTCC)n	2
(CCCCAG)n	1
(CCCCCG)n	1
(TAAAAA)n	1
(TCTCTG)n	1
(TTCGGG)n	1
TOTAL	70

Na Tabela 17, os microssatélites identificados pelo programa *RepeatMasker* em ESTs de *Eucalyptus* estão discriminados de acordo com o número encontrado e tamanho médio da região microssatélite, de acordo com o tamanho do motivo.

Tabela 17: número total de microssatélites identificados em ESTs de Eucalyptus pelo programa RepeatMasker por tamanho de motivo, com o respectivo tamanho médio das regiões microssatélites.

	n	tamanho médio
uninucleotídeos	753	47,2
dinucleotídeos	202	38,4
trinucleotídeos	266	49,6
tetranucleotídeos	69	75,5
pentanucleotídeos	169	19,0
hexanucleotídeos	70	60,6
Total/Média geral	1529	45,2

5.1.1.2.2 Análise de SSRs em ESTs pelo programa TROLL

Na análise dos mesmos ESTs utilizando o programa *TROLL* e, portanto, sendo mais conservador, houve uma redução drástica nos números de locos microssatélites identificados, em comparação com a análise com o *RepeatMasker*, principalmente dentre os tetra- e pentanucleotídicos.

Foram encontradas 481 regiões microssatélites, dentre as quais 41,9 % eram dinucleotídicas, 56,7 % trinucleotídicas, 1,5 % tetranucleotídicas e 0,2 % pentanucleotídica (Figura 8).

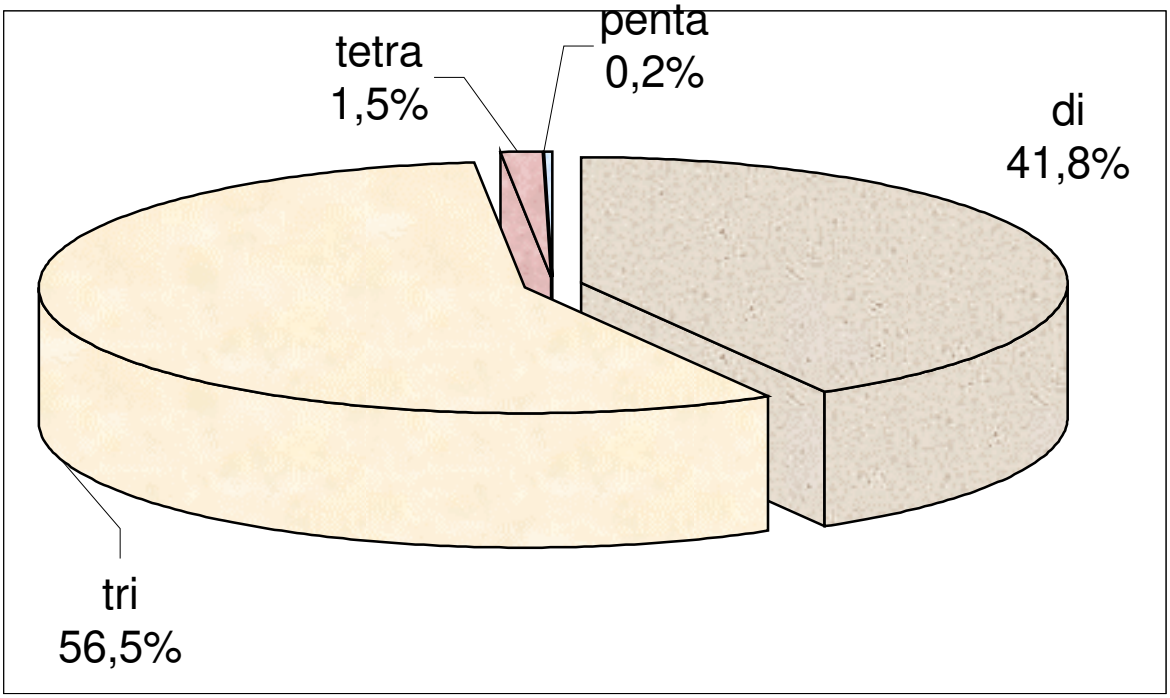

Figura 8: Proporções de motivos microssatélites identificados pelo programa *TROLL* em ESTs de *Eucalyptus*

Dentre os dinucleotídeos, as classes mais comuns foram (TC)n e (AG)n, e dentre os trinucleotídicos, mais uma vez o tamanho de motivo mais freqüente (correspondendo a 56,5 % de todos microssatélites identificados em ESTs pelo programa *TROLL*), foi (GAA)n, correspondendo a 40,1 % de todos os trinucleotídeos observados. Poucos tetra- e pentanucleotídeos foram encontrados, sem destaque para qualquer motivo em particular. O número total de microssatélites em ESTs presentes em *Eucalyptus* identificados pelo programa *TROLL*, bem como o tamanho médio das regiões microssatélites e o número médio de repetições, de acordo com o tipo de motivo, estão discriminados na Tabela 18.

Tabela 18: número de microssatélites di-, tri-, tetra- e pentanucleotídeos identificados em ESTs de *Eucalyptus* pelo programa TROLL, com tamanho médio e o número médio de repetições.

	n	tamanho médio	média de repetições
dinucleotídeos	201	23,0	11,5
trinucleotídeos	272	21,0	7
tetranucleotídeos	7	18,9	7,8
pentanucleotídeos	1	20,0	4
Total/ Médias	481	21,8	8,9

Dentre os 481 microssatélites encontrados em ESTs foram selecionadas 22 seqüências microssatélites para as quais foram desenhados pares de *primers* flanqueadores para amplificação via PCR. Destas 22 regiões, 11 eram compostas por dinucleotídicos, 8 eram trinucleotídicas e três eram tetranucleotídicas.

5.1.1.3 Caracterização e otimização de amplificação de marcadores microssatélites

Todos os locos microssatélites identificados (156 definidos a partir de DNA genômico e 22 a partir de ESTs) foram testados a partir de uma temperatura de anelamento inicial de 54 ºC. Aqueles que não apresentaram produtos amplificados de PCR ou que apresentavam bandas não-específicas foram testados com uma temperatura de anelamento 2 ºC superior, e assim por diante, até que todos tivessem sido testados e otimizados até uma temperatura máxima de anelamento de 64 ºC.

De 178 pares de *primers* para regiões microssatélites desenvolvidos e testados, foi obtido produto de amplificação específico para 111, dos quais 12 foram provenientes de ESTs e 99 de DNA genômico. O sucesso de amplificação foi respectivamente de 63,5 % e 54,5 %, para microssatélites desenvolvidos a

partir de clones genômicos e clones de cDNA (Tabelas 19 e 20). Não houve diferença significativa neste sucesso avaliada por um teste exato de Fisher bilateral (p = 0,48).

Tabela 19: Sucesso de amplificação e detecção de polimorfismo dentre os pares de *primers* desenhados a partir de seqüências genômicas.

	Desenhados	Amplificados	% amplificados	Polimórficos	% polimórficos*	% polimórficos**
di	102	67	66,00	32	48,00	31,00
tri	31	18	58,00	11	61,00	35,00
tetra	19	13	68,00	2	15,00	11,00
penta	4	1	25,00	0	0,00	0,00
	156	**99**	**63,00**	**45**	**45,00**	**29,00**

***** pares de *primers* polimórficos em relação ao total de *primers* amplificados; ** pares de *primers* polimórficos em relação ao número total de pares de *primers* desenhados.

Tabela 20: Sucesso de amplificação e detecção de polimorfismo dentre os pares de *primers* desenhados a partir de ESTs.

	Desenhados	Amplificados	% amplificados	Polimórficos	% polimórficos*	% polimórficos**
di	11	6	55,00	5	83,00	45,00
tri	8	4	50,00	2	50,00	25,00
tetra	3	2	67,00	1	50,00	33,00
penta	0	0	0,00	0	0,00	0,00
	22	**12**	**55,00**	**8**	**67,00**	**36,00**

***** pares de *primers* polimórficos em relação ao total de *primers* amplificados; ** pares de *primers* polimórficos em relação ao número total de pares de *primers* desenhados

Dos 12 pares de *primers* provenientes de ESTs que amplificaram, 8 (66,6 %) mostraram polimorfismo entre os 12 parentais testados. Dentre os 99 pares de *primers* provenientes de DNA genômico, 45 mostraram polimorfismo (45,5 %). Desta forma, obtivemos um sucesso final de 8 pares de *primers* polimórficos dentre 22 desenhados a partir de ESTs (36,4 % de aproveitamento), enquanto que entre pares de *primers* desenhados a partir de seqüências de DNA genômico,

esse sucesso caiu para 45 pares de *primers* polimórficos dentre 156 desenhados (28,8 %). Embora exista uma diferença em sucesso na detecção de polimorfismos em valores absolutos, esta diferença foi declarada não significativa (Tabela 21).

Dentre os microssatélites encontrados em regiões genômicas, observou-se que, enquanto os pares de *primers* desenhados flanqueando motivos dinucleotídicos e tetranucleotídicos mostraram maior eficiência de amplificação, pares de *primers* desenhados flanqueando motivos trinucleotídicos mostraram maior nível de polimorfismo. Dentre os microssatélites encontrados em ESTs, foram observadas diferenças quando foi analisado o polimorfismo dos locos, sendo que uma maior proporção de microssatélites compostos por motivos dinucleotídicos apresentou polimorfismo em relação aos outros tamanhos de microssatélites. Na avaliação estatística destas potenciais diferenças, não foram encontradas diferenças significativas na proporção de locos polimórficos derivados de clone genômico ou EST (Tabela 21).

Tabela 21. Resultados dos testes exatos de Fisher para a comparação de proporção de locos microssatélites polimórficos derivados de seqüências genômicas e de EST.

Tipo de seqüência	**Comparação**	**Valor p**
Genômica	Dinucleotídicos x tri/tetra/penta	0,52
Genômica	Dinucleotídicos x trinucleotídicos	0,43
EST	Dinucleotídicos x tri/tetra/penta	0,54

No final do desenvolvimento de *primers* para regiões microssatélites identificadas em DNA genômico, obtivemos um sucesso de 14,1 %, o que equivale a 45 pares de *primers* desenvolvidos dentre 319 regiões microssatélites identificadas (Figura 9).

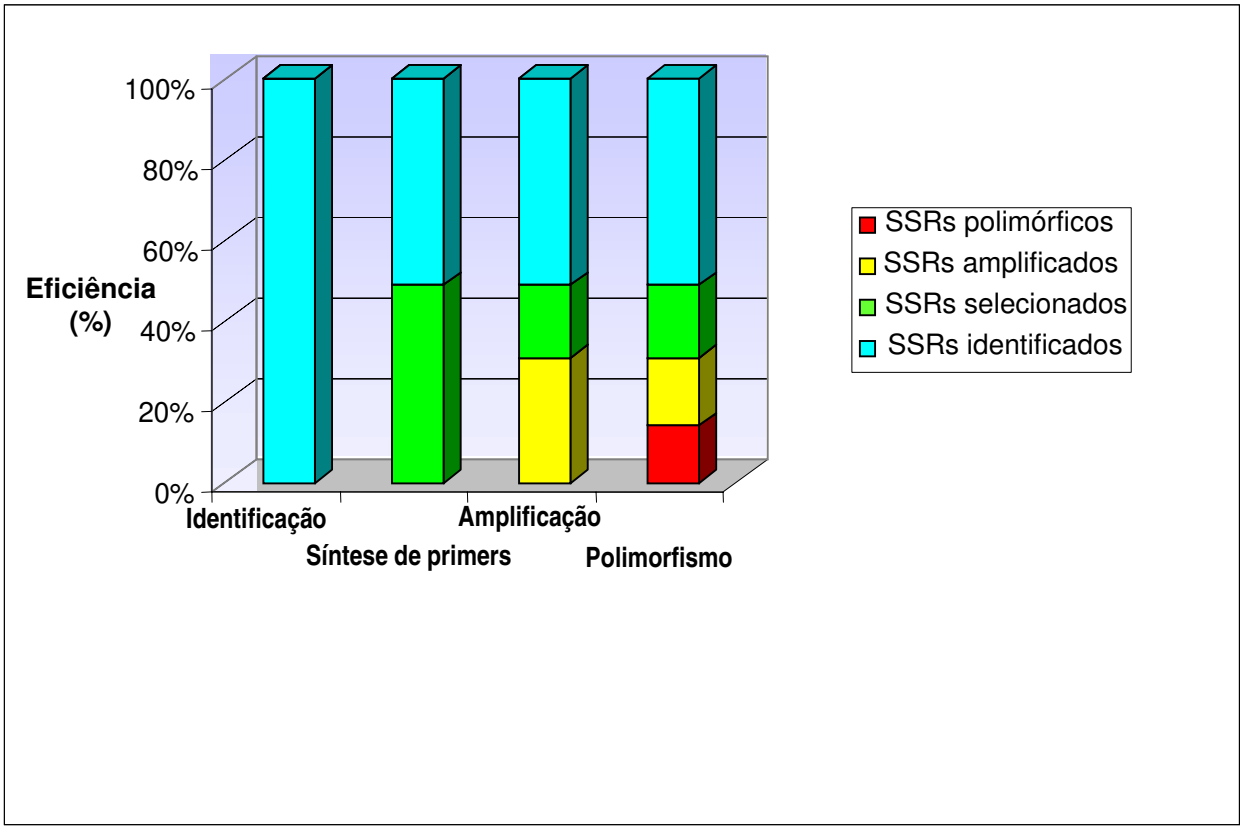

Figura 9: Diagrama mostrando as diferentes fases de desenvolvimento de pares de *primers* para amplificação de locos microssatélites.

Devemos ressaltar aqui que este número é conservador. Isto porque quando foram selecionadas as 156 regiões microssatélites dentre 319, isto foi realizado visando à maximização potencial do sucesso no trabalho de bancada por uma questão de tempo e despesas. É possível obter uma quantidade de pares

de *primers* maior do que a descrita neste trabalho ao analisar melhor as seqüências restantes buscando o desenho de *primers*.

Além disto, muitas outras seqüências microssatélites podem ser identificadas pelo programa *TROLL* a partir de reseqüenciamento e seqüenciamento da outra fita daqueles clones que continham microssatélites e para os quais não foram desenhados pares de *primers* por não conterem região flanqueadora suficientemente grande de um ou ambos os lados ou de seqüência sem qualidade para isto.

Também devemos considerar o número de pares de *primers* polimórficos como sendo um número mínimo de sucesso obtido devido ao fato de que ainda há otimizações e aperfeiçoamentos a serem feitos em muitos dos pares de *primers* que foram considerados não-amplificáveis neste trabalho, e que podem aumentar ainda mais o número de pares de *primers* polimórficos desenvolvidos através de seqüenciamento aleatório de biblioteca genômica por fragmentação (*shotgun*).

5.1.1.3.1 Avaliação de polimorfismo e correlações com dados estruturais

Por meio da discriminação de todos os locos microssatélites amplificados e polimórficos através da atribuição de um escore, baseado no número de alelos diferentes apresentados por cada loco, foi possível procurar correlações entre polimorfismo e características estruturais. Dentre as características testadas, o tamanho da seqüência microssatélite e o número de repetições parecem mostrar uma pequena correlação com nível de polimorfismo do loco microssatélite (r=0,24 e r=0,30, respectivamente), ambas positivas e significativas (teste de t bi-caudal,

α=0,05; t_{calc}=2,65 e 3,31, respectivamente; t_{tab}=1,66). A correlação entre o nível de polimorfismo com o tamanho do motivo também mostrou ser significativa (t_{calc}=-2,38; t_{tab}=1,66), embora ainda menor do que as duas anteriores (r=0,22). Não houve correlação entre o conteúdo nucleotídico da região flanqueadora com polimorfismo (r=0,04), nem entre o conteúdo nucleotídico do microssatélite e o nível de polimorfismo (r=0,09).

5.1.2 Elementos Genéticos Repetitivos

Os 310 elementos genéticos encontrados nas 7.395 seqüências validas analisadas correspondem a 63.019 pares de bases, ou 1,4 % de todas as bases seqüenciadas. Observou-se que há uma grande disparidade entre o número de elementos retrotransponíveis presentes no genoma de *Eucalyptus* (299 retroelementos) e o número de elementos transponíveis (11 transposons) (Tabela 22). Dentre os elementos genéticos retrotransponíveis, notou-se que há uma presença muito maior de elementos do tipo *Long Terminal Repeat* (LTR) e uma menor quantidade de elementos não-LTR, sendo que dos 299 elementos retrotransponíveis, 296 foram do tipo LTR e apenas 3 do tipo não-LTR.

Também notou-se uma disparidade relativamente grande entre os retroelementos do tipo LTR, onde aqueles do tipo *copia-like* compunham a maior parte (244), e os elementos do tipo *gypsy-like* compunham a menor fração (52 elementos). Foi observada elevada diversidade entre 244 retroelementos LTR *copia-like*, e mais uma vez a presença de um tipo mais freqüente sobrepujando os outros, ATCOPIA43, para o qual foram encontrados 47 elementos, quase 20% de

todos retroelementos do tipo *copia-like* (Tabela 23). Dentre os 52 retroelementos LTR do tipo *gypsy-like* uma menor discrepância foi notada, mas o elemento do tipo ATGAGPOL1 é o mais numeroso (Tabela 24). Poucos elementos genéticos repetitivos não-LTR foram identificados (Tabela 25), bem como elementos transponíveis (Tabela 26).

Tabela 22: Número total de seqüências repetitivas identificadas na biblioteca de DNA genômico de *Eucalyptus.*

RETROTRANSPOSONS		
NÃO-LTRs	LINEs	3
	SINEs	0
LTRs	Gypsy-type	52
	Copia-type	244
Total		299
ELEMENTOS DE DNA		
HAT-type		3
MuDR-type		8
En-Spm-type		0
Helicop-typ		0
Total		11
TOTAL GERAL		**310**
MICROSSATÉLITES		
Simple repeats:		**987**
BAIXA COMPLEXIDADE		
Low complexity:		**1636**
Total EGTs		**2933**

Pela comparação de todas as seqüências genômicas obtidas, com o banco não-redundante de seqüências protéicas do NCBI (blastx-nr), com o intuito de procurar seqüências similares a elementos genéticos transponíveis, foi possível identificar 254 seqüências similares a estes elementos, um número relativamente próximo ao encontrado pelo programa *RepeatMasker*. Foi notada, novamente, uma maior proporção de elementos genéticos retrotransponíveis LTR, em

comparação com não-LTRs e transposons, corroborando o resultado obtido pelo programa *RepeatMasker* (Tabela 27, Anexo I).

A análise paralela de um número próximo de ESTs com o programa *RepeatMasker* mostrou que regiões codificantes contem uma complexidade bem inferior em número e tipos de elementos genéticos repetitivos, com exceção de microssatélites, principalmente no que tange a elementos genéticos transponíveis e retrotransponíveis (Tabela 28).

Tabela 23: Elementos genéticos retrotransponíveis do tipo LTR-*copia* identificados em seqüências genômicas pelo programa *RepeatMasker*.

Copia	
ATCOPIA43I	47
ENDOVIR1_I	19
ATCOPIA95_I	12
ATCOPIA63_I	7
ATCOPIA28_I	6
ATCOPIA34_I	6
ATCOPIA78_I	6
ATCOPIA24I	5
ATCOPIA37_I	5
ATCOPIA55_I	5
ATCOPIA1I	4
ATCOPIA32_I	4
ATCOPIA44_I	4
ATCOPIA70_I	4
ATCOPIA10_I	3
ATCOPIA22I	3
ATCOPIA23_I	3
ATCOPIA25I	3
ATCOPIA31_I	3
ATCOPIA33_I	3
ATCOPIA36_I	3
ATCOPIA41I	3
ATCOPIA58_I	3
ATCOPIA59_I	3
ATCOPIA62_I	3
ATCOPIA64_I	3
ATCOPIA68_I	3
ATCOPIA77_I	3
ATCOPIA94_I	3
ATCOPIA11I	2
ATCOPIA12I	2
ATCOPIA13I	2
ATCOPIA17I	2
ATCOPIA20I	2
ATCOPIA26I	2
ATCOPIA3I	2
ATCOPIA40_I	2
ATCOPIA46_I	2
ATCOPIA48_I	2
ATCOPIA49_I	2
ATCOPIA4I	2
ATCOPIA50_I	2
ATCOPIA52	2
ATCOPIA53_I	2
ATCOPIA54_I	2
ATCOPIA56_I	2
ATCOPIA66_I	2
ATCOPIA69_I	2
ATCOPIA74_I	2
ATCOPIA76_I	2
ATCOPIA7I	2
ATCOPIA81_I	2
ATCOPIA85_I	2
ATCOPIA86_I	2
ATCOPIA8AI	2
ATCOPIA19I	1
ATCOPIA2I	1
ATCOPIA32B_I	1
ATCOPIA38A_I	1
ATCOPIA65_I	1
ATCOPIA67_I	1
ATCOPIA73_I	1
ATCOPIA75_I	1
ATCOPIA82_I	1
ATCOPIA83_I	1
ATCOPIA90_I	1
ATCOPIA93_I	1
TA1-2_I	1
TOTAL	244

Tabela 24: Elementos genéticos retrotransponíveis do tipo LTR-*gypsy* identificados em seqüências genômicas pelo programa *RepeatMasker*.

Gypsy	
ATGAGPOL1	19
ATGP8I	10
ATGP2I	7
ATGP5I	5
ATGP3I	3
ATLANTYS3_I	3
ATGP7I	1
ATHILA4C_I	1
ATHILA5_I	1
ATHILA6A_I	1
ATHILA7A_I	1
TOTAL	52

Tabela 25: Elementos genéticos transponíveis identificados em seqüências genômicas pelo programa *RepeatMasker*.

Elementos de DNA	
HAT-type	3
ATHATN6	1
ATHAT1	1
MuDR-type	3
ATMU5	1
ATDNAI26T9	1
ATMU1	1
ATMU10	1
ATMU4	1
ATMU6	1
TOTAL	14

Tabela 26: Elementos genéticos retrotransponíveis do tipo não-LTR identificados em seqüências genômicas pelo programa *RepeatMasker*.

Não-LTRs	
ATLINE1_6	2
ATLINE1_5	1
TOTAL	3

Dentre as ESTs analisadas foram encontradas apenas 1.038 regiões contendo seqüências de baixa complexidade, desta vez mostrando um equilíbrio bastante razoável entre regiões ricas em AT, ricas em GA e ricas em GC, com exceção de regiões ricas em CT. Outra característica observada é um equilíbrio entre ricas em T, ricas em G e ricas em C, apesar de ter sido encontrado um grande número de seqüências ricas em A (Tabela 29).

Tabela 28: Número total de seqüências repetitivas identificadas em ESTs de *Eucalyptus.*

RETROTRANSPOSONS		
NÃO-LTRs	LINEs	0
	SINEs	0
LTRs	Gypsy-type	0
	Copia-type	2
Total		**2**
ELEMENTOS DE DNA		
HAT-type		0
MuDR-type		2
En-Spm-type		1
Helicop-typ		0
Total		**3**
MICROSSATÉLITES		
Simple repeats:		**1529**
BAIXA COMPLEXIDADE		
Low complexity:		**1038**
Total		**2572**

Foram identificadas 1.529 regiões contendo seqüências microssatélites, sendo que a maior parte era composta por uninucleotídeos (A)n. Excluindo-se todos os uninucleotídeos, observou-se a presença de 777 seqüências microssatélites em ESTs, em um total de 4.451 seqüências, enquanto que, em DNA genômico, encontramos 807 microssatélites, dentre 7.395 seqüências.

Tabela 29: Número total de seqüências de baixa complexidade identificadas em 4451 ESTs de *Eucalytpus*.

Baixa complexidade	nº ocorrências
Ricas em A	310
Ricas em GA	229
Ricas em AT	181
Ricas em GC	157
Ricas em CT	46
Ricas em C	39
Ricas em G	34
Ricas em T	32
Polipirimidínicas	8
Polipurínicas	2
TOTAL	1038

Em ESTs, foi observado um número de elementos de DNA transponíveis (8) relativamente maior do que de elementos retrotransponíveis (2), mas estes números são muito pequenos para que se possa detalhar qualquer diferença real que exista entre estes tipos de elementos em ESTs.

5.2 Genes, Fragmentos de Genes e Pseudogenes

5.2.1 Comparação com ESTs

Comparando-se as seqüências genômicas com o banco de ESTs de *Eucalyptus* foi possível obter dados interessantes sobre regiões codificantes deste gênero. Na primeira análise, comparando todas as seqüências com 11.000 ESTs, cerca de 44 seqüências mostraram similaridade com algum EST, e pela

comparação da seqüência da EST com o banco de dados não-redundante de proteínas do NCBI foi possível identificar funções putativas para parte dos possíveis genes (Tabela 30). Contabilizando a quantidade de pares de bases em DNA genômico das seqüências que mostraram similaridade a ESTs de *Eucalyptus*, observou-se que cerca de 2 % das bases seqüenciadas estavam presentes em regiões codificantes.

Tabela 30: Genes identificados em seqüências genômicas de *Eucalyptus* através de comparação com ESTs de duas espécies deste gênero, e posterior comparação dos ESTs com o banco não-redundante de proteínas do NCBI através de "blastx".

Gene	e-value	score
gi|1169534|ENO_RICCO Enolase	7,90E-157	1419
|cl|At1g56600 galactinol synthase, putative	9,00E-36	359
Cytochrome P450 like_TBP (EC 1.14.14.1).	2,80E-49	526
93276 ORF107A	1,00E-20	101
gi|10039641|gb|AAG12204.1|AF287482_5 Orf122 [ChloroBium tepidum]	3,40E-29	323
|cl|Atg213560 malate oxidoreductase(malic enzyme)	8,70E-52	510
Q8VNN2 LacZ protein.	1,00E-23	109
O22405 Cytosolic glucose-6-phosphate dehydrogenase 1(G6PD).	3,00E-49	194
AAQ18140 Enolase.	8,00E-85	173
Q9SKT3 Putative secretory carrier-associated membrane protein	4,00E-33	141
|Q8W1E6] AT5g67540/K9I9_10	2,00E-32	139
tr|AB049723|(SSA-13)Putative senescence-associated protein [P. sativum]	4,00E-28	127
|O64509| Expressed protein (Hypothetical protein) (At2g02730).	2,00E-20	84

Dentre as 44 seqüências identificadas pela comparação com ESTs, 13 puderam ter sua função putativa identificada a partir desta estratégia, dentre as quais 2 enolases e uma glucose-6-fosfatase desidrogenase (G6DP). Outro resultado interessante foi a identificação de um gene putativo envolvido na senescência. Os outros genes identificados foram, em sua maioria, genes responsáveis pela manutenção do metabolismo celular (*housekeeping*) com

atividade catalítica, como malato oxirredutase e galactinol sintase, além de genes que codificam proteínas hipotéticas ou sem função descrita.

Também foi possível observar características de éxons e íntrons. Foram identificados 23 íntrons e apenas 2 éxons inteiros. O tamanho médio dos íntrons foi de 121,2 pares de bases (s=64,7).

5.2.2 Análise utilizando o *Gene Projects*

Pela comparação global de todos os fragmentos de DNA genômico seqüenciados com seqüências de proteínas depositadas no banco de dados não-reduntante do NCBI, foram observadas 166 seqüências que mostraram similaridade elevada, com escore superior a 100 e valor *e* inferior a 10^{-20}, a diversas proteínas de diferentes organismos. Dos 166 possíveis genes identificados nas seqüências genômicas, cerca de 76 eram genes *housekeeping* (Tabela 32), sendo que grande parte destes genes são codificados em cloroplastos. Diversos genes ligados à cadeia de transporte de elétrons e do complexo fotossintético foram identificados. Nesta última categoria foi identificado um número bastante grande de genes ycf1 e ycf2, NADH, PSI 700 apoproteína, citocromos e genes para proteínas ribossomais (Tabela 31, Anexo I).

Vale ressaltar a identificação de pelo menos 5 seqüências genômicas contendo genes inteiros, todos genes envolvidos na manutenção do metabolismo celular (Tabela 36), os quais foram de extrema importância quando procurou-se identificar regiões promotoras.

Cerca de 18 seqüências genômicas mostraram similaridade com genes de função desconhecida, proteína hipotética ou ainda não isolada e estudada,

principalmente de *A. thaliana* e *O. sativa*. Vale acrescentar que, destas, 8 tratavam-se de proteínas hipotéticas, ou seja, preditas por análise computacional de seqüências, e não experimentalmente.

Tabela 32: Genes identificados a partir de comparações de todas seqüências genômicas de *Eucalyptus* utilizando "blastx-nr" no *pipeline* do *Gene Projects.*

Genes	Total
Housekeeping	76
Elementos Genéticos Repetitivos	35
Proteínas desconhecidas ou hipotéticas	18
Sinalização celular	5
Transporte celular	5
Formação de madeira	4
Metabolismo H2O2	3
Resistência à doenças	3
Orf	2
Permease	2
Proteinase	2
Anticorpo	1
Chaperona	1
Ciclo do ácido tricarboxilico	1
Crescimento/Alongamento celular	1
Formação/Maturação fruto	1
Metabolismo sacarose	1
Orf cloroplasto	1
Tolerância Alumínio	1

Os resultados mais interessantes, do ponto de vista prático para o melhoramento da espécie, foram aqueles relativos à resistência a doenças, formação da madeira e alguns genes de plantas que podem ser de interesse para *Eucalyptus*. Foram identificados 10 genes possivelmente ligados à formação da madeira, dentre os quais 5 seqüências que mostraram elevada similaridade a genes de síntese de celulose, 2 seqüências que mostraram similaridade a genes

do metabolismo de peróxido de hidrogênio, 1 seqüência similar a uma pectina metilesterase e 1 seqüência similar a pectato liase, estes 2 últimos genes ligados à diferenciação no câmbio. Além destes genes, ligados à formação de madeira, foram encontrados mais 2 genes de resistência, 2 de proteases, 1 gene de tolerância ao alumínio e 1 gene modulador de resposta a giberelina.

5.2.3 RNAs transportadores

A análise de presença de genes para RNAs transportadores pelo programa *tRNAScan-SE* identificou, dentre as seqüências genômicas de *Eucalyptus*, 11 genes putativos nos singletos e 5 nos agrupamentos contíguos (Tabela 33).

Tabela 33: Possíveis genes de tRNAs identificados nas seqüências genômicas de *Eucalyptus*.

tRNA	Anti-codon	Escore
Asn	GTT	77,16
Cys	GCA	58,87
Tyr	GTA	47,84
Met	CAT	62,47
Leu	CAA	52,25
Unknown	TAN	49,24
Ser	GCT	75,26
Pseudo	GTT	23,43
Glu	TTC	52,15
Tyr	GTA	51,49
Met	CAT	51,61
Arg	TCT	62,51
Thr	GGT	59,05
Trp	CCA	64,59
Val	GAC	51,67
Ser	GGA	49,89

5.2.4 Regiões Promotoras

O alinhamento de genes inteiros e éxons iniciais de genes identificados dentre as seqüências genômicas, com o objetivo de identificar regiões controladoras, mostrou poucos resultados conclusivos. Em um primeiro alinhamento com o programa *Clustal X*, observou-se que havia 3 grupos distintos de seqüências, e para cada um destes grupos um novo alinhamento foi feito, a partir dos quais foi possível observar que os 3 grupos apresentavam resultados pouco conclusivos, mas suas regiões promotoras eram, definitivamente, distintas.

No primeiro grupo (Figura 9a), foi identificada na região 25 a montante do códon de início de transcrição uma região com leve conservação entre as 10 seqüências analisadas, similar a uma *TATA-Box*. Outras duas regiões, uma 40 e outra 60, à montante, também mostraram alguma similaridade e podem estar envolvidas no controle de expressão destes genes.

Entre as 11 seqüências alinhadas no segundo grupo, foram observadas 4 regiões com leve conservação (Figura 9b). Uma delas, bem próximo do códon de início de transcrição, a 10 pares de bases à montante, outra a 25 pares de bases a montante, outra a 45 pares de bases à montante e uma última a 60 pares de bases a montante, com destaque para a segunda, que também se assemelha a uma *TATA-box*.

Nas 10 últimas seqüências, pertencentes ao terceiro grupo, foram observadas 2 regiões que podem estar envolvidas no controle da expressão destes genes, uma a 40 pares de bases a montante e outra a 65 (Figura 9c).

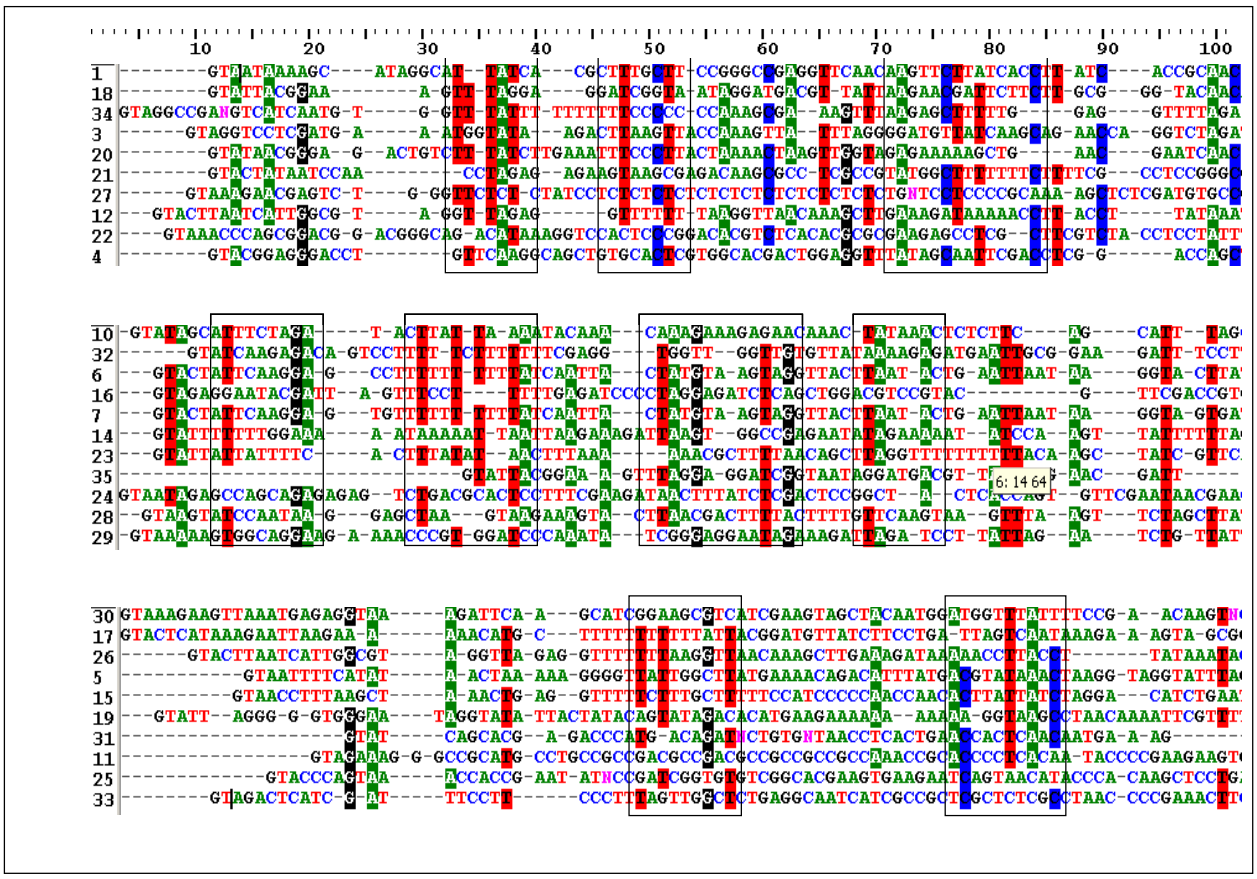

Figura 10: Alinhamento feito pelos programas *BioEdit* e *Clustal X* a partir de regiões a montante de genes identificados em *Eucalyptus*.

5.3 Composição Nucleotídica

5.3.1 Composição GC

O programa *RepeatMasker* calculou uma composição GC geral do genoma de *Eucalyptus* igual a 40,15 %, uma proporção bastante diferente daquela calculada com este mesmo programa em um número equivalente de seqüências de ESTs de *Eucalyptus*, com 49,64 % de GC.

Esta diferença também foi notada entre íntrons e éxons. Íntrons apresentaram, em média, 38,95 % de GC (s=3,88), enquanto que éxons continham em média 45,51 % de GC (s=4,56). Estas médias foram comparadas e consideradas significativamente diferentes (p=0,00064 no teste de t, bi-caudal). Comparando-se a média de GC nos íntrons com a média de conteúdo GC fora dos íntrons, observamos que estas podem ser consideradas iguais (p=0,83292).

Regiões microssatélites também foram avaliadas quanto ao conteúdo nucleotídico de suas regiões flanqueadoras. Em média, regiões microssatélites apresentaram uma média de 43,68 % de GC (s=0,06), o que as tornam estatisticamente diferentes de seqüências genômicas, com média de 40,15 % (p=0,007).

6 DISCUSSÃO

6.1 Seqüências repetitivas

Um número bastante expressivo de seqüências de baixa complexidade foi observado, sendo que o genoma de *Eucalyptus* mostrou ser bastante rico em seqüências ricas em AT. Regiões ricas em AT são comuns a outras dicotiledôneas.

A presença em maior escala de retroelementos do tipo LTR é também consistente com os resultados observados em *Arabdopsis* e *Z. mays*. Em plantas, retroelementos LTR são encontrados em maior quantidade (Meyers *et al.*, 2001) e , em *Eucalyptus*, também observou-se este padrão. Deve-se ressaltar que uma

grande parte dos elementos genéticos transponíveis em plantas é inativa. Todavia, há indícios de que estresses bióticos e abióticos podem causar a ativação e conseqüente propagação destes elementos.

A grande quantidade de elementos transponíveis em *Eucalyptus*, e em quase todas as plantas estudadas até o momento, parece ter relação com a evolução de plantas em geral. Aparentemente os primeiros ancestrais de plantas já possuíam elementos genéticos transponíveis, e é possível que estes elementos possuam papéis importantes na organização do genoma e em outras possíveis funções.

A presença de poucos elementos SINEs e LINEs mostrou que *Eucalyptus* assemelha-se mais a *Arabidopsis*, apesar desta possuir um genoma mais compacto, do que a outras plantas analisadas com extensão de genoma semelhante a *Eucalyptus*, como *Brassica napus* e outras brassicáceas, pertencentes à mesma família de *Arabidopsis*. Desta forma, a quantidade destes elementos não parece estar relacionada à expansão do genoma, como o são retrotransposons, e nem mesmo a parentesco, mas sim a outros fatores. É possível, dentre outros, que a pequena quantidade de SINEs esteja ligada a pequena quantidade de LINEs, o que também foi observado em *Arabidopsis* (Kumar e Bennetzen, 1999) e a fatores como composição nucleotídica e composição de transposons do genoma.

Seqüências poli-purínicas e poli-pirimidínicas não parecem ser comuns em *Eucalyptus*, de acordo com os resultados observados. Apenas 87 destas estruturas foram identificadas, enquanto que ricas em AT e em GC contabilizaram mais de 1.200 seqüências.

6.1.1 Microssatélites

6.1.1.1 Abundância e riqueza de regiões microssatélites

O número de regiões microssatélites identificadas em *Eucalyptus* pelas análises aqui descritas mostrou que este tipo de seqüência é bastante numerosa. O programa *RepeatMasker* indicou 986 seqüências microssatélites, enquanto que o programa *TROLL* permitiu a identificação de 319 destas estruturas. Isto se dá, primariamente, devido a diferenças nos parâmetros para qualificar ou não uma seqüência como sendo microssatélite. Enquanto o programa *RepeatMasker* considera como microssatélite qualquer região com mais de 18 nucleotídeos contendo repetições, independente de serem repetições perfeitas, o programa *TROLL* procura apenas microssatélites perfeitos, di- a pentanucleotídicos.

Com relação à gênese de microssatélites imperfeitos, é possível que regiões microssatélites, sob ausência de pressão de seleção, sofram amplificações errôneas mais freqüentemente que amplificações corretas, gerando um número maior de regiões microssatélites imperfeitas.

Um experimento interessante seria procurar identificar esta mesma relação de microssatélites imperfeitos e perfeitos em bibliotecas enriquecidas, com o intuito de descrever qual estratégia fornece um número relativo maior de regiões microssatélites perfeitas. Isto seria de grande utilidade para a contabilidade geral de regiões microssatélites identificadas e passíveis de definição de *primers* para desenvolvimento de marcadores microssatélites e, por conseqüência, na

descrição de uma estratégia ótima para este fim. Vale ressaltar que isto já foi feito em *Eucalyptus*, mas para poucos locos (Brondani *et al.*, 1998)

Admitindo-se que a porcentagem de 1,3% de nucleotídeos contidos em SSRs calculada para o experimento de *sample sequencing* feitas neste trabalho valha para todo o genoma de *Eucalyptus*, sabendo-se que *E. grandis* tenha cerca de 640 Mpb de extensão e que a média de seqüências microssatélites seja de 35,2 pb, calculou-se que há neste genoma mais de 230.00 seqüências microssatélites, somando-se seqüências perfeitas e imperfeitas. Isto equivale a 1 seqüência microssatélite a cada 3 kb. Por outro lado, estes números mudam bastante quando considera-se os valores encontrados pela análise com o programa *TROLL* foram utilizados, onde foi encontrado que 0,09 % dos nucleotídeos seqüenciados estavam em regiões microssatélites, e que estas eram, em média, 11,3 pb longas. Com base nesta análise, que resumiria então o número total de microssatélites perfeitos, conclui-se que o genoma de *Eucalyptus* pode conter mais de 51.000 seqüências microssatélites seriam identificadas, o que equivale a um microssatélite perfeito a cada 12 kb.

A proporção de microssatélites observada para o genoma de *Eucalyptus* não parece estar de acordo com a tendência observada por Morgante *et al.* (2002), onde genomas menores possuem maior porcentagem de seu genoma composta por regiões microssatélites do que genomas maiores, em plantas. Pode-se citar, por exemplo, *Arabidopsis*, com cerca de 150 Mbp e 0,85% deste total em seqüências microssatélites, enquanto que em *Z. mays*, com cerca de 3.000 Mbp, 0,37% deste total está em seqüências microssatélites (Toth *et al.*, 2000). Era esperado que *Eucalyptus*, com um genoma de 640 Mpb, mostrasse uma

porcentagem entre as descritas acima, já que possui um genoma mais extenso que *Arabidopsis* e menor do que *Z. mays*.

Observou-se relação entre o número de microssatélites encontrado de acordo com o tamanho do motivo, com motivos menores sendo mais numerosos. Esta regra pareceu ser bem adequada para todos os tamanhos de motivo, com exceção de pentanucleotídicos, na análise pelo programa *RepeatMasker*. O número de microssatélites pentanucleotídicos identificado foi muito superior ao esperado quando comparamos com as proporções dos outros motivos. Como este padrão não foi observado na análise com o programa *TROLL*, onde os pentanucleotídicos seguiram o padrão esperado, parece que este tipo de microssatélite é mais suscetível a amplificações imperfeitas.

Foi possível determinar que o motivo (TA)n é quase tão numeroso quanto (TC)n. Assim, a estratégia de identificação de regiões microssatélites e desenvolvimento de pares de *primers* para sua amplificação pela estratégia descrita neste trabalho já mostra resultados bastante expressivos e importantes, já que este motivo não pode ser amostrado pela construção de bibliotecas enriquecidas.

Dentre os outros tipos de microssatélites há uma predominância de composição dos motivos nucleotídeos com as bases nitrogenadas timina e adenina. Isto é observado desde tri- até hexanucleotídeos, onde os motivos mais numerosos identificados foram, em geral, compostos majoritariamente por estes dois nucleotídeos. Provavelmente este padrão de composição de motivos microssatélites é um reflexo da composição nucleotídica geral do genoma de *Eucalyptus* (ver “Composição Nucleotídica” abaixo), que apresenta uma grande

quantidade de regiões de baixa complexidade ricas em AT. Como microssatélites têm origem aleatória, é de se pensar que, se há maior quantidade de nucleotídeos A e T e grande quantidade de regiões AT em *Eucalyptus*, o normal seria apresentar o padrão que foi observado.

A descrição dos motivos mais numerosos é um indicativo bastante significativo para o desenvolvimento de pares de *primers* para regiões microssatélites. Grupos de pesquisa com menor quantidade de recursos podem se valer destas indicações para construir bibliotecas enriquecidas apenas para os motivos mais numerosos, diga-se (TAA)n, (TAAA)n, (GAAAA)n e (TAAAAA)n, dentre di-, tri-, tetra-, penta- e hexanucleotídeos, respectivamente.

Outra observação interessante é a riqueza de motivos identificados dentre os microssatélites. Com exceção de microssatélites hexanucleotídicos, todos os outros tipos de microssatélites apresentaram uma riqueza bastante significante em termos de número de motivos. Como microssatélites hexanucleotídicos são menos comuns do que microssatélites com motivos menores, provavelmente o pequeno número de motivos hexanucleotídicos seja proveniente de um viés que ocorre na amplificação preferencial de apenas alguns motivos microssatélites hexanucleotídicos.

Como esperado, as seqüências microssatélites são mais raras quanto maiores forem suas extensões. Como há indícios de que o nível de polimorfismo de um loco está intimamente ligado ao tamanho da seqüência microssatélite, o fato de que seqüências microssatélites maiores são mais raras indica que futuros desenvolvimentos de pares de *primers* devem considerar quantos locos e qual o nível de polimorfismo desejado são necessários para cada projeto.

Em geral, os motivos mais numerosos identificados pelo programa *RepeatMasker* também foram os mais numerosos na análise do programa *TROLL*. Este fato levanta duas conclusões importantes. Primeiro, como dissemos anteriormente, a análise das seqüências com o programa *TROLL* deve ser considerada com uma subamostragem do total de seqüências microssatélites contidas na biblioteca genômica para seqüenciamento por fragmentação randômica de DNA (*shotgun*). Assim, a identificação dos mesmos motivos mostra, em parte, que esta subamostragem por este programa foi realmente aleatória. Segundo, conclui-se que não há qualquer tendência de determinados motivos serem preferencialmente perfeitos ou imperfeitos, ou seja, a imperfeição dos motivos é aleatória e não parece ter relação com motivos específicos. Uma exceção deve ser colocada, com relação a todos os pentanucleotídicos, a qual já foi descrita neste trabalho, os quais parecem mostrar uma certa tendência a se tornarem mais imperfeitos ou a amplificarem imperfeitamente. Deve-se salientar aqui que esta conclusão não foi estendida para microssatélites hexanucleotídicos porque este tipo de motivo não é normalmente analisado pelo programa *TROLL*.

A análise dos microssatélites pelo programa *TROLL* mostrou ser mais interessante para caracterizá-los com relação ao número médio de repetições de cada tipo e com relação ao tamanho médio da região microssatélite, já que muitos microssatélites identificados pelo programa *RepeatMasker* foram imperfeitos, o que poderia gerar erros nestas estimativas.

Assim, utilizando o programa *TROLL*, foi observado que, em média, microssatélites contendo motivos menores contém um maior número de repetições que microssatélites com motivos maiores. Entretanto, quando o

tamanho final das seqüências microssatélites foi analisado, percebeu-se que os valores se equilibravam.

Um fato interessante foi um padrão observado nessa última análise. Não foram identificadas diferenças significativas em tamanho de fragmento entre microssatélites formados por motivos di- e tetranucleotídicos (p=0,30), nem entre tri- e pentanucleotídicos (p=0,49). Também não foi observada diferença significativa entre o tamanho médio das seqüências microssatélites entre locos com motivos pares (di- e tetranucleotídicos) e locos com motivos ímpares (tri- e pentanucleotídicos), utilizando os valores observados pelo programa *TROLL*.

Entre os microssatélites identificados pelo programa *RepeatMasker*, a situação foi diferente. Microssatélites di- e tetranucleotídicos não apresentaram diferenças significativas de média de tamanho (p=0,05), bem como tri- e pentanucleotídicos (p=0,91). Por outro lado, quando comparou-se di- + tetranucleotídicos com tri- + pentanucleotídicos, notou-se que havia diferenças significativas ($p=7,26x\ 10^{-7}$) , sendo que motivos pares compunham fragmentos menores do que motivos ímpares. Isto vem reforçar ainda mais a hipótese de que motivos pentanucleotídicos e, talvez, trinucleotídicos, mostram maiores imperfeições, já que a análise feita pelo programa *TROLL* não identifica microssatélites imperfeitos e nesta análise não foram detectadas diferenças significativas entre os tamanhos médios dos fragmentos.

Como os resultados mostrados pela análise utilizando o programa *TROLL* mostram apenas microssatélites perfeitos (319), e os resultados provenientes da análise com o programa *RepeatMasker* mostram todos motivos microssatélites

(767 di-, tri-, tetra- e pentanucleotídicos), perfeitos e imperfeitos, pode-se estimar que cerca de 41% dos microssatélites são perfeitos, e 59% são imperfeitos.

6.1.1.2 Comparação entre microssatélites em DNA genômico e em EST

A análise dos ESTs de *Eucalyptus* mostrou que microssatélites nesta espécie parecem conter uma associação preferencial com regiões codificantes, como constatado por Morgante *et al.* (2002) e proposto como modelo geral de ocorrência de microssatélites em genomas de plantas. O número absoluto de microssatélites (1.539) identificados em ESTs (4.451 seqüências) é muito superior ao encontrado (986 microssatélites) na biblioteca por *shotgun* de DNA genômico (7.395 seqüências).

Outro padrão interessante é uma sensível diferença na composição dos motivos microssatélites. Há uma leve tendência de formação de motivos com Citosinas e Adeninas, ao contrário de microssatélites provenientes de DNA genômico, onde havia uma predominância de Timinas e Adeninas. Desta forma, é possível que mesmo em projetos com poucos recursos, que não contam com bibliotecas de ESTs numerosas, seja possível obter microssatélites preferencialmente ligados a genes, apenas utilizando sondas para microssatélites comuns em ESTs e raros em DNA genômico, como (CACAAA)n.

A cada 33 ESTs foi identificado pelo menos 1 microssatélite perfeito. Do ponto de vista prático estes microssatélites são bem interessantes, pois permitirão o mapeamento destes ESTs no mapa genético de *Eucalyptus*, auxiliando a construção de um mapa transcricional.

Assim, bibliotecas de ESTs de *Eucalyptus* parecem ser uma boa fonte de microssatélites. Apesar de poucos microssatélites para esta região terem sido testados (ver abaixo), não parece haver qualquer tipo de limitação em relação ao seu nível de polimorfismo.

6.1.1.3 Desenvolvimento de pares de *primers* para regiões microssatélites

A estratégia de identificação de microssatélites e desenvolvimento de pares de *primers* para estes utilizando bibliotecas genômicas por seqüenciamento por fragmentação randômica de DNA (*shotgun*) e seqüenciamento aleatório (*sample sequencing*) permitiu a geração de 156 pares de *primers*. Juntamente com estes 156, outros 22 pares de *primers* desenvolvidos a partir dos ESTs foram testados.

Aparentemente, microssatélites presentes em ESTs mostraram melhores padrões de amplificação e melhores níveis de polimorfismo (Tabelas 19 e 20) embora não tenha sido detectada diferença estatisticamente significativa em relação àqueles derivados de clones genômicos. Entretanto, esta observação está de acordo com análises de genomas de plantas (Morgante *et al.*, 2002), que mostraram maior polimorfismo em microssatélites derivados de ESTs. Os resultados desta análise sugerem que regiões microssatélites teriam surgido originalmente próximas a genes em um genoma ancestral desprovido de seqüências repetitivas. Assim, a expansão dos microssatélites em regiões transcritas, porém não traduzidas, resultaram no amplo polimorfismo alélico observado hoje em regiões gênicas, porém não traduzidas . Em contraste, os microssatélites em regiões de DNA repetitivo evolutivamente mais recentes, os

microssatélites também se originaram mais recentemente. Pelo menor tempo transcorrido desde sua origem, a oportunidade de geração de variabilidade alélica tem sido menor, resultando , portanto, em menor nível de polimorfismo.

Foi obtido um sucesso total de 45 (14,1%) locos microssatélites polimórficos identificados em DNA genômico, que adicionados a outros 8 locos microssatélites polimórficos identificados em ESTs de *Eucalyptus*, somam 53 novos locos microssatélites para *Eucalyptus*, para os quais *primers* fluorescentes foram encomendados para que estes possam ser acrescentados aos mapas genéticos que estão sendo gerados no âmbito do Projeto Genolyptus.

Devem ser incluídas, neste ponto, diversas observações que fazem com que o sucesso da técnica aqui apresentada deva ser considerado com um número mínimo de obtenção. Primeiramente, deve-se ressaltar que apenas 156 seqüências microssatélites foram selecionadas dentre 319 possíveis. Esta limitação foi feita apenas por motivos operacionais, como tempo e recursos, e porque ainda não se sabia quão interessantes estes resultados se mostrariam. Alguns microssatélites não foram selecionados porque não foi possível desenhar *primers* ótimos, devido à ausência ou baixa qualidade das regiões flanqueadoras, e para estas regiões microssatélites será possível desenhar *primers* a partir do seu re-seqüenciamento ou seqüenciamento da outra fita.

Um segundo ponto a ser colocado é que os testes para amplificação foram feitos em larga escala, onde diversos *primers* falharam por motivos que podem ser ultrapassados. Desta forma, novas reações de otimização estão sendo testadas com ótimos resultados, o que faz acreditar que a expectativa de locos amplificáveis será ampliada.

6.1.1.4 Avaliação de possíveis correlações entre polimorfismo e características estruturais de regiões microssatélites

Pouca ou nenhuma correlação foi encontrada entre os parâmetros estruturais das seqüências microssatélites com o nível de polimorfismo detectado no loco (conteúdo GC das regiões flanqueadoras, conteúdo GC do microssatélite, tamanho da região microssatélite, número de repetições e tamanho do motivo). Apenas tamanho da região, número de repetições e tamanho do motivo parecem mostrar algum resultado positivo.

Entretanto, deve-se ressaltar que o escore utilizado para quantificar o nível de polimorfismo pode estar subestimando o real polimorfismo destes locos, já que, como citado anteriormente, todos os pares de *primers* foram testados em agarose 3 %, o que restringe o poder de visualização de alelos.

6.1.2 Elementos Genéticos Repetitivos

O número de seqüências pertencentes a elementos genéticos repetitivos identificados na biblioteca de DNA genômico ficou bastante aquém do esperado. Como em geral uma maior quantidade destes elementos é observada com o aumento da extensão do genoma, quando comparamos genomas diversos de plantas, esperava-se encontrar um valor próximo aquele observado em *Arabidopsis*, de cerca de 10%. Ao invés disto, foi observado que apenas 1,4% das bases seqüenciadas pertencia a este tipo de seqüência. Mesmo na análise feita comparando-se todas seqüências com o banco de proteínas não-redundante do

NCBI não foi possível obter estimativas maiores do que esta. Isto provavelmente deve-se a presença de elementos genéticos repetitivos ainda não descritos, e a muitos elementos genéticos repetitivos presentes apenas em *Eucalyptus*.

De qualquer forma, foi possível obter diversas observações sobre a presença de elementos genéticos repetitivos em *Eucalyptus*, as quais poderão ser úteis durante a montagem do mapa físico da espécie pelo mascaramento destes elementos mais comuns para redução de alinhamentos errôneos e redução da necessidade de capacidade de cálculo para montagem deste mapa.

Deve-se ressaltar que *Eucalyptus* mostrou padrão de presença de elementos genéticos retrotransponíveis semelhantes a outras plantas já analisadas, como *Arabidopsis* e *Z. mays*, onde há uma predominância de retrotransposons. Foi observada uma quantidade muito maior de elementos retrotransponíveis em relação a DNA transposons. Isto era esperado, já que DNA transposons são mais comuns em mamíferos, mais especificamente em humanos. Provavelmente o número de ambos está subestimado, mas não se espera um incremento muito maior no número de DNA transposons em *Eucalyptus* em detrimento a retroelementos.

Um fato interessante foi notado quando apenas os elementos retrotransponíveis foram analisados. Ao contrário de *Arabidopsis*, foi observado um grande desvio no número de elementos LTR quando comparado com o número de elementos não-LTR. Este mesmo padrão foi observado em milho, o que torna esta análise ainda mais complicada, já que era esperado observar em *Eucalyptus* um padrão semelhante a *Arabidopsis*, uma dicotiledônea, e não a uma monocotiledônea.

Outro desvio interessante foi observado dentro dos elementos LTR. Enquanto em *Arabidopsis* e em *Z. mays* um certo equilíbrio na distribuição de elementos LTR *copia-like* em relação a elementos LTR *gypsy-like* foi observado, em *Eucalyptus* notou-se a presença de uma grande disparidade entre estes elementos, com elementos LTR *copia-like* sendo quase 5 vezes mais freqüentes que *gypsy-like*. Provavelmente há em *Eucalyptus* uma tendência a maior amplificação de *copia-like*. Há evidências, por exemplo, de que o gene da transciptase reversa de elementos *gypsy-like* é relativamente diferente da RT de elementos *copia-like* (Marin e Llorens, 2000), o que poderia causar diferentes taxas de amplificação. Esta diferença pode ser também proveniente de inserção diferencial destes elementos durante a evolução de *Eucalyptus*, e posterior amplificação a taxas iguais. Neste caso, um maior número de elementos *copia-like* deveriam ter sido inseridos. Tanto uma quanto a outra hipótese necessitam de experimentação, mas o fato é que esta discrepância ainda não foi identificada em *Arabidopsis* e *Oryza*.

A análise das seqüências de ESTs utilizando o *RepeatMasker* mostrou poucos resultados no que se refere a elementos genéticos repetitivos, o que leva a crer que elementos transponíveis estão transcricionalmente pouco ativos em *Eucalyptus*, como esperado e como observado em outras espécies (AGI, 2000).

6.2 Genes, Fragmentos de Genes e Pseudogenes

6.2.1 Comparação das seqüências genômicas com ESTs

A partir da comparação dos 766 agrupamentos contíguos de DNA genômico de *Eucalyptus* foi possível obter uma estimativa mínima de DNA codificante para este gênero. Cerca de 44 agrupamentos contíguos mostraram similaridade elevada a ESTs das duas bibliotecas utilizadas, os quais totalizaram 2 % de todas as bases seqüenciadas.

Os resultados da comparação dos ESTs com o banco não-redundante de proteínas (blastx-nr) do NCBI permitiu identificar função putativa para 13 destes agrupamentos contíguos. Os ESTs semelhantes aos outros 31 agrupamentos mostraram resultados irrelevantes (com escore ou valor *e* insuficientes) ou então não mostraram similaridade a nenhuma proteína.

6.2.2 Análise utilizando o *Gene Projects*

Por meio da comparação global de todas seqüências de DNA genômico com todas seqüências depositadas no banco de proteínas do NCBI foi identificado um número consideravelmente maior de possíveis genes. Cerca de 166 seqüências mostraram elevada similaridade com proteínas descritas, desconhecidas ou hipotéticas, o que faz acreditar que o número de genes identificados através da análise por comparação com ESTs está realmente

subestimado, o que reforça a delimitação mínima de DNA codificante em *Eucalyptus* em 2%.

Provavelmente isto ocorreu porque a análise utilizando comparação com ESTs foi limitada apenas aos agrupamentos contíguos, que eram minoria, reduzindo consideravelmente o número de genes encontrados por não terem sido utilizados os singletos obtidos, e por termos comparado as seqüências com poucos ESTs de *Eucalyptus*.

Mesmo a comparação de todas as seqüências genômicas com o banco de proteínas do NCBI através do *Gene Projects* está subestimada, já que seqüências genômicas contendo íntrons muito extensos não conseguem ser analisadas através da comparação padrão feita pelo *blastx*, pois este protocolo não consegue identificar possíveis íntrons, aumentando o número de resultados vazios (*no hits*) ou irrelevantes (escore menor do que 100 e/ou valor *e* maior do que 10^{-20}).

Com relação às possíveis funções identificadas para as seqüências genômicas, resultados mais contundentes foram observados. Como esperado a maior parte dos genes identificados eram genes responsáveis pela manutenção do metabolismo celular (*housekeeping genes*) e genes de vias catabólicas comuns, mas de um número considerável de genes diretamente e indiretamente ligados à formação de madeira foi observado. Foram encontrados 4 genes diretamente ligados á formação de madeira, dentre eles alguns genes de celulose sintase, e outros 6 indiretamente ligados, como genes ligados ao metabolismo de peróxido de hidrogênio e ao metabolismo de sacarose.

6.2.3 Genes de RNAs transportadores

Alguns genes de RNAs transportadores em *Eucalyptus* foram identificados. Entretanto, este número ainda é pouco significativo para que possamos chegar a qualquer conclusão sobre *codon bias* em *Eucalyptus*.

6.2.4 Identificação de regiões controladoras

Poucos resultados conclusivos puderam ser obtidos através do alinhamento de regiões localizadas a montante a genes inteiros ou éxons iniciais de genes, identificados pelo *pipeline Gene Projetcs*. Com exceção de alguns indícios de regiões similares a *TATA-box* e identificação de outras possíveis regiões envolvidas no controle da expressão destes genes, não foi possível encontrar padrões que pudessem identificar, categoricamente, que estas são regiões promotoras.

6.3 Composição nucleotídica

O genoma de *Eucalyptus* parece mostrar composição nucleotídica semelhante a outras angiospermas dicotiledôneas. O conteúdo GC total (40,15%), bastante parecido com aquele apresentado em *Arabidopsis* e em *Lycopersicum*, incrementa ainda mais a discussão sobre as diferenças de composição nucleotídica entre dicotiledôneas e monocotiledôneas, que normalmente apresentam conteúdo GC acima de 50%.

Entre íntrons e éxons também foi possível observar um padrão semelhante a *Arabidopsis*, onde íntrons apresentaram menor conteúdo GC quando

comparados com éxons. Quando o conteúdo nucleotídico de íntrons foi comparado com o conteúdo nucleotídico de éxons, cada um destes também foi comparado com o conteúdo GC geral de *Eucalyptus*, concluindo que éxons são mais ricos em GC do que íntrons e do que seqüências genômicas (aqui considerando todas seqüências genômicas como intergênicas), e que íntrons não possuem diferenças no seu conteúdo nucleotídico quando comparados com seqüências genômicas.

Também interessante foi o resultado observado em regiões microssatélites. Apesar de correlações entre o conteúdo GC e o nível de polimorfismo em locos microssatélites não terem sido observadas, foi possível observar que, em média, regiões adjacentes a microssatélites contém um maior conteúdo GC do que regiões genômicas. Balloux *et al.* (1998) descrevem que, se há correlação negativa entre polimorfismo e conteúdo GC das regiões flanqueadoras, pode-se concluir que níveis mais baixos de GC tornariam a região mais susceptível à mutações. Assim, proporções de conteúdo GC entre diferentes regiões poderiam ser utilizadas para estimar diferentes taxas de mutação para estas regiões. Entretanto, foi observado que regiões flanqueadoras de microssatélites, consideradas altamente polimórficas e, portanto, com elevadas taxas de mutação, contém conteúdo GC mais elevado do que o resto do genoma, com exceção de regiões codificantes, o que tornaria a hipótese levantada, de que o conteúdo GC pode ser utilizado para se estimar a taxa de mutação presente em determinados locais, implausível.

7 CONCLUSÕES E PERSPECTIVAS

7.1 Seqüências repetitivas

A partir da análise do genoma de *Eucalyptus* por *sample sequencing* de uma biblioteca para seqüenciamento por fragmentação randômica de DNA (*shotgun*) foi possível identificar as seqüências repetitivas mais comuns, e estimar a riqueza destas seqüências no genoma de *Eucalyptus*. Além disto os resultados destas estimativas foram utilizados em aplicações imediatas dentro do Projeto Genolyptus, através do desenvolvimento e caracterização de pares de *primers* microssatélites.

As duas metodologias utilizadas para identificar regiões microssatélites pareceram ser bastante eficientes, permitindo a identificação dos principais tipos e motivos microssatélites presentes nesta espécie. Estes dados serão de fundamental importância para futuros projetos de desenvolvimento de novos marcadores moleculares microssatélites, os quais permitirão aumentar o número de marcas nos mapas genéticos de *Eucalyptus*, permitindo a construção de mapas genéticos detalhados.

Como na caracterização dos locos microssatélites desenvolvidos nesta tese foram utilizados os 12 parentais que estão sendo utilizados em cruzamentos no campo, dentro do Projeto Genolyptus, há certa garantia de que todos os pares de *primers* desenvolvidos irão, no mínimo, amplificar os locos em questão, mesmo sabendo que se tratam de 12 indivíduos distintos, de espécies ou híbridos diferentes. Alguns destes pares de *primers* poderão também ser testados em

outras espécies de *Eucalyptus* e até mesmo em outras espécies da família *Myrtaceae*, com o intuito de observar a transferibilidade destes marcadores.

A partir do método descrito neste trabalho para desenvolvimento de *primers* microssatélites, onde considerável tempo e trabalho são economizados, é possível reduzir a pressão sobre o quesito transferibilidade de locos microssatélites quando no desenvolvimento e otimização destes *primers*. Como o seqüenciamento aleatório é mais rápido e atualmente o custo do seqüenciamento em geral mostra uma tendência de redução, é possível desenvolver *primers* através desta estratégia, mesmo para espécies cuja conservação de locos microssatélites não esteja próximo do ideal, ou então desenvolver pares de *primers in silico*, a partir do alinhamento de seqüências microssatélites comuns a diversas espécies e desenho de *primers* degenerados para regiões conservadas, garantindo assim amplificação em diversas espécies próximas.

Dentre quase 8 mil seqüências de DNA genômico analisadas foi possível identificar centenas de regiões microssatélites. Para pelo menos 45 destas regiões pares de *primers* polimórficos foram desenvolvidos. Um dos pontos positivos do desenvolvimento de marcadores moleculares microssatélites através desta técnica é que estes locos estão dispersos aleatoriamente e presentes em regiões antes não analisadas devido a restrições das técnicas anteriormente utilizadas no desenvolvimento destes marcadores. Um outro ponto positivo foi o desenvolvimento de marcadores moleculares microssatélites para motivos (AT)n, os quais nunca haviam sido desenvolvidos, também por restrições das técnicas utilizadas anteriormente.

Pares de *primers* para diversos locos microssatélites contendo repetições tri- e tetranucleotídeas foram desenvolvidos. Alguns marcadores trinucleotídicos já haviam sido desenvolvidos, mas para apenas alguns motivos, e os locos compostos por microssatélites com motivos tetranucleotídicos desenvolvidos neste trabalho são os primeiros deste tipo para *Eucalyptus*.

Deve-se ressaltar que 45 pares de *primers*, em quase 8 mil seqüências, representam um número mínimo de marcadores moleculares baseados em microssatélites desenvolvidos a partir de *sample sequencing*. Como dito anteriormente, uma maior quantidade de marcadores é esperada com a otimização dos *primers* que ainda não amplificaram e com reseqüenciamento de clones considerados ruins pelo programa *TROLL*. Além disto, deve-se lembrar que, por uma limitação do programa, nenhuma região microssatélite com motivos hexanucleotídeos foi analisada quanto à sua amplificação e quanto ao seu nível de polimorfismo, e espera-se que estas possam aumentar ainda mais o número de marcadores moleculares desenvolvidos.

Correlações encontradas entre características estruturais de seqüências microssatélites com o nível de polimorfismo foram, em geral, insatisfatórias. Entretanto, o tamanho da região microssatélite, bem como o número de repetições e o tamanho do motivo, mostraram correlação positiva com o nível de polimorfismo.

Foi possível observar que ESTs são mais ricos em regiões microssatélites que regiões genômicas, e que locos microssatélites presentes nestas seqüências são, no mínimo, tão polimórficos quanto qualquer outro loco microssatélite presente no genoma. Devido ao grande número de ESTs já depositados no banco

de seqüências do Projeto Genolyptus, será possível obter centenas, até milhares de marcadores moleculares microssatélites para *Eucalyptus*, e possivelmente a construção de um mapa transcricional para o gênero, e talvez para as várias espécies que estão sendo utilizadas pelo projeto.

A identificação das principais famílias de elementos genéticos transponíveis mostrou que o genoma de *Eucalyptus* contém um número significativo destes elementos, com destaque para elementos retrotransponíveis do tipo LTR *copia-like*.

Análises mais apuradas e dispendiosas devem ser feitas, já que notou-se que o número total de elementos genéticos transponíveis parece estar bem subestimado. Será possível, por exemplo, comparar todas seqüências com elas mesmas, permitindo a identificação de seqüências repetitivas dentro da própria biblioteca. Selecionando-se seqüências semelhantes e analisando-se sua estrutura, será possível identificar elementos transponíveis ainda não estudados e talvez elementos presentes apenas em *Eucalyptus*.

7.2 Genes, Fragmentos de Genes e Pseudogenes

Poucos genes puderam ser identificados nas seqüências genômicas analisadas neste trabalho, e notou-se que o número de genes também foi subestimado. Entretanto, foi possível identificar alguns genes interessantes e estimar o limite mínimo do genoma de *Eucalyptus* que contém genes, um pouco acima de 2%.

Todos os clones que mostraram similaridade a genes interessantes para *Eucalyptus* são ponto de partida para outros diversos experimentos, principalmente aqueles que objetivem a identificação do gene completo. Como a biblioteca por *shotgun* construída neste trabalho contém clones com insertos entre 0,5 e 4,0 kb, é possível re-seqüenciar e/ou "caminhar" pelo clone até que toda sua seqüência seja obtida, para então serem comparadas com todas genes depositados em bancos de seqüências públicos. Uma outra alternativa tendo o mesmo objetivo seria a utilização de clones genômicos desta biblioteca como sondas para identificação de BACs contendo a seqüência completa destes genes.

As regiões a montante de genes completos e éxons iniciais de genes também merecem maiores análises. É possível que novos genes sejam identificados através da comparação destas seqüências com o banco de ESTs do Projeto Genolyptus, a qual já possui mais de 70 mil seqüências depositadas, o que permitiria a identificação de um maior número de seqüências contendo regiões controladoras e, portanto, melhores definições destas regiões, além de permitir estimar com maior confiabilidade qual o percentual do genoma de *Eucalyptus* que contém regiões codificantes.

7.3 Conteúdo nucleotídico

Além de diferenças de conteúdo nucleotídico entre DNA genômico e ESTs e entre íntrons e éxons, verificou-se que regiões microssatélites diferem significativamente de regiões genômicas, possuindo uma maior quantidade de nucleotídeos guanina e citosina.

Paralelamente, foi observado que motivos microssatélites localizados em DNA genômico também são preferencialmente compostos por adenina e timina, enquanto que microssatélites presentes em ESTs contém uma composição mais rica em citosinas e adeninas. Isto provém, provavelmente, das diferenças de composição nucleotídica nas regiões genômicas e regiões codificantes.

8 Considerações finais

Além de todas estas descrições e contribuições para o desenvolvimento de marcadores moleculares microssatélites, deve-se destacar a importância das seqüências geradas neste trabalho no âmbito do Projeto Genolyptus. Na prática, tanto as seqüências em si, como os resultados extraídos de suas análises, serão brevemente aplicadas em outros projetos.

As seqüências genômicas podem ser utilizadas na detecção de contaminação de bibliotecas de ESTs por DNA genômico, o que será de grande utilidade na aferição da qualidade destas bibliotecas, economizando tempo e dinheiro quando da eliminação de bibliotecas com taxas elevadas de contaminação.

Um outro projeto que se beneficiará dos resultados gerados é o projeto de montagem do mapa físico de *Eucalyptus* por meio da ligação de pontas de BACs, uma das estratégias que podem ser utilizadas na construção do mapa físico. Devido ao elevado número de seqüências repetitivas presentes no genoma de plantas superiores, bem como devido à grande similaridade que estas apresentam

umas em relação às outras, é necessário que se tenha conhecimento de quais são os tipos e quantas destas seqüências repetitivas há no genoma de *Eucalyptus* para que durante o processo de montagem dos contíguos de BACs estas seqüências possam ser mascaradas, simplificando o processo e reduzindo montagens errôneas.

Desta forma, a análise do genoma de *Eucalyptus* utilizando *sample sequencing* de uma biblioteca genômica construída para seqüenciamento por fragmentação randômica de DNA (*shotgun*) mostrou ser bastante informativa, possibilitando acessar e descrever aspectos nunca ou pouco estudados no genoma do gênero, além de ter mostrado ser uma estratégia alternativa para se obter panoramas gerais de genomas muito extensos.

9 Referências Bibliográficas

AGI. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, v.408,n.(6814): p. 796-815, 2000.

Aguero, F., R. E. Verdun*, et al.* A random sequencing approach for the analysis of the *Trypanosoma cruzi* genome: general structure, large gene and repetitive DNA families, and gene discovery. Genome Res, v.10,n.(12): p. 1996-2005, 2000.

Altschul, S. F., T. L. Madden*, et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, v.25,n.(17): p. 3389-3402, 1997.

Andrade, E. d. O eucalipto.ed. S. Paulo,: Chacaras e quintais, 1939.

Aubourg, S. and P. Rouzé. Genome Annotation. Plant Physiol. Biochem., v.39: p. 181-193, 2001.

Ausubel, F. M. Short protocols in molecular biology : a compendium of methods from Current protocols in molecular biology. 5th.ed. New York: Wiley, 2002.

Balloux, F., E. Ecoffey*, et al.* Microsatellite conservation, polymorphism, and GC content in shrews of the genus *Sorex* (*Insectivora*, *Mammalia*). Mol Biol Evol, v.15,n.(4): p. 473-475, 1998.

Bancroft, I. Duplicate and diverge: the evolution of plant genome microstructure. Trends Genet, v.17,n.(2): p. 89-93, 2001.

Bankier, A. T., K. M. Weston*, et al.* Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. Methods Enzymol, v.155: p. 51-93, 1986.

Bell, K. S., A. O. Avrova*, et al.* Sample sequencing of a selected region of the genome of *Erwinia carotovora* subsp. *atroseptica* reveals candidate phytopathogenicity genes and allows comparison with *Escherichia coli*. Microbiology, v.148,n.(Pt 5): p. 1367-1378, 2002.

Bernardi, G. The isochore organization of the human genome. Annu Rev Genet, v.23: p. 637-661, 1989.

Bracelpa. Avaliação do Setor de Papel e Celulose. http://www.bracelpa.org.br/#. 2003.

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

Brondani, R. and D. Grattapaglia. Discovery and development of microsatellite-based markers in *Eucalyptus*. IUFRO Conference on Silviculture and Improvement of Eucalyptus, Salvador, Colombo: EMBRAPA- Centro Nacional de Pesquisa de Florestas, (1997).

Bruskiewich, R. GeeCee. Cambridge, EMBOSS, 1999.

Byrne, N., M. Marquez-Garcia*, et al.* Conservation and genetic diversity of microsatellite loci in the genus *Eucalyptus*. Aust. J. Bot., v.44: p. 331-341, 1996.

Castelo, A. T., W. Martins*, et al.* TROLL--tandem repeat occurrence locator. Bioinformatics, v.18,n.(4): p. 634-636, 2002.

Cavalcanti, G. Identificação das Principais Espécies de Eucalyptus Existentes no Brasil. Silvicultura em São Paulo, v.1,n.(2): p., 1963.

Cho, Y. G., Y. G. Cho*, et al.* Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). Theor Appl Genet, v.100: p. 713-722, 2000.

Chou, H. H. and M. H. Holmes. DNA sequence quality trimming and vector removal. Bioinformatics, v.17,n.(12): p. 1093-1104, 2001.

Comings, D. E. Mechanisms of chromosome banding and implications for chromosome structure. Annu Rev Genet, v.12: p. 25-46, 1978.

Creste, S., A. Neto*, et al.* Detection of single sequence polymorphisms in denaturing polyacrylamide sequencig gels by silver staning. Plant Molecular Biology Reporter, v.19: p. 299-306, 2001.

Doyle, J. and J. Doyle. Isolation of Plant DNA from Fresh Tissue. Focus, v.12,n.(1): p. 13-15, 1986.

Edwards, Y. J., G. Elgar*, et al.* The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analyses. J Mol Biol, v.278,n.(4): p. 843-854, 1998.

Eldridge, K. Eucalypt domestication and breeding.ed. Oxford, New York: Clarendon Press ;Oxford University Press, 1994.

Elgar, G., M. S. Clark*, et al.* Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning. Genome Res, v.9,n.(10): p. 960-971, 1999.

Ewing, B. and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res, v.8,n.(3): p. 186-194, 1998.

Ewing, B., L. Hillier*, et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res, v.8,n.(3): p. 175-185, 1998.

FAO. Eucalypts for Planting: Forestry Ser.ed.: Fao, 1981.

Ferreira, M. and D. Grattapaglia. Introdução ao Uso de Marcadores Moleculares am Análise Genética. 3.ed. Brasília: EMBRAPA-Cenargen, 1998.

Fleischmann, R. D., M. D. Adams*, et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, v.269,n.(5223): p. 496-512, 1995.

Forests. FORESTs: Eucalyptus Genome Sequencing Project Consortium. http://watson.fapesp.br/PITE/Agrarias/engflor2.htm. 2003.

Freimer, N. B. and M. Slatkin. Microsatellites: evolution and mutational processes. Ciba Found Symp, v.197: p. 51-67; discussion 67-72, 1996.

Garner, T. W. Genome size and microsatellites: the effect of nuclear size on amplification potential. Genome, v.45,n.(1): p. 212-215, 2002.

Genolyptus, P. Projeto Genolyptus. www.genolyptus.ucb.br. 2003.

Glenn, T. C., W. Stephan*, et al.* Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. Mol Biol Evol, v.13,n.(8): p. 1151-1154, 1996.

Goff, S. A., D. Ricke*, et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science, v.296,n.(5565): p. 92-100, 2002.

Goldstein, D. and C. Schlotterer. Microsatellites: Evolution and Applications.ed. Oxford: Oxford University Press, 1999.

González, E., A. d. Andrade*, et al.* Transformação genética do Eucalipto. Biotecnologia Ciência & Desenvolvimento, v.5,n.(26): p. 18-22, 2002.

Goodall, G. J. and W. Filipowicz. Different effects of íntron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. Embo J, v.10,n.(9): p. 2635-2644, 1991.

Grattapaglia, D. Marcadores moleculares em espécies florestais: *Eucalyptus* como modelo. In: L. L. Nass and A. C. C. Valois Recursos Genéticos e Melhoramento de Plantas. 1.ed. Rondonópolis, MT: Fundação MT. 2001.p. 1183.

Grattapaglia, D. Genolyptus. In: A. Borém, M. Giúdice and T. Sediyama Melhoramento Genômico.ed. Viçosa: Universidade Federal de Viçosa. 2003.p. 224.

Grattapaglia, D. and H. J. Bradshaw. Nuclear DNA content of commercially important Eucalyptus species and hybrids. Canadian Journal of Forest Research, v.24,n.(5): p. 1074-1078, 1994.

Grattapaglia, D. and R. Sederoff. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics, v.137,n.(4): p. 1121-1137, 1994.Hall, T. BioEdit. Raleigh, 2001.

Hancock, J. M. The contribution of slippage-like processes to genome evolution. J Mol Evol, v.41,n.(6): p. 1038-1047, 1995.

Huang, X. and A. Madan. CAP3: A DNA sequence assembly program. Genome Res, v.9,n.(9): p. 868-877, 1999.

Karlyshev, A. V., J. Henderson*, et al.* Procedure for the investigation of bacterial genomes: random shot-gun cloning, sample sequencing and mutagenesis of *Campylobacter jejuni*. Biotechniques, v.26,n.(1): p. 50-52, 54, 56, 1999.

Kumar, A. and J. L. Bennetzen. Plant retrotransposons. Annu Rev Genet, v.33: p. 479-532, 1999.

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

Lowe, T. M. and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, v.25,n.(5): p. 955-964, 1997.

Marck, C. and H. Grosjean. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. Rna, v.8,n.(10): p. 1189-1232, 2002.

Marin, I. and C. Llorens. Ty3/Gypsy retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. Mol Biol Evol, v.17,n.(7): p. 1040-1049, 2000.

McClintock, B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A, v.36: p. 344-355, 1950.

Meyers, B. C., S. V. Tingey*, et al.* Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res, v.11,n.(10): p. 1660-1676, 2001.

Mizuno, M. and M. Kanehisa. Distribution profiles of GC content around the translation initiation site in different species. FEBS Lett, v.352,n.(1): p. 7-10, 1994.

Mora, A. and C. Garcia. A Cultura do Eucalipto no Brasil. 1.ed. São Paulo: Sociedade Brasileira de Silvicultura, 2000.

Morgante, M., M. Hanafey*, et al.* Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet, v.30,n.(2): p. 194-200, 2002.

Pearce, S. R., G. Harrison*, et al.* The Ty1-copia group retrotransposons in *Vicia* species: copy number, sequence heterogeneity and chromosomal localisation. Mol Gen Genet, v.250,n.(3): p. 305-315, 1996.

Pertsemlidis, A. and J. W. Fondon, 3rd. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol, v.2,n.(10): p. REVIEWS2002, 2001.

Pryor, L. *Eucalyptus*. In: A. Halevy CRC Handbook of Flowering. 1.ed. Boca Raton: Franklin Book Company, Incorporated. 1985. v.2:p. 476-482.

Ramsay, L., M. Macaulay*, et al.* Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J, v.17,n.(4): p. 415-425, 1999.

Rozen, S. and H. Skaletsky. *Primer*3 on the WWW for general users and for biologist programmers. Methods Mol Biol, v.132: p. 365-386, 2000.

Sato, S., T. Kaneko*, et al.* Structural analysis of *Arabidopsis thaliana* chromosome 5. IV. Sequence features of the regions of 1,456,315 bp covered by nineteen physically assigned P1 and TAC clones. DNA Res, v.5,n.(1): p. 41-54, 1998.

SBS. Sociedade Brasileira de Silvicultura. www.sbs.org.br. 2002.

Scharf, R. Falta madeira na terra do Pau-Brasil. Revista Galileu, May. p. 52-60.

Schlotterer, C. and D. Tautz. Slippage synthesis of simple sequence DNA. Nucleic Acids Res, v.20,n.(2): p. 211-215, 1992.

Schlueter, S. D., Q. Dong*, et al.* GeneSeqer@PlantGDB: Gene structure prediction in plant genomes. Nucleic Acids Res, v.31,n.(13): p. 3597-3600, 2003.

Slater, G. ESTate - Expressed Sequence Tag Analysis Tools. Cambridge, Slater, GStC, 1999.

Smith, A. and P. Green. Repeatmasker Web Server, 2001.

Southerton, S. G., S. H. Strauss*, et al.* Eucalyptus has a functional equivalent of the Arabidopsis floral meristem identity gene LEAFY. Plant Mol Biol, v.37,n.(6): p. 897-910, 1998.

Sperling, L., P. Dessen*, et al.* Random sequencing of *Paramecium* somatic DNA. Eukaryot Cell, v.1,n.(3): p. 341-352, 2002.

Thompson, J. D., T. J. Gibson*, et al.* The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, v.24: p. 4876-4882, 1997.

Toth, G., Z. Gaspari*, et al.* Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res, v.10,n.(7): p. 967-981, 2000.

Vinogradov, A. E. DNA helix: the importance of being GC-rich. Nucleic Acids Res, v.31,n.(7): p. 1838-1844, 2003.

Watson, J. D. Recombinant DNA. 2nd.ed. New York: Scientific American Books : Distributed by W.H. Freeman, 1992.

Weber, J. L. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics, v.7,n.(4): p. 524-530, 1990.

White, S. E., L. F. Habera*, et al.* Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. Proc Natl Acad Sci U S A, v.91,n.(25): p. 11792-11796, 1994.

Wong, R. M., K. K. Wong*, et al.* Sample sequencing of a *Salmonella typhimurium* LT2 lambda library: comparison to the Escherichia coli K12 genome. FEMS Microbiol Lett, v.173,n.(2): p. 411-423, 1999.

Yu, J., S. Hu*, et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science, v.296,n.(5565): p. 79-92, 2002.

10 Anexos

Tabela 27: Resultado da quantificação de elementos genéticos transponíveis pela comparação com o banco de dados não-redundante de proteínas do NCBI, pelo *Gene Projects*.

Genes putativos	e-value
putative retrotransposon-related protein [Oryza sativa]	6,00E-53
similar to reverse transcriptase, putative; protein id: At5g17725.1 [Arabidopsis thaliana]	5,00E-09
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	8,00E-58
putative polyprotein [Kalanchoe top-spotting virus]	2,00E-05
putative transposon protein [Arabidopsis thaliana]	9,00E-15
(AF053008) gag-pol polyprotein [Glycine max]	2,00E-33
putative polyprotein [Oryza sativa]	4,00E-43
Putative polyprotein [Oryza sativa]	3,00E-48
(AC079852) Putative polyprotein [Oryza sativa]	2,00E-08
putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	7,00E-18
similar to copia-type pol polyprotein [Oryza sativa (japonica cultivar-group)]	2,00E-46
(D85597) polyprotein [Oryza australiensis]	5,00E-15
(AY016208) gag/pol polyprotein [Arabidopsis thaliana]	2,00E-48
putative Ty3-gypsy-like retroelement pol polyprotein [Arabidopsis thaliana]	0.003
putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	1,00E-05
protein Ty1/copia-element polyprotein [imported] - Arabidopsis thaliana	9,00E-08
polyprotein [Arabidopsis thaliana]	2,00E-29
(AC079852) Putative polyprotein [Oryza sativa]	2,00E-23
mutator-like transposase-like [Arabidopsis thaliana]	1,00E-17
(AC037197) Putative copia-type polyprotein [Oryza sativa]	0.003
(AF116598) gag-pol polyprotein [Vitis vinifera]	7,00E-36
reverse transcriptase homolog T7M24.7 - Arabidopsis thaliana	0.0004
gag-protease polyprotein [Cucumis melo]	6,00E-09
(Y12432) polyprotein [Ananas comosus]	6,00E-41
(AP002521) putative retroelement pol polyprotein (AC006920) [Oryza sativa (japonica	2,00E-17
putative gag-pol precursor [Oryza sativa (japonica cultivar-group)]	2,00E-35
putative polyprotein [Oryza sativa (japonica cultivar-group)]	2,00E-39
putative gag-pol polyprotein [Zea mays]	2,00E-70
(NM_148832) similar to putative retroelement pol polyprotein; protein id: At3g43955.1	2,00E-37
(D85597) polyprotein [Oryza australiensis]	4,00E-24
copia-like retroelement pol polyprotein [Arabidopsis thaliana]	7,00E-28
(AC007018) putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	4,00E-42
(AF275345) putative copia-like polyprotein [Lycopersicon esculentum]	5,00E-51
non-LTR retroelement reverse transcriptase-like protein [Arabidopsis thaliana]	2,00E-17
(AB072492) integrase [Silene latifolia]	8,00E-62
retrotransposon del1-46 [Lilium henryi]	0.0001
(NM_103299) gag-pol polyprotein, putative [Arabidopsis thaliana]	0.0001
(AF053008) gag-pol polyprotein [Glycine max]	2,00E-07
putative retroelement pol polyprotein; protein id: At2g21310.1 [Arabidopsis thaliana]	3,00E-27

(AC007730) putative non-LTR retrolelement reverse transcriptase [Arabidopsis thaliana]	6,00E-15
(D85597) polyprotein [Oryza australiensis]	1,00E-25
(AC006248) putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	5,00E-05
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	3,00E-47
(AF369930) pol polyprotein [Citrus x paradisi]	8,00E-28
(Y12432) polyprotein [Ananas comosus]	1,00E-11
copia-type polyprotein - Arabidopsis thaliana	1,00E-23
integrase [Nicotiana tabacum]	2,00E-43
reverse transcriptase homolog T7M24.7 - Arabidopsis thaliana	2,00E-68
probable copia-type polyprotein T18I24.5 [imported] - Arabidopsis thaliana	1,00E-23
putative retroelement pol polyprotein; protein id: At2g21310.1 [Arabidopsis thaliana]	9,00E-23
reverse transcriptase homolog T7M24.7 - Arabidopsis thaliana	2,00E-52
(AP000969) Similar to copia-type pol polyprotein. (AF105716) [Oryza sativa (japonica cultivar-	1,00E-92
(AF077409) similar to reverse transcriptases (PFam: rvt.hmm, score: 60.13) [Arabidopsis	3,00E-44
putative non-LTR retroelement reverse transcriptase [Oryza sativa (japonica cultivar-group)]	2,00E-10
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	9,00E-75
putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	1,00E-54
probable reverse transcriptase, 100033-105622 [imported] - Arabidopsis thaliana	0.002
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	6,00E-46
(U22103) gag-protease polyprotein [Glycine max]	0.002
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	3,00E-36
gag-pol polyprotein - maize copia-like retrotransposon Sto-4	6,00E-59
(AF053008) gag-pol polyprotein [Glycine max]	3,00E-53
gag-pol polyprotein [Vitis vinifera]	1,00E-05
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	3,00E-07
putative gag-pol polyprotein, 3'-partial [Oryza sativa]	1,00E-16
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	6,00E-20
polyprotein, putative [Arabidopsis thaliana]	8,00E-28
(AF369930) pol polyprotein [Citrus x paradisi]	1,00E-10
(AC091735) Putative retroelement [Oryza sativa] [Oryza sativa (japonica cultivar-group)]	1,00E-16
Putative retroelement [Oryza sativa (japonica cultivar-group)]	4,00E-26
retrotransposon like protein; protein id: At4g10690.1 [Arabidopsis thaliana]	7,00E-16
(AC005560) putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	0.005
(AC005825) putative retroelement pol polyprotein [Arabidopsis thaliana]	8,00E-41
(AC005825) putative retroelement pol polyprotein [Arabidopsis thaliana]	2,00E-41
putative gag-pol polyprotein [Zea mays]	9,00E-53
polyprotein [Lycopersicon esculentum]	4,00E-19
putative polyprotein [Oryza sativa (japonica cultivar-group)]	3,00E-15
retrotransposon del1-46 [Lilium henryi]	6,00E-27
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	4,00E-30
polyprotein [Lycopersicon esculentum]	2,00E-09
putative retroelement pol polyprotein; protein id: At2g21310.1 [Arabidopsis thaliana]	1,00E-07
polyprotein-like [Oryza sativa (japonica cultivar-group)]	1,00E-07
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	5,00E-38
putative retroelement pol polyprotein [Arabidopsis thaliana]	2,00E-14

polyprotein-like [Oryza sativa (japonica cultivar-group)]	1,00E-07
ORF I polyprotein [petunia vein clearing virus]	1,00E-48
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	0.007
ORF I polyprotein [petunia vein clearing virus]	1,00E-48
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	5,00E-38
pol protein [Cucumis melo]	8,00E-64
putative retroelement pol polyprotein [Arabidopsis thaliana]	2,00E-14
pol protein [Cucumis melo]	6,00E-51
mutator-like transposase-like protein [Arabidopsis thaliana]	2,00E-84
(U22103) gag-protease polyprotein [Glycine max]	7,00E-20
gag-pol polyprotein [Vitis vinifera]	4,00E-32
putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	2,00E-09
(AY016208) gag/pol polyprotein [Arabidopsis thaliana]	8,00E-30
putative retroelement pol polyprotein [Arabidopsis thaliana]	0.0003
pol protein [Cucumis melo]	8,00E-64
(AY016208) gag/pol polyprotein [Arabidopsis thaliana]	2,00E-33
copia-type polyprotein, putative [Arabidopsis thaliana]	7,00E-09
putative non-LTR retroelement reverse transcriptase [Oryza sativa (japonica cultivar-group)]	1,00E-08
Putative retroelement [Oryza sativa (japonica cultivar-group)]	6,00E-12
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	3,00E-57
similar to putative retroelement pol polyprotein; protein id: At4g09425.1 [Arabidopsis thaliana]	2,00E-14
putative polyprotein [Oryza sativa (japonica cultivar-group)]	5,00E-48
putative polyprotein [Oryza sativa (japonica cultivar-group)]	2,00E-42
(AF082134) pol polyprotein [Zea mays]	3,00E-51
polyprotein [Lycopersicon esculentum]	2,00E-16
probable LTR retrotransposon - Arabidopsis thaliana	2,00E-22
(AC092548) putative reverse transcriptase [Oryza sativa (japonica cultivar-group)]	5,00E-08
putative gag-pol polyprotein [Zea mays]	8,00E-38
Putative Zea mays retrotransposon Opie-2 [Oryza sativa (japonica cultivar-group)]	8,00E-08
Putative Zea mays retrotransposon Opie-2 [Oryza sativa (japonica cultivar-group)]	2,00E-41
(Y12432) polyprotein [Ananas comosus]	1,00E-11
(Y12432) polyprotein [Ananas comosus]	1,00E-23
copia-like retroelement pol polyprotein [Arabidopsis thaliana]	1,00E-31
(AF369930) pol polyprotein [Citrus x paradisi]	4,00E-33
(NM_148832) similar to putative retroelement pol polyprotein; protein id: At3g43955.1	7,00E-15
(AC005825) putative retroelement pol polyprotein [Arabidopsis thaliana]	2,00E-37
putative gag-pol polyprotein [Zea mays]	4,00E-49
putative non-LTR retroelement reverse transcriptase [Oryza sativa (japonica cultivar-group)]	6,00E-38
(U22103) gag-protease polyprotein [Glycine max]	0.002
(AC092548) putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	2,00E-13
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	9,00E-32
(AF369930) pol polyprotein [Citrus x paradisi]	7,00E-11
Putative 22 kDa kafirin cluster; Ty3-Gypsy type [Oryza sativa]	3,00E-10
(NM_127718) putative retroelement pol polyprotein [Arabidopsis thaliana]	0.005
putative gag-pol polyprotein [Zea mays]	2,00E-55

(Y13368) reverse transcriptase [Beta vulgaris]	1,00E-20
Putative Zea mays retrotransposon Opie-2 [Oryza sativa (japonica cultivar-group)]	3,00E-13
putative AP endonuclease/reverse transcriptase [Brassica napus]	0.0001
(AC005724) putative reverse transcriptase [Arabidopsis thaliana]	2,00E-17
ORF I polyprotein [petunia vein clearing virus]	8,00E-47
reverse transcriptase homolog T7M24.7 - Arabidopsis thaliana	1,00E-05
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	3,00E-29
(AC092548) putative reverse transcriptase [Oryza sativa (japonica cultivar-group)]	3,00E-17
(AF053008) gag-pol polyprotein [Glycine max]	3,00E-60
(AF082134) pol polyprotein [Zea mays]	1,00E-51
(AB018120) non-LTR retroelement reverse transcriptase-like [Arabidopsis thaliana]	2,00E-22
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	2,00E-30
similar to putative retroelement pol polyprotein; protein id: At2g01034.1 [Arabidopsis thaliana]	3,00E-52
gag-pol polyprotein [Vitis vinifera]	3,00E-21
(NM_113875) transposon related protein [Arabidopsis thaliana]	1,00E-53
(AB018120) non-LTR retroelement reverse transcriptase-like [Arabidopsis thaliana]	1,00E-05
Putative Zea mays retrotransposon Opie-2 [Oryza sativa (japonica cultivar-group)]	5,00E-59
putative reverse transcriptase [Oryza sativa]	0.0005
similar to reverse transcriptase, putative; protein id: At5g17725.1 [Arabidopsis thaliana]	0.007
(AC079852) Putative polyprotein [Oryza sativa]	8,00E-12
putative polyprotein [Oryza sativa (japonica cultivar-group)]	7,00E-28
ORF I polyprotein [petunia vein clearing virus]	2,00E-51
gag-pol polyprotein - maize copia-like retrotransposon Sto-4	3,00E-24
putative retroelement pol polyprotein [Arabidopsis thaliana]	2,00E-14
(Y12432) polyprotein [Ananas comosus]	3,00E-07
mutator-like transposase-like [Arabidopsis thaliana]	3,00E-22
(AF053008) gag-pol polyprotein [Glycine max]	2,00E-17
(AF053008) gag-pol polyprotein [Glycine max]	4,00E-26
(AC079852) Putative polyprotein [Oryza sativa]	1,00E-21
(D85597) polyprotein [Oryza australiensis]	6,00E-23
putative polyprotein [Oryza sativa (japonica cultivar-group)]	2,00E-67
(AF053008) gag-pol polyprotein [Glycine max]	2,00E-11
retrovirus-related pol polyprotein from transposon TNT 1-94-like [Oryza sativa (japonica	8,00E-10
Putative Zea mays retrotransposon Opie-2 [Oryza sativa (japonica cultivar-group)]	8,00E-74
pol protein [Cucumis melo]	7,00E-67
(Y12432) polyprotein [Ananas comosus]	2,00E-60
(D85597) polyprotein [Oryza australiensis]	2,00E-16
pol protein [Cucumis melo]	1,00E-33
pol protein [Cucumis melo]	5,00E-70
ORF I polyprotein [petunia vein clearing virus]	2,00E-15
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	6,00E-52
(AF391808) Fourf gag/pol protein [Zea mays]	2,00E-10
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	1,00E-72
(AF082134) pol polyprotein [Zea mays]	6,00E-55
AU081502(C10364)putative retrovirus-related Pol polyprotein from transposon TNT 1-94	5,00E-34

putative gag-pol polyprotein, 3'-partial [Oryza sativa]	4,00E-21
(AF369930) pol polyprotein [Citrus x paradisi]	2,00E-44
Putative Zea mays retrotransposon Opie-2 [Oryza sativa (japonica cultivar-group)]	0.0005
Putative polyprotein [Oryza sativa]	0.0001
putative polyprotein [Oryza sativa (japonica cultivar-group)]	4,00E-25
gag-pol polyprotein - maize copia-like retrotransposon Sto-4	9,00E-51
non-LTR retrolelement reverse transcriptase-like [Arabidopsis thaliana]	0.0005
putative polyprotein [Kalanchoe top-spotting virus]	0.004
mutator-like transposase-like [Arabidopsis thaliana]	6,00E-10
retrovirus-related pol polyprotein from transposon TNT 1-94-like [Oryza sativa (japonica	0.006
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	1,00E-11
putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	0.004
(AF128395) contains similarity to retrovirus-related polyproteins and to CCHC zinc finger	6,00E-13
gag-pol polyprotein [Zea mays]	3,00E-11
(AF053008) gag-pol polyprotein [Glycine max]	2,00E-25
ORF I polyprotein [petunia vein clearing virus]	2,00E-33
(AC079179) Putative Tam3-like transposon protein [Oryza sativa]	2,00E-38
gag-pol polyprotein [Vitis vinifera]	2,00E-05
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	2,00E-11
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	2,00E-39
putative gag-pol polyprotein [Oryza sativa]	7,00E-08
copia-type reverse transcriptase-like protein [Arabidopsis thaliana]	1,00E-33
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	2,00E-61
(AY081469) putative exonuclease [Arabidopsis thaliana]	7,00E-18
(AC006248) putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	3,00E-25
(AC093180) putative polyprotein [Oryza sativa (japonica cultivar-group)]	1,00E-14
(AJ411813) gag polyprotein [Cicer arietinum]	0.005
pol protein [Cucumis melo]	2,00E-40
(AF275345) putative copia-like polyprotein [Lycopersicon esculentum]	1,00E-43
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	4,00E-19
(AF369930) pol polyprotein [Citrus x paradisi]	7,00E-10
(AF229252) reverse transcriptase [Picea glauca]	4,00E-31
(AF053008) gag-pol polyprotein [Glycine max]	8,00E-24
probable retroelement polyprotein F25P12.89 [imported] - Arabidopsis thaliana	6,00E-25
putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	5,00E-05
(Y12432) polyprotein [Ananas comosus]	3,00E-89
polyprotein [Oryza sativa (japonica cultivar-group)]	1,00E-51
(AY016208) gag/pol polyprotein [Arabidopsis thaliana]	3,00E-45
putative polyprotein [Oryza sativa (japonica cultivar-group)]	1,00E-21
(AC018460) Putative retroelement polyprotein [Arabidopsis thaliana]	1,00E-09
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	9,00E-12
(AP002538) Similar to Zea mays retrotransposon Opie-2 (T04112) [Oryza sativa (japonica	3,00E-40
(AF082134) pol polyprotein [Zea mays]	2,00E-46
(AF369930) pol polyprotein [Citrus x paradisi]	8,00E-12
(AC006528) putative retroelement pol polyprotein [Arabidopsis thaliana]	2,00E-27

(AC087797) putative gag-pol polyprotein [Oryza sativa]	0.001
(AC005724) putative reverse transcriptase [Arabidopsis thaliana]	1,00E-11
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	2,00E-43
putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	8,00E-05
copia-type reverse transcriptase-like protein [Arabidopsis thaliana]	3,00E-32
pol protein [Cucumis melo]	8,00E-18
(AC006528) putative retroelement pol polyprotein [Arabidopsis thaliana]	1,00E-08
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	4,00E-41
putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	7,00E-05
copia-type polyprotein, putative [Arabidopsis thaliana]	1,00E-30
probable gag-proteinase polyprotein [imported] - Arabidopsis thaliana	0.006
(AC091238) Putative gag-pol polyprotein [Oryza sativa]	3,00E-23
pol protein [Cucumis melo]	8,00E-07
(AF090447) copia-type pol polyprotein [Zea mays]	5,00E-53
putative gag polyprotein [Cicer arietinum]	1,00E-08
putative transposon protein [Oryza sativa]	7,00E-06
(AF053008) gag-pol polyprotein [Glycine max]	3,00E-22
putative polyprotein [Arabidopsis thaliana]	8,00E-07
(AF275345) putative copia-like polyprotein [Lycopersicon esculentum]	5,00E-49
putative gag-pol polyprotein [Zea mays]	1,00E-57
(AF053008) gag-pol polyprotein [Glycine max]	5,00E-45
putative gag-pol polyprotein [Zea mays]	1,00E-57
(AF053008) gag-pol polyprotein [Glycine max]	5,00E-45
polyprotein, putative; protein id: At1g37110.1 [Arabidopsis thaliana]	3,00E-42
(NM_103266) polyprotein, putative [Arabidopsis thaliana]	2,00E-53
(AP003372) putative gag-pol polyprotein [Oryza sativa (japonica cultivar-group)]	6,00E-41
copia-type polyprotein, putative; protein id: At1g48710.1 [Arabidopsis thaliana]	3,00E-56
copia-type polyprotein, putative; protein id: At1g32590.1 [Arabidopsis thaliana]	8,00E-38
copia-type polyprotein, putative; protein id: At1g48710.1 [Arabidopsis thaliana]	6,00E-34
Enzymatic polyprotein [Contains: Aspartic protease ; Endonuclease; Reverse transcriptase]	2,00E-28
(NM_104561) polyprotein, putative; protein id: At1g57640.1 [Arabidopsis thaliana]	3,00E-14
(AJ411800) putative polyprotein [Cicer arietinum]	2,00E-06
(NM_129721) putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	2,00E-38
(NM_127220) putative non-LTR retroelement reverse transcriptase [Arabidopsis thaliana]	1,00E-09
(AC079179) Putative Tam3-like transposon protein [Oryza sativa]	5,00E-18

Tabela 31: Lista de possíveis genes identificados entre as seqüências de DNA genômico de *Eucalyptus* utilizando o *pipeline Gene Projects* do Laboratório de Genômica e Expressão, Universidade Estadual de Campinas (Unicamp).

Clone	Gene	escore
038-F05-DF.F	gi|28261781|ref|NP_783294.1| ycf2 protein [Atropa belladonna] >g...	e-105
038-E04-DF.F	gi|28261781|ref|NP_783294.1| ycf2 protein [Atropa belladonna] >g...	2,00E-91

093-F08-DF.F	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	3,00E-90
096-G07-DF.F	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	4,00E-89
109-D05-DF.F	gi|28261781|ref|NP_783294.1| ycf2 protein [Atropa belladonna] >g...	2,00E-86
018-A07-DF.R	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	3,00E-85
112-F06-DF.R	gi|7525076|ref|NP_051101.1| ycf2 [Arabidopsis thaliana] >gi|6686...	3,00E-80
104-G11-DF.F	gi|7525076|ref|NP_051101.1| ycf2 [Arabidopsis thaliana] >gi|6686...	3,00E-80
096-F06-DF.F	gi|7525076|ref|NP_051101.1| ycf2 [Arabidopsis thaliana] >gi|6686...	2,00E-78
095-C12-DF.F	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	1,00E-77
104-G11-DF.R	gi|28261781|ref|NP_783294.1| ycf2 protein [Atropa belladonna] >g...	9,00E-77
027-F06-DF.F	gi|7525076|ref|NP_051101.1| ycf2 [Arabidopsis thaliana] >gi|6686...	2,00E-74
050-A03-DF.F	gi|28261781|ref|NP_783294.1| ycf2 protein [Atropa belladonna] >g...	1,00E-72
112-E01-DF.R	gi|7525076|ref|NP_051101.1| ycf2 [Arabidopsis thaliana] >gi|6686...	5,00E-71
093-H12-DF.R	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	7,00E-70
098-C11-DF.F	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	2,00E-69
095-E02-DF.F	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	2,00E-67
092-C03-DF.R	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	9,00E-67
093-F06-DF.F	gi|7525076|ref|NP_051101.1| ycf2 [Arabidopsis thaliana] >gi|6686...	4,00E-65
104-F09-DF.R	gi|13518377|ref|NP_084736.1| (NC_002693) Ycf2 protein [Oenothera...	7,00E-65
098-C11-DF.R	gi|5881742|dbj|BAA84433.1| ycf1 [Arabidopsis thaliana] >gi|75250...	8,00E-61
102-D11-DF.R	gi|7525093|ref|NP_051117.1| ycf1 [Arabidopsis thaliana] >gi|5881...	2,00E-59
102-D11-DF.R	gi|7525076|ref|NP_051101.1| ycf2 [Arabidopsis thaliana] >gi|6686...	6,00E-58
096-G12-DF.F	gi|2924274|emb|CAA77427.1| Ycf2 protein [Nicotiana tabacum] >gi|...	7,00E-58
102-D11-DF.F	gi|2764732|emb|CAA05498.1| ndhB [Arabidopsis thaliana]	e-118
029-D01-DF.F	gi|6723806|emb|CAB67217.1| NADH-plastoquinone oxidoreductase sub...	6,00E-99
015-D02-DF.F	gi|28261774|ref|NP_783287.1| NADH dehydrogenase ND1 subunit [Atr...	9,00E-67
011-E11-DF.F	gi|20143878|sp|P58419|NU4C_OENHO NAD(P)H-quinone oxidoreductase	3,00E-65
084-E12-DF.R	gi|282888|pir||S26870 NADH2 dehydrogenase (ubiquinone) (EC 1.6.5...	4,00E-64
052-B02-DF.F	gi|20143878|sp|P58419|NU4C_OENHO NAD(P)H-quinone oxidoreductase	9,00E-64
045-C04-DF.F	gi|1695963|gb|AAB37149.1| NADH dehydrogenase [Mentha x rotundifo...	2,00E-63
087-F09-DF.F	gi|225268|prf||1211235DD NADH dehydrogenase 2-like ORF 180	1,00E-60
093-F06-DF.R	gi|7525033|ref|NP_051059.1| photosystem I P700 apoprotein A1 [Ar...	e-102
107-A01-DF.F	gi|225197|prf||1211235AB photosystem I P700 apoprotein A2 >gi|11...	e-100
004-E07-SP.F	gi|7525033|ref|NP_051059.1| photosystem I P700 apoprotein A1 [Ar...	3,00E-88
102-C09-DF.R	gi|2119840|pir||S60184 photosystem I protein A2 - garden snapdra...	7,00E-85
082-A08-DF.F	gi|6723761|emb|CAB67170.1| cytochrome f [Oenothera elata subsp. ...	3,00E-78
092-C03-DF.F	gi|13518392|ref|NP_084751.1| cytochrome c biogenesis protein [Oe...	3,00E-65
101-B04-DF.F	gi|2119840|pir||S60184 photosystem I protein A2 - garden snapdra...	4,00E-64
102-C08-DF.R	gi|7525033|ref|NP_051059.1| photosystem I P700 apoprotein A1 [Ar...	1,00E-61
112-F06-DF.F	gi|225115|prf||1209190A photosystem II protein D2	9,00E-61
106-E01-DF.F	gi|23821995|sp|Q9BBT0|PSBD_LOTJA Photosystem II D2 protein (Phot...	2,00E-59
097-F08-DF.F	gi|13358970|dbj|BAB33187.1| PSI P700 apoprotein A2 [Lotus japoni...	2,00E-59
112-C06-DF.R	gi|3015520|gb|AAC98470.1| cytochrome oxidase subunit 1 [Betula p...	3,00E-58
080-E02-DF.F	gi|7430671|pir||T00934 probable cytochrome P450 [imported] - Ara...	1,00E-57
038-D06-DF.F	gi|3850940|gb|AAC72162.1| ATP synthase beta subunit [Triunia mon...	7,00E-88
086-E02-DF.F	gi|224351|prf||1102209E ORF 5 >gi|552956|gb|AAA84683.1| ATPase s...	1,00E-71

104-H03-DF.F	gi|738937|prf||2001457E GTP-binding protein >gi|303744|dbj|BAA02...			2,00E-64
106-A11-DF.F	gi|18407327|ref|NP_566099.1| (NM_130301) putative ATP-dependent ...			4,00E-62
021-D07-DF.R	gi|10717339|gb|AAD34165.2|AF153048_1 ATPase subunit 6 [Glycine max]			4,00E-61
105-E10-DF.F	gi|738937|prf||2001457E GTP-binding protein >gi|303744|dbj|BAA02...			1,00E-59
094-H09-DF.F	gi|6624215|dbj|BAA88492.1| ABC protein [Pseudomonas fluorescens]			2,00E-72
036-B03-DF.F	gi|12437|emb|CAA29000.1| alternate petB gene product [Zea mays]			e-101
104-D09-DF.F	gi|22796414|emb|CAC87704.1| thylakoid structural protein [Liquid...			3,00E-86
105-A01-DF.F	gi|12191|emb|CAA39759.1| petA [Pisum sativum]			9,00E-63
002-B11-SP.F	gi|225210|prf||1211235AQ	ORF	512	1,00E-63
074-F09-DF.F	gi|7636118|emb|CAB88738.1| acetyl-coA carboxylase beta subunit [...			5,00E-85
075-F09-DF.F	gi|2924258|emb|CAA77410.1| RNA polymerase beta" subunit [Nicoti...			e-121
065-A10-DF.F	gi|2924258|emb|CAA77410.1| RNA polymerase beta" subunit [Nicoti...			e-115
067-A10-DF.F	gi|133421|sp|P11703|RPOB_SPIOL DNA-directed RNA polymerase beta ...			e-106
026-B06-DF.F	gi|28261709|ref|NP_783224.1| RNA polymerase beta subunit [Atropa...			e-100
111-B11-DF.F	gi|11822|emb|CAA77346.1| RNA polymerase beta subunit [Nicotiana ...			3,00E-99
105-A01-DF.R	gi|133421|sp|P11703|RPOB_SPIOL DNA-directed RNA polymerase beta ...			2,00E-95
100-G06-DF.F	gi|28261749|ref|NP_783263.1| RNA polymerase alpha subunit [Atrop...			3,00E-94
015-C02-DF.F	gi|2924258|emb|CAA77410.1| RNA polymerase beta" subunit [Nicoti...			1,00E-93
112-B07-DF.R	gi|133421|sp|P11703|RPOB_SPIOL DNA-directed RNA polymerase beta ...			6,00E-86
102-A10-DF.R	gi|13518366|ref|NP_084725.1| RNA polymerase alpha chain [Oenothe...			5,00E-83
051-C11-DF.F	gi|2924258|emb|CAA77410.1| RNA polymerase beta" subunit [Nicoti...			6,00E-79
011-A02-DF.F	gi|13518329|ref|NP_084688.1| RNA polymerase beta" chain [Oenoth...			5,00E-72
120-D02-DF.F	gi|13518329|ref|NP_084688.1| RNA polymerase beta" chain [Oenoth...			3,00E-61
121-D02-DF.F	gi|7459622|pir||T01920 probable RNA-directed RNA polymerase (EC ...			2,00E-58
090-B01-DF.F	gi|225282|prf||1211235T RNA polymerase beta			7,00E-56
066-A10-DF.R	gi|6851008|emb|CAA69754.3| ribosomal protein L16 [Arabidopsis th...			2,00E-71
015-E02-DF.F	gi|7440878|pir||T09650 ribosomal protein L16 - Vigna unguiculata...			4,00E-64
013-B10-DF.F	gi|6723832|emb|CAB67244.1| ribosomal protein L2 [Oenothera elata...			2,00E-63
098-E03-DF.F	gi|5881732|dbj|BAA84423.1| ribosomal protein S3 [Arabidopsis tha...			5,00E-56
111-B09-DF.F	gi|2995405|emb|CAA73042.1| polyprotein [Ananas comosus] >gi|7489...			3,00E-89
118-A12-DF.F	gi|16648945|gb|AAL24324.1| mutator-like transposase-like protein...			2,00E-84
038-E07-DF.F	gi|28269414|gb|AAO37957.1| putative gag-pol polyprotein [Oryza s...			9,00E-75
118-A10-DF.F	gi|16648945|gb|AAL24324.1| mutator-like transposase-like protein...			1,00E-74
068-C07-DF.F	gi|24418044|gb|AAN60494.1| Putative Zea mays retrotransposon Opi...			8,00E-74
026-G03-DF.F	gi|28269414|gb|AAO37957.1| putative gag-pol polyprotein [Oryza s...			1,00E-72
102-G01-DF.R	gi|23928444|gb|AAN40030.1| putative gag-pol polyprotein [Zea mays]			2,00E-70
108-C09-DF.R	gi|28558781|gb|AAO45752.1| pol protein [Cucumis melo]			5,00E-70
068-G04-DF.F	gi|3779021|gb|AAC67200.1| putative retroelement pol polyprotein ...			2,00E-69
072-A07-DF.F	gi|7488299|pir||T01860 reverse transcriptase homolog T7M24.7 - A...			2,00E-68
097-H06-DF.F	gi|22213212|gb|AAM94552.1| putative polyprotein [Oryza sativa (j...			2,00E-67
105-H05-DF.F	gi|28558781|gb|AAO45752.1| pol protein [Cucumis melo]			7,00E-67
120-H07-DF.F	gi|19979587|dbj|BAB88749.1| (AB072492) integrase [Silene latifolia]			1,00E-64
092-F11-DF.R	gi|28558781|gb|AAO45752.1| pol protein [Cucumis melo]			8,00E-64
027-E01-DF.F	gi|28558781|gb|AAO45752.1| pol protein [Cucumis melo]			8,00E-64
054-H11-DF.F	gi|28558781|gb|AAO45752.1| pol protein [Cucumis melo]			2,00E-63

105-H06-DF.F	gi|2995405|emb|CAA73042.1| polyprotein [Ananas comosus] >gi|7489...	7,00E-62
057-D08-DF.F	gi|19979587|dbj|BAB88749.1| integrase [Silene latifolia]	8,00E-62
065-D01-DF.F	gi|28269414|gb|AAO37957.1| putative gag-pol polyprotein [Oryza s...	2,00E-61
084-E04-DF.F	gi|2995405|emb|CAA73042.1| polyprotein [Ananas comosus] >gi|7489...	2,00E-60
092-F11-DF.F	gi|3777527|gb|AAC64917.1| gag-pol polyprotein [Glycine max]	3,00E-60
071-E01-DF.F	gi|23928444|gb|AAN40030.1| putative gag-pol polyprotein [Zea mays]	4,00E-60
112-A07-DF.R	gi|22748369|gb|AAN05371.1| Putative retroelement [Oryza sativa (...	2,00E-59
065-C08-DF.F	gi|24418044|gb|AAN60494.1| Putative Zea mays retrotransposon Opi...	5,00E-59
113-G08-DF.F	gi|7489738|pir||T17429 gag-pol polyprotein - maize copia-like re...	6,00E-59
080-D05-DF.F	gi|28269414|gb|AAO37957.1| putative gag-pol polyprotein [Oryza s...	8,00E-58
001-H02-DF.F	gi|23928444|gb|AAN40030.1| putative gag-pol polyprotein [Zea mays]	1,00E-57
108-E03-DF.F	gi|23928444|gb|AAN40030.1| putative gag-pol polyprotein [Zea mays]	1,00E-57
086-D05-DF.F	gi|28269414|gb|AAO37957.1| putative gag-pol polyprotein [Oryza s...	3,00E-57
066-A09-DF.F	gi|14716946|emb|CAC44142.1| putative polyprotein [Cicer arietinum]	6,00E-57
084-D07-DF.R	gi|23928444|gb|AAN40030.1| putative gag-pol polyprotein [Zea mays]	7,00E-57
120-A08-DF.F	gi|15221940|ref|NP_175303.1| (NM_103766) copia-type polyprotein,...	3,00E-56
068-C08-DF.F	gi|23928439|gb|AAN40025.1| putative gag-pol polyprotein [Zea mays]	2,00E-55
018-D05-DF.R	gi|25301688|pir||H96650 protein T3P18.3 [imported] - Arabidopsis...	3,00E-62
037-F07-DF.F	gi|18767374|gb|AAL79340.1|AC099402_4 Putative 22 kDa kafirin clu...	5,00E-58
050-G05-DF.F	gi|82207|pir||A05205 hypothetical protein 1708 - common tobacco ...	3,00E-96
106-F02-DF.F	gi|6630685|dbj|BAA88531.1| unnamed protein product [Oryza sativa...	1,00E-92
107-F05-DF.F	gi|11466011|ref|NP_054553.1| hypothetical protein [Nicotiana tab...	1,00E-76
112-C07-DF.F	gi|23297654|gb|AAN13002.1| unknown protein [Arabidopsis thaliana...	7,00E-73
049-D03-DF.F	gi|15228384|ref|NP_190419.1| hypothetical protein; protein id: A...	1,00E-71
112-D06-DF.R	gi|7444414|pir||T02206 hypothetical protein - common tobacco ret...	1,00E-66
105-E01-DF.F	gi|82207|pir||A05205 hypothetical protein 1708 - common tobacco ...	1,00E-65
105-H11-DF.F	gi|20198297|gb|AAM15511.1| hypothetical protein [Arabidopsis tha...	2,00E-64
017-E03-DF.F	gi|7459476|pir||T02186 hypothetical protein At2g47010 [imported]...	7,00E-58
094-A06-DF.F	gi|15218625|ref|NP_174702.1| hypothetical protein; protein id: A...	7,00E-58
091-H09-DF.F	gi|9758461|dbj|BAB08990.1| gene_id:MUK11.18~unknown protein [Ara...	3,00E-57
051-A02-DF.F	gi|6630685|dbj|BAA88531.1| unnamed protein product [Oryza sativa...	2,00E-55
038-G02-DF.F	gi|24461860|gb|AAN62347.1|AF506028_14 CTV.20 [Poncirus trifoliata]	5,00E-62
008-G02-DF.F	gi|7485903|pir||T04609 hypothetical protein F20O9.70 - Arabidops...	6,00E-60
108-F02-DF.F	gi|9759345|dbj|BAB10000.1| dbj|BAA90805.1~gene_id:MAH20.7~strong...	3,00E-59
092-C06-DF.R	gi|9759345|dbj|BAB10000.1| dbj|BAA90805.1~gene_id:MAH20.7~strong...	3,00E-57
105-G10-DF.R	gi|23463065|gb|AAN33202.1| At5g58430/mqj2_20 [Arabidopsis thalia...	7,00E-72
090-C02-DF.F	gi|602034|emb|CAA34729.1| unidentified reading frame (AA 1-510) ...	3,00E-67
115-F08-DF.F	gi|2335097|gb|AAC02766.1| putative receptor-like protein kinase ...	8,00E-67
114-F08-DF.F	gi|15229473|ref|NP_189475.1| P-glycoprotein, putative; protein i...	2,00E-65
093-A11-DF.F	gi|11034602|dbj|BAB17126.1| putative receptor kinase [Oryza sati...	4,00E-63
113-G03-DF.F	gi|11034602|dbj|BAB17126.1| putative receptor kinase [Oryza sati...	6,00E-61
045-A01-DF.F	gi|1272349|gb|AAA97903.1| secreted glycoprotein 3	2,00E-60
028-H06-DF.F	gi|25052804|gb|AAN65180.1| mitogen-activated protein kinase 4 [P...	2,00E-60
002-C08-SP.F	gi|13540262|gb|AAK29382.1|AF312745_1 MAP kinase phosphatase [Ara...	7,00E-58
029-B03-DF.F	gi|15226381|ref|NP_178304.1| leucine-rich repeat transmembrane p...	2,00E-55

101-A11-DF.F	gi|18398836|ref|NP_565441.1| (NM_127431) putative calmodulin-bin...	6,00E-73
011-A11-DF.F	gi|9758246|dbj|BAB08745.1| contains similarity to selenium-bindi...	7,00E-59
113-D07-DF.F	gi|11908088|gb|AAG41473.1|AF326891_1 putative phi-1 protein [Ara...	9,00E-66
066-B08-DF.R	gi|21552981|gb|AAM62410.1|AF480488_1 NPR1 [Nicotiana tabacum]	9,00E-57
013-G07-DF.F	gi|9622884|gb|AAF89966.1|AF200530_1 cellulose synthase-6 [Zea mays]	1,00E-66
027-C07-DF.F	gi|6446577|gb|AAD39534.2| cellulose synthase catalytic subunit [...	1,00E-58
050-E03-DF.F	gi|15231828|ref|NP_188048.1| putative pectin methylesterase	3,00E-36
095-A08-DF.F	gi|2463509|emb|CAA70735.1| pectate lyase [Zinnia elegans]	8,00E-32
066-B08-DF.R	gi|6446577|gb|AAD39534.2| cellulose synthase catalytic subunit [Gossypium	1,00E-58
006-E06-DF.F	gi|9622886|gb|AAF89967.1|AF200531_1 (AF200531) cellulose synthase-7	6,00E-42
065-B08-DF.F	gi|9759258|dbj|BAB09693.1| cellulose synthase catalytic subunit [Arabidopsis	9,00E-35
057-C12-DF.F	gi|1345684|sp|P49317|CAT3_NICPL CATALASE ISOZYME 3	9,00E-89
095-E01-DF.F	gi|15231718|ref|NP_190864.1| peroxiredoxin - like protein; prote...	1,00E-59
110-F03-DF.F	gi|3442|emb|CAA27416.1| put. CAN1 protein (aa 1-590) [Saccharomy...	6,00E-78
082-E09-DF.F	gi|3442|emb|CAA27416.1| put. CAN1 protein (aa 1-590) [Saccharomy...	5,00E-76
096-B07-DF.F	gi|16124255|gb|AAG28706.2| single chain antibody against rice st...	1,00E-84
036-G02-DF.F	gi|100335|pir||S18181 dnaK-type molecular chaperone Nthsp70 - co...	e-114
106-G01-DF.F	gi|7437043|pir||T10101 aconitate hydratase (EC 4.2.1.3) - cucurb...	3,00E-65
092-A05-DF.R	gi|20161620|dbj|BAB90540.1| gibberellin response modulator-like ...	1,00E-57
106-H05-DF.F	gi|550319|emb|CAA57393.1| beta-fructofuranosidase [Beta vulgaris...	7,00E-67
027-C07-DF.F	gi|15231718|ref|NP_190864.1| peroxiredoxin - like protein supported by	1,00E-59
067-B04-DF.F	gi|459535|emb|CAA54967.1| orf250~homology to orf228 from Marchan...	3,00E-79
038-B10-DF.F	gi|20045|emb|CAA32025.1| ORF [Nicotiana tabacum] >gi|130582|sp|P...	9,00E-73
108-E12-DF.F	gi|225280|prf||1211235Q rpoC-like ORF 548	5,00E-71
080-A09-DF.F	gi|9279572|dbj|BAB01030.1| subtilisin proteinase-like protein [A...	4,00E-60
068-G01-DF.F	gi|7435774|pir||S22502 cysteine proteinase (EC 3.4.22.-) - kidney bean	1,00E-41
102-D07-DF.F	gi|23477203|emb|CAD36200.1| TIR-NBS disease resistance protein [Populus	8,00E-44
067-C04-DF.F	gi|13359451|dbj|BAB33421.1| putative senescence-associated prote...	6,00E-61
059-G02-DF.F	gi|13124313|sp|Q9SWE5|HL3A_ARATH Halotolerance protein Hal3a	2,00E-42

Tabela 35: Lista de pares de *primers* desenvolvidos a partir de seqüências genômicas e de ESTs de *Eucalyptus*.

EMBRA	SSR	*Primer* F/ *Primer* R	FRAGMENTO (pb)	Tm (ºC)	Localizaç
604	(GAAA)16	TGGGTGAGGCTGTGATGAT	260	60	Genômico
		TGGAGGTTCTTTTATTTCGGTG			
605	(AG)49	TGGTTTTAGGCAACACGAGA	200	60	Genômico
		ATCCTCACATGTGCACCTCA			
609	(GAA)8	CAAATCTAGGCGAGGACTGA	280	60	Genômico
		GGAACAATTTCAAGAATCAAGCC			
610	(AAT)15	ATCTTCTTCACCATGAGCGG	170	60	Genômico
		CTCACTAGCACTGGTCCACCT			
615	(TAA)12	AAAGGTTATGCCCAAATGGA	270	60	Genômico
		TTTTCAACCCATTGCCAAAC			
616	(CCG)10	CTTCTCCTTCTCCTTCGCCT	110	60	Genômico

		GTTCGTGTTCGTGTCGATGT			
EG Aldolase	(CT)21	AGGAGCCAAGAAGGGAATGT	230	60	EST
		ACGGAGATTCGACAGATTGG			
EG Ripening	(CT)12	CATGAGCACCACGAGAAGAA	180	60	EST
		TGCTCGTGCGTACTTGGTTA			
618	(CT)15	CCGAAGATGGAGAAGGACAC	160	56	EST
		TGCTCGTGAAGAACATCAGC			
624	(TCTG)9	CTCGGTTAGCGTTACGAAGG	140	56	EST
		AAAAGAAGCAACCCTGTGCT			
627	(CT)18	TTCTCTTCATTCATCGCTTCC	240	56	EST
		GGATTCCCGAAGAGGAACTC			
629	(CT)16	GCCCCAACGTGGTAAGTCTA	180	56	EST
		CCCAGAAATCGAGCCAACT			
631	(AGG)9	GCGCGAGAGAGAGAGAGAGA	180	60	EST
		GGGCCTTGTACTCCTCCTTC			
632	(CTT)10(CT)15	AAGAAAGGCTGGTCCAAGGT	240	58	EST
		CCCCCAAAACAAGAACACAA			
640	(ATA)11	AATATGTTCGAAAGCGCCAT	200	56	Genômico
		GACCCGTTCTCTTTCCTTCTA			
641	(TC)14	TCCCCTCATCGTTGAGATTC	270	56	Genômico
		CCATTGCCTAGCCTAGATGC			
644	(TC)21	GCCACTCACTTTCCCTCACT	230	56	Genômico
		GGTGGTCATACGACAGAGCA			
645	(AG)11	CAAAATGGTCCATGGTGTTG	220	56	Genômico
		TCGCTGTAGCATTGAGCTTG			
646	(GA)17	AAAGCGTTACGTGCGACTCT	160	58	Genômico
		GGTACAGAAGAGGGCGTCAA			
648	(TC)13	CACCTCTCTACCCGTTTCCA	160	56	Genômico
		CCAAATCCTCTTCGATTCCA			
650	(TAA)12	AAAGGTTATGCCCAAATGGA	270	56	Genômico
		TTTTCAACCCATTGCCAAAC			
651	(AG)17	CGAGCTCCAAAACCATAAA	120	56	Genômico
		GCATTTTCCTCCCTCATTTG			
652	(CT)15	TTCTTCACATCTCCCCTTCC	160	58	Genômico
		TGAGGCGAAAGATCTGGACT			
653	(CT)12	TTTCCCCGAAGCAGAAACTA	260	60	Genômico
		AGAAGCAGGTGCAGAGGTTG			
654	(AG)14	GGGGCAAAATACAATCCAAA	250	56	Genômico
		CAAAATTGGCAAAATCACGA			
655	(TA)18	ATGTGGCTAAACCGCAAAAC	150	56	Genômico
		CAAATAGGTCGTCGATTTGTCA			
658	(CT)10	AAGGTCAGAGGCTGACGGTA	150	56	Genômico
		CACTTATCAATCCGGGCGTA			
659	(AT)9	CATGCCCAAAGTGGATATATGTT	250	56	Genômico
		CGGTTGCTAACCATCCAAGT			
661	(AT)15	GCAGAGGTCATCAATCATCG	240	58	Genômico
		GATCTCTGCAATGTCCGGTT			
662	(AG)16	TCTCCTCCTGTGTTGCTCCT	270	56	Genômico
		ATTTGCGGATTCTTTAGCGA			

665	(GA)13	GACGATTTGGGCAGAACAAT	140	56	Genômico
		TTGCAACACATTCCCACATT			
670	(TG)14	GCATCCCAATAGGCAGATCA	240	56	Genômico
		TGGCAAGACGACTTTGAGTG			
676	(AG)9	CGCAGGAAATGGGTAGAGAG	270	56	Genômico
		GGAATTTGGTGGTGAGGAGA			
680	(GA)15	TACTTCCATTGGCAGCATGA	240	60	Genômico
		AGTCCTCACATGTCTTGCCC			
681	(AG)13	GAGTTCATCGCCGAAGAGAG	210	60	Genômico
		TGGTTGACAAAGAACAGCCA			
682	(CT)13	TTGTTGCCCCTAAAGCAAAC	220	56	Genômico
		CCTAGCGTCAAGTAGGCAAAA			
687	(CT)9	GGTTTGGTTTGGTTCGGTTA	210	56	Genômico
		GCAAATAAATGGAGGTTGCG			
692	(AAT)14	CCTGTGGATCCTTGATCACC	250	56	Genômico
		GGAAGCACCCGAAGTTTTTA			
693	(CT)9	GCCGTTAACTGCACCACTTT	160	56	Genômico
		TTCCACGGATCTCATAGCTG			
695	(GA)21	TCGAGGGAGAATGGAGAAAA	160	56	Genômico
		TCTGGCTGTTAAGTATAAGCAGTGA			
706	(GA)14	CAATCCTTCCCCTAACCTCA	230	56	Genômico
		GCAGTGGCTTGTCCTGTCTT			
710	(AT)20	CGGAATACATGGGAAGCAAT	210	56	Genômico
		TCCCATGAATCCATCCTGTT			
711	(CT)10	CCTCTTCCCTCCCTTCTTTG	160	56	Genômico
		GTCGAAAGAAGCAAGCAGGT			
727	(AT)12	CAGATAAATCCAGGGGCTCA	160	56	Genômico
		ACGCTCATGATTTTAACGCC			
729	(AAG)6	ACCCCATAAGTGGGATTCAA	260	58	Genômico
		GACGGCGTGCTTAAATTACC			
731	(GGA)7	ACTGTAAATGATGCATGCGG	250	58	Genômico
		ATATGCTGCAGCCAACACAG			
746	(TAT)8	GCCAGTAGTGTTTTCCTCGG	180	56	Genômico
		TTGCCCTCCTCATGGTATTC			
747	(AG)10	GGACACTTCTGAGTCGAAGGA	260	56	Genômico
		CTCCCTCACGGTTATGGAAA			
749	(AAG)6	CGAAAATGAAAGCCTACCCA	270	56	Genômico
		CGTGAAGTAGCAGGCAATCA			
753	(ATAC)7	TGGGAGCTAACGAAAAGAAAA	210	56	Genômico
		AGTAACCCAGCCTACCCCAT			
762	(AT)17	ATGTGGCAGCATAATGCAAA	170	56	Genômico
		GGAGTCCAACAAATGGAAGC			
764	(GA)18	ATTCGGCCAAAACAACAGAG	140	56	Genômico
		CGCAAATGTGTTAGCTGTCAA			
765	(AG)14	TGTCGGATCGTCCTTCTTCT	170	56	Genômico
		TCGTCAGCATGGTGTAGAGC			