SUZANA NEIVA SANTOS

Genes de Lignificação em *Eucalyptus*: estrutura e diversidade genética dos genes *4cl* e *ccoaomt*.

Orientador: Dario Grattapaglia

Brasília
2005

SUZANA NEIVA SANTOS

Genes de Lignificação em *Eucalyptus*: estrutura e diversidade genética dos genes *4cl* e *ccoaomt*

"Dissertação apresentada ao Programa de Pós-Graduação "Stricto Sensu" em Ciências Genômicas e Biotecnologia da Universidade Católica de Brasília, como requisito para a obtenção do Título de Mestre em Ciências Genômicas e Biotecnologia."

Orientador: Dario Grattapaglia

Brasília
2005

S237g Santos, Suzana Neiva.
Genes de lignificação em Eucalyptus: estrutura e diversidade genética dos genes 4cl e ccoaomt / Suzana Neiva Santos ; orientador Dario Grattapaglia – 2005.
xix, 208 f. : il. ; 30 cm.

Dissertação (mestrado) – Universidade Católica de Brasília, 2005.

1. Eucalipto - genética. 2. Genes. I. Grattapaglia, Dari, orient. II. Título.

CDU 582.883.4:575

Ficha elaborada pela Divisão de Processamento do Acervo do SIBI – UCB.

TERMO DE APROVAÇÃO

Dissertação defendida e aprovada como requisito parcial para a obtenção do Título de Mestre em Ciências Genômicas e Biotecnologia, defendida e aprovada, em 26 de agosto de 2005, pela banca examinadora constituída por:

_____

Dr. Alexandre Siqueira Guedes Coelho
Universidade Federal de Goiás
Examinador Externo

_____

Dr. Rinaldo Wellerson Pereira
Universidade Católica de Brasília
Examinador Interno

_____

Dr. Georgios Joannis Pappas Júnior
Universidade Católica de Brasília
Examinador Interno

_____

Dario Grattapaglia
Universidade Católica de Brasília
Orientador

Dedico à minha mãe, Antonia Neiva, e ao meu pai, Cloves Santos, por tudo que são e representam para mim.

Ofereço ao meu irmão Tiago e à “Minha Linda”, Laura Neiva, como uma forma de incentivo para o crescimento pessoal e profissional.

AGRADECIMENTOS

À Deus, minha força e fortaleza, pela inspiração com que me conduziu.

Aos meus pais, incentivadores, guias e mestres, e ao meu irmão Tiago, pelo apoio incondicional e pelo incentivo durante todo o curso.

À todos os meus familiares, inclusive aqueles que estão longe, que me deram força e demonstraram carinho e orgulho.

Ao Professor Dr. Dario Grattapaglia pela idealização do trabalho e orientação.

Ao Professor Dr. Georgios J. Pappas Júnior pelo auxilio e preocupação nas análises de bioinformática.

Ao Professor Dr. Rinaldo Wallace, pelo auxílio nas análises estatísticas, pela colaboração técnica, pela orientação efetiva, pela amizade e estímulo, meu sincero obrigada.

Aos professores do Programa de Pós-graduação em Ciências Genômicas e Biotecnologia da Universidade Católica de Brasília, pelo aprimoramento da minha formação acadêmica.

Aos técnicos e colegas do Laboratório de Biotecnologia da Universidade Católica de Brasília, Alessandra Reis, André Luiz Ramos, Idacuy Mundim, Márcia Araújo e Willian Baião Reis pela agradável convivência e auxílio.

A todos os funcionários da UCB, em especial a Francisco Fabio Gomes da Costa e Danielle Cordeiro, pela presteza.

Às amigas Luciana Retori, Sandra Elisa, Luciana Tagliaferro, Andréa Tagliaferro e em especial à Letícia Fagundes (Tilda!!!), pelas alegrias, carinho, companheirismo, convivência e pelas ProTeHomicas.

Aos amigos que conquistei no mestrado, Guilherme Jacques, Elisa Pavin, Jorge Castro, Alexandre Póvoa, Patrícia Pellegrini, Suzanila Santos, Camila Chaves, Joaquim, Juliana Nardelli, pelos sufocos que passamos nas matérias e pela recompensa nas reuniões de fim de ano.

Ao amigo inesquecível Alexandre Peixoto Figueira pela paciência, auxílio, presença e pelos momentos de descontração e carinho.

Aos colegas do laboratório de Genética Vegetal da Embrapa - Cenargen, Eva, Nathalia, Rodrigo Tristan, Alexandre Missiaggia e Marilia Pappas, pela ajuda nos trabalhos de bancada.

Aos colegas do laboratório 3 da Embrapa - Cenargen, Eduardo Romano (Dudu), Karina Proite, Ana Carolina Vilarinho e Maria Laine Tinoco pelo incentivo e apoio.

À Dell Computadores e ao CNPq, pela bolsa de estudos e custeio do mestrado na UCB.

À todos que contribuíram de alguma forma para o êxito deste trabalho.

Deus nos dá as nozes. Mas não as quebra. (Provérbio Alemão)

"Por que cometer erros antigos se há tantos erros novos a escolher?"
Bertrand Russel, filósofo inglês

"Se podes olhar, vê. Se podes ver, repara."
Jose Saramago

SUMÁRIO

Lista de Figuras e Tabelas

Página

Lista de Abreviações

%	Porcentagem
°C	Graus Celsius
4CL	4-Comarato-CoA Ligase
AldOMT	5-Hidroxiconiferaldeído *o*-Metiltransferase
bp	Par de base
BSA	Albumina de Soro Bovino
C3H	4-Coumarato 3-Hidroxilase
C4H	Cinamato 4-Hidroxilase
CAD	Cinamil Álcool Dehidrogenase
CCoAOMT	Caffeoil CoA *o*-Methiltransferase
CCR	Cinamoil-CoA Redutase
COMT	Ácido Cafeico *o*-Metiltransferase;
CTAB	*Cetyltrimethylammonium Bromide*
CAP3	*Sequence Assembly Program*
DNA	Ácido Desoxirribonucléico
dNTP	Deoxirribonucleosídeo Trifosfato
EDTA	Ácido etileno Diamono Tetracético
EST	*Expressed Sequence Tag*
F5H	ferrulato 5-Hidroxilase;
HCT	*p*-Hidroxicinamoil-CoA:
Kb	Kilobase
min	Minuto
ml	Mililitro
mM	Milimolar
µl	Microlitro
PAL	Fenilalanina amônia Liase;
PCR	*Polimerase Chain Reaction*
QTL	*Quantitative Trait Loci*
RNAse	Ribonuclease
rpm	Rotações por minuto
SAD	Sinapil Álcool dehidrogenase
SAM	*S*-Adenosil-*L*-Metionina
SNP	*Single Nucleotide Polymorphism*
Taq	*Thermus aquaticus*
TE	Tris-EDTA

RESUMO

O presente trabalho teve como objetivo o estudo da diversidade nucleotídica em dois genes chave na via de biossíntese de lignina. Os genes *4cl* e *ccoaomt* foram estudados em amostras de populações naturais de três espécies comerciais de eucalipto, *E. grandis*, *E. globulus* e *E. urophylla.* A mineração de um banco de dados de ESTs gerado no projeto Genolyptus revelou a presença de diferentes isoformas destes genes e uma riqueza de seqüências suficientes para detecção de SNPs. Para o gene *4cl*, as tentativas de resseqüenciamento sugeriram a existência de várias cópias do gene no genoma do eucalipto indicando a necessidade de clonagem prévia de amplicons para futuros estudos de diversidade de seqüência. A análise de um trecho de 440 pb do gene *ccoaomt* revelou uma freqüência de 1 SNP a cada 55 pb, 63 pb e 220 pb respectivamente para *E. grandis*, *E. urophylla* e *E. globulus*. *E. grandis* (π = 0,00356) apresentou o dobro de diversidade nucleotídica do que *E. globulus* (π = 0,00168) e cerca de 1,4 vezes mais que *E. urophylla* (π = 0,00254). Observa-se ainda que *E. grandis*, a espécie com a maior distribuição geográfica e portanto maior oportunidade de fluxo gênico, apresentou, de fato, valores mais altos de diversidade nucleotídica e haplotípica. SNPs fixados ou quase fixados bem como indel privativos foram observados em *E. urophylla*, a espécie disjunta que ocorre nas ilhas ao norte da Austrália. Foram detectados polimorfismos não sinônimos, potenciais alvos interessantes para estudos de variabilidade na atividade da enzima. A análise da extensão do desequilíbrio de ligação, embora limitada a apenas um gene e com base em poucos sítios polimórficos, sugere que dentro de um gene e a distâncias menores do que ~250 pb, SNPs tendem a se encontrar em forte desequilíbrio de ligação. O seqüenciamento e montagem de um clone BAC resultaram na obtenção da seqüência completa da região codante do gene *4cl* com 5.203 pb. A obtenção desta seqüência com o início de transcrição, inédita para *Eucalyptus*, abre possibilidades

interessantes de estudos detalhados da diversidade nucleotídica e padrões de DL ao longo deste gene em populações de clones fenotipados, no sentido de buscar associações entre haplótipos específicos e variação quantitativa em propriedades químicas da madeira.

PALAVRAS-CHAVE: *Eucalyptus*, diversidade nucleotídica, CCoAOMT, 4CL, BAC.

ABSTRACT

The objective of this work was to study the nucleotide diversity in two key genes in the pathway of lignin biosynthesis. The genes *4cl* and *ccoaomt* were studied in samples of natural populations of three commercial species of eucalypt, *E. grandis, E. globulus* and *E. urophylla*. The mining of the Genolyptus EST database disclosed the presence of different isoforms of these genes and a variety of sequences for SNP detection. For the gene *4cl*, the ressequencing attempts had suggested the existence of multiple copies of the gene in the eucalypt genome indicating the necessity of further cloning of amplicons for studies of sequence diversity. The analysis of a fragment of 440 pb of the gene *ccoaomt* disclosed a frequency of 1 SNP each 55 pb, 63 pb and 220 pb respectively for *E. grandis*, *E. urophylla* and *E. globulus*. *E. grandis* (π = 0,00356) presented the double of nucleotide diversity of *E. globulus* (π = 0,00168) and about 1,4 times more diversity than *E. urophylla* (π = 0,00254). Furthermore, we observed that *E. grandis*, the species with the greatest geographic distribution presented higher values of nucleotide and haplotype diversity. Fixed SNPs or almost fixed as well as indel privative were observed in *E. urophylla*, the disjoint species that occurs in the islands at the north of Australia. Nonsynonymous polymorphisms were detected and are potential targets for enzyme activity studies. The analysis of the extension of the linkage disequilibrium, although limited to few polymorphic sites along the sequenced region, suggested that at distances smaller than 250 pb, SNPs tends to be in strong linkage disequilibrium. The sequencing and assembly of one BAC clone resulted in a complete sequence of the 4cl gene with 5203 pb. The sequence of the translation of this gene, previously unknown for *Eucalyptus*, opens interesting possibilities directed to a more general characterization of the nucleotide diversity and DL extension around this gene in populations to the detection of associations between specific haplotypes and quantitative variation in

chemical properties of the wood.

KEYWORDS: *Eucalyptus*, nucleotide diversity, CCoAOMT, 4CL, BAC

1.Introdução

1.1 A história do eucalipto no Brasil

O plantio de florestas de eucalipto constitui uma das melhores alternativas para atender às diversas demandas da sociedade no que diz respeito ao consumo de produtos de base florestal, seja na área de papel, celulose e derivados, seja na área de siderurgia. Entretanto, esse consumo vem atrelado às exigências de sustentabilidade da produção de biomassa florestal.

O Brasil começou a se destacar como grande plantador mundial de *Eucalyptus* a partir de 1910. Inicialmente, o eucalipto foi plantado com a finalidade de ornamentação ou para servir de quebra-ventos, pelo seu extraordinário porte. Após estudos mais profundos, o silvicultor Edmundo Navarro de Andrade, com sementes trazidas de Portugal da espécie *Eucalyptus globulus* inicialmente e mais tarde de *E. citriodora* e *E. tereticornis*, introduziu a plantação econômica do eucalipto. Comparado às espécies nativas do Brasil, o eucalipto foi o que mais se destacou, sendo então escolhido para a produção de lenha para as locomotivas da Companhia Paulista de Estradas de Ferro, hoje conhecida como Ferrovia Paulista S.A.

Em 1966, a área plantada constava de cerca de 600 a 700 mil hectares. Após a implementação da lei 5106, a qual define programas de incentivos fiscais ao reflorestamento, a área plantada do eucalipto aumentou em mais de 3 vezes até o ano de 1992 (LIMA, 1993).

Atualmente, o Brasil está entre os maiores plantadores mundiais, com pouco mais de 3 milhões de hectares, detendo o maior índice médio de produtividade, $40m^3$ por hectare/ano (Revista Madeira, 2003). Por ser uma árvore de crescimento rápido – com idade média de corte de 6 a 8 anos – quando comparada ao ciclo das árvores de matas nativas que levam em média 25 anos, além de fácil adaptação às mais diferentes condições de clima e solo, o eucalipto passou a ser uma alternativa racional contra a devastação das florestas nativas em diversas regiões do planeta, propiciando a preservação do meio ambiente (CEIMA, 2002).

O gênero *Eucalyptus* é conhecido por sua grande variabilidade genética englobando

mais de 600 espécies descritas. Esta decorre, entre outros, do seu hábito alógamo, da resposta à pressão de seleção causada pelas alterações do meio ambiente e do próprio processo de deriva e especiação. O mosaico formado pela distribuição das espécies reflete diferentes adaptações a uma grande variação de clima e solo. São centenas de espécies com propriedades físicas e químicas e características mecânicas e estéticas tão diversas que fazem com que os eucaliptos sejam usados para as mais diversas finalidades, dispensando o uso das várias espécies latifoliadas nativas (PEREIRA *et al.*, 2000).

Além de matéria prima para a produção de celulose, papel, chapas de fibra, aglomerados, madeira serrada, casas, estruturas e móveis, outros benefícios e produtos adicionais podem ser citados tais como a proteção de solos contra erosões, a capacidade de captação de CO_2, a geração de energia (lenha e carvão), o tanino (curtimento de couro), tecidos sintéticos, cápsulas de remédios, óleos essenciais e mel (pólen). Destarte, o *Eucalyptus* pode ser uma fonte de riqueza econômica e social, gerando empregos e mantendo o homem no campo (REVISTA MADEIRA, 1997).

Entretanto, mesmo diante das inúmeras possibilidades de aproveitamento do eucalipto e das riquezas disponíveis, poucas espécies têm sido plantadas em escala comercial. Isto porque faz-se necessário a produção de madeira de alta qualidade e para tanto, duas estratégias podem ser empregadas. A primeira, mais utilizada até hoje, consiste em melhorar geneticamente a qualidade da madeira das espécies mais plantadas, como *Eucalyptus grandis* e *Eucalyptus urophylla*. A segunda alternativa é a identificação de espécies produtoras de madeira de características satisfatórias para o uso que se pretende, com programas posteriores destinados a aumentar a produtividade (CEIMA, 2002).

As espécies utilizadas em reflorestamento apresentam, entre outras vantagens, alta produtividade, redução da idade de corte, segurança de abastecimento, homogeneidade de matéria-prima, custo competitivo, produção regionalizada, além da possibilidade de múltiplos

usos da floresta e seus produtos. Existe unanimidade entre os pesquisadores da área de produtos florestais que a qualidade da madeira para determinados usos pode ser melhorada, modificada ou ter alguns fatores minimizados ou controlados, em considerável extensão. Atualmente, busca-se a interação entre os atributos desejados da matéria-prima e a qualidade do produto final, através do trabalho conjunto dos setores de produção florestal e industrial.

No Brasil, os programas de melhoramento genético tiveram início na década de 70 com o objetivo de encontrar espécies melhor adaptadas para as condições ambientais e propiciar o aumento do conjunto genético e qualidade das plantas (BRUNE & ZOBEL, 1981). Clones elite de híbridos de *E. grandis* e *E. urophylla* foram gerados e são amplamente utilizados pela indústria de papel e celulose, devido à qualidade da madeira, o rápido crescimento e os grandes volumes de madeira que são produzidos (BERTOLUCCI *et al*., 1995).

Dentre as centenas de espécies de eucalipto, as mais plantadas no mundo são: *Eucalyptus grandis, E. saligna, E. urophylla, E. camaldulensis, E. tereticornis, E. globulus, E. viminalis, E. nitens, E. deglupta, E. citriodora, E. exserta, E. paniculata* e *E. robusta*. No Brasil, as espécies mais utilizadas estão ilustradas na figura 1 (REVISTA MADEIRA, 2003).

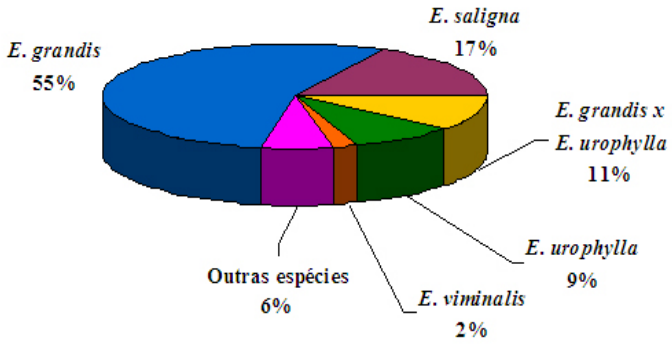


Figura 1. Proporção das espécies de *Eucalyptus* cultivadas no Brasil.

1.2 Características das espécies de *Eucalyptus*

Da família das *Myrtaceae* e com mais de 600 espécies descritas, o eucalipto é uma árvore originária da Austrália, tipicamente de clima temperado a sub-tropical mas de fácil adaptação em várias outras condições climáticas. Possui fibras delgadas, curtas, rígidas, de granulação reduzida, o número de fibras por grama é alto e as microfibrilas apresentam uma pequena angulação em torno do eixo da fibra. Todas essas características, proporcionam uma estrutura adequada para a produção volumosa de papel de boa qualidade e de alta opacidade, o que é de grande interesse para o comércio.

As árvores são perenes e apresentam aromas, dependendo da espécie. Além da importância na estrutura da paisagem, elas se destacam no quesito qualidade da madeira e na produção de óleos. Os óleos possuem um grande valor comercial, sendo utilizados na medicina, indústria e aromáticos.

O eucalipto possui um genoma de aproximadamente 630 Mb com um conjunto haplóide de 11 cromossomos. Devido ao grande interesse econômico da espécie pela indústria papeleira, estudos estão sendo realizados com o foco principal na qualidade da madeira. Desta forma, vários genes que participam na formação da madeira, como o gene *ccr*, *ccoaomt*, *cad*, *comt* entre outros, bem como genes de florescimento, são os principais genes em estudo atualmente.

1.2.1 Eucalyptus grandis

Sua área de ocorrência natural estende-se em forma descontínua e fragmentada por uma longa faixa costeira do continente Australiano, desde Newcastle até Atherton (Figura 2). O clima em toda a área varia de temperado-quente a subtropical-moderado com invernos suaves e chuvas abundantes e bem distribuídas. É sem dúvida uma espécie que possui qualidades excelentes, superando qualquer outra em incremento, quando em condições

ambientais adequadas. A madeira de *E. grandis* é leve e fácil de ser trabalhada. Utilizada intensivamente, na Austrália e na República Sul Africana, como madeira de construção, quando oriunda de plantações de ciclo longo. A madeira produzida em ciclos curtos é utilizada para caixotaria. Plantações, convenientemente manejadas, podem produzir madeira excelente para serraria e laminação. É a principal fonte de matéria prima para celulose e papel do Estado de São Paulo. É susceptível ao cancro do eucalipto (*Cryphonectria cubensis* Bruner) e é a espécie mais plantada fora da Austrália (IPEF, 2004).

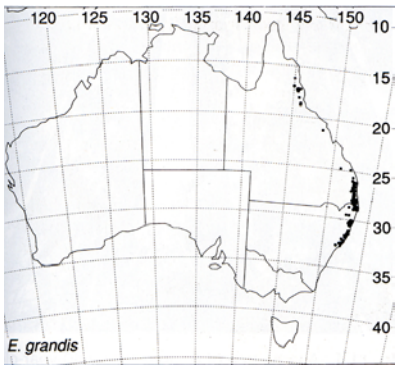


Figura 2. Distribuição da espécie *E. grandis* pelo continente Australiano. Adaptado de Eldridge *et al*., 1994.

1.2.2 Eucalyptus globulus

Eucalyptus globulus é um das árvores nativas mais amplamente cultivada na Austrália. Ela pode ser encontrada em parques e jardins por toda a Austrália e é também bem estabelecida em várias outras partes do mundo, incluindo o Brasil. É natural da parte leste da Tasmânia em populações distribuídas ao longo da costa (Figura 3). Também são encontrados nas montanhas do Estados de Victoria e Nova Gales do Sul.

É considerada uma árvore de médio à grande porte a qual pode alcançar até 70 metros de altura, com média variando de 15 a 25 metros. Possui uma madeira rígida, dura,

acinzentada, com tronco ereto e galhos de caule longo e fino. Apresenta textura aberta com anéis de crescimento distintos. As toras são fortes e duráveis e utilizadas para os mais diversos fins, como construções de estradas de ferro, produção de papel, óleo e mel.

Apresentam inflorescências brancas que ocorrem do inverno ao início do verão. Possui um forte e vigoroso sistema de raiz o qual pode causar danos a construções e tubulações subterrâneas se a planta não for adequadamente localizada (WALTERS, 1998).

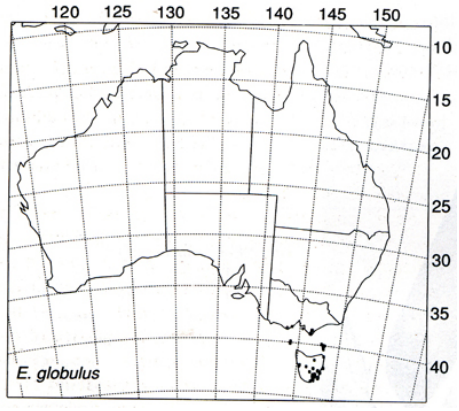


Figura 3. Distribuição da espécie *E. globulus* pelo continente Australiano. Adaptado de Eldridge *et al*., 1994.

1.2.3 Eucalyptus urophylla

O *Eucalyptus urophylla* é uma espécie do subgênero *Symphyomyrthus*, secção *Transversaria*, série *Salignae*, sub-série *Salignae*.

Sua área de ocorrência natural situa-se em algumas ilhas orientais do arquipélago de Sonda: Timor, Flores, Adonara, Lomblem, Pantar, Alor e Wetar, situadas ao norte da Austrália, além de outras ilhas a leste do arquipélago Indonesiano (Figura 4). Na área de ocorrência natural, a madeira é utilizada para construções e estruturas que demandem alta resistência. No Brasil, a madeira é para utilização geral (FERREIRA, 1979).

É a espécie que tem o maior potencial de crescimento em termos de área plantada em

função da tolerância ao fungo causador do cancro do eucalipto (*Cryphonectria cubensis*). Apresenta boa produtividade e potencial de utilização para os mais diversos fins, como fabricação de papel e celulose, chapas duras, serraria e produção de carvão (SCANAVACA JUNIOR, 2001).

Figura 4. Distribuição da espécie *E. urophylla* pelo continente Australiano. Adaptado de Eldridge *et al*., 1994

1.3 A lignina

Em vista da importância da madeira como matéria prima e fonte renovável de energia, investigações bioquímicas e moleculares sobre a formação da madeira estão sendo conduzidas em vários laboratórios do mundo. Se por um lado existe conhecimento relativamente extenso sobre a via biossintética da lignina, é ainda limitado o que se conhece sobre a bioquímica existente na produção de celulose e hemicelulose e os genes envolvidos nesses processos. O rápido avanço na pesquisa molecular da lignina possibilitou a descoberta de outros componentes da madeira que por sua vez aumentaram os conhecimentos e interesses da química da lignina para a indústria de polpa e papel (BAUCHER *et al.*, 2003).

A lignina é, depois da celulose, o segundo mais abundante biopolímero terrestre, agregando aproximadamente 30% do carbono orgânico da biosfera. A capacidade de sintetizar lignina foi uma essencial adaptação evolucionária das plantas do meio aquático para o terrestre (BOERJAN *et al.*, 2003; NICHOLSON & HAMMERSCHMIDT, 1992). Crucial

para a integridade estrutural da parede celular, a lignina permite a rigidez e condição ereta da planta (CHABANNES *et al.*, 2001; JONES *et al*., 2001). Além disso, fornece à parede celular a capacidade de impermeabilidade, permite o transporte de solutos através do sistema vascular, impede a perda de água excessiva por transpiração e atua na proteção da planta contra o ataque de patógenos (HUMPHREYS *et al*., 1999).

Os três maiores componentes de formação da parede celular são a celulose, a hemicelulose e a lignina. Moléculas longas de celulose proporcionam o esqueleto das paredes. Cadeias lineares de celulose são alinhadas conjuntamente em estruturas conhecidas como fibras elementares ou protofibrilas que, associadas a estruturas mais complexas, são chamadas de microfibrilas. Hemicelulose e outros carboidratos produzem a matriz da parede celular onde a lignina, um polímero fenólico hidrofóbico heterogêneo, incrusta outros componentes da parede tornando-a resistente e impermeável (DELMER & AMOR, 1995).

A madeira é a principal matéria prima utilizada na produção de polpa e papel (FAO, 2001). Durante a produção do papel, a lignina é quimicamente separada dos outros componentes polissacarídicos presentes na madeira através de reações de polpagem seguida de branqueamento do papel, o que leva a um grande consumo de reagentes químicos e energia, tornando-a uma atividade onerosa para a indústria, além da questão da poluição do meio ambiente (BIERMANN, 1996).

A etapa de remoção da lignina é delicada e influencia diretamente na qualidade final do papel a ser produzido. Resíduos de lignina na polpagem causam redução do branco do papel, mas é de fundamental importância manter um balanço final para evitar perda da qualidade e quantidade de polpa por degradação da celulose. Por estas razões, novas tecnologias biológicas têm sido criadas e testadas a fim de evitar tais problemas. Uma das novas tecnologias é o pré-tratamento da madeira com enzimas ou fungos que degradam a lignina (MESSNER & SREBOTNIK, 1994; VIIKARI *et al.*, 1994). Outra alternativa,

complementar à primeira, é modificar a constituição da lignina ou alterar árvores geneticamente para reduzir a produção de lignina ou tornar mais simples a sua extração. Obviamente, para atingir esse objetivo, faz-se necessário um profundo entendimento sobre a biossíntese da lignina em níveis bioquímicos e moleculares. Com base nestas necessidades, pesquisas recentes no assunto têm sido motivadas pelo interesse das indústrias de polpagem (BAUCHER *et al*., 2003).

Estudos recentes de produtos derivados da lignina constataram um potencial valor comercial, inclusive para as ligninas alquiladas, que possuem propriedades úteis como agentes dispersantes e emulsificantes (KOSIKOVA *et al*., 2000), e para as ligninas usadas como co-polímeros termoplásticos (LI & SARKANEN, 2000) e hidrocarbonetos aromáticos líquidos obtidos através de clivagem da lignina (THRING *et al*., 2000). Modificações na via de biossíntese da lignina que resultem na incorporação de diferentes precursores poderão potencialmente expandir a variedade de produtos derivados da lignina.

1.3.1 Composição e estrutura da lignina

A lignina é um heteropolímero aromático complexo composto principalmente de três monômeros derivados do álcool hidroxicinamílico que diferem entre si por graus de metoxilação nas posições dos carbonos C3 e C5 do anel aromático (FREUDENBERG & NEISH, 1968).

A biossíntese desses monômeros de lignina, ou monolignóis, inicia-se com a deaminação do aminoácido fenilalanina e envolve sucessivas reações de hidroxilação do anel aromático, seguido de *o*-metilação fenólica e conversão do grupo carboxil em um grupo hidroxil, resultando nos álcoois *p*-coumaril, coniferil e sinapil (BOERJAN *et al*., 2003) (Figura 5).

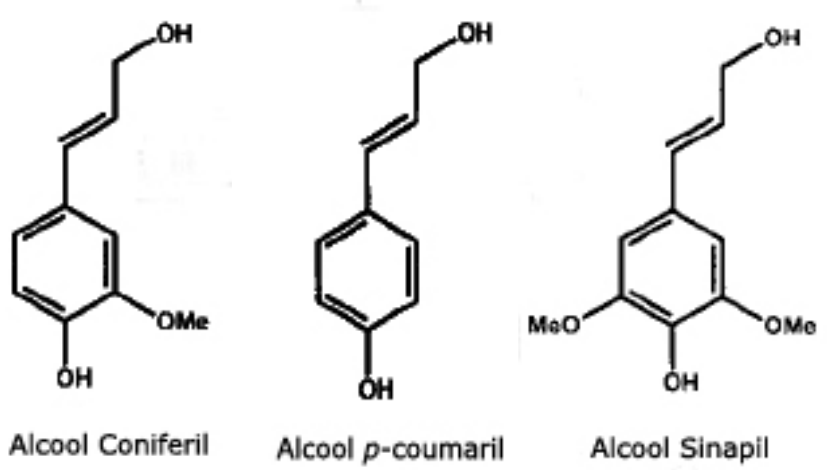


Figura 5. Estrutura molecular dos álcoois constituintes da lignina. Extraído e modificado de Boerjan *et al*., 2003.

Inicialmente imaginava-se que as reações de hidroxilação e metilação ocorriam em nível de ácidos cinâmicos e que os ácidos *p*-coumárico, ferúlico e sinápico eram subseqüentemente convertidos para os correspondentes monolignóis pela ação seqüencial das enzimas 4CL, CCR e CAD. Entretanto, a descoberta de todos os passos para a formação desses monômeros foi possível através de ensaios enzimáticos *in vitro* que permitiram a identificação dos genes envolvidos na via. A partir de análises de mutantes e plantas transgênicas modificadas na biossíntese de monolignóis, uma via de formação dos monolignóis foi inicialmente sugerida mas já passou por diversas revisões realizadas por vários autores. Para os nossos estudos, adotaremos a descrita por Boerjan *et al.* (2003), conforme a figura 6.

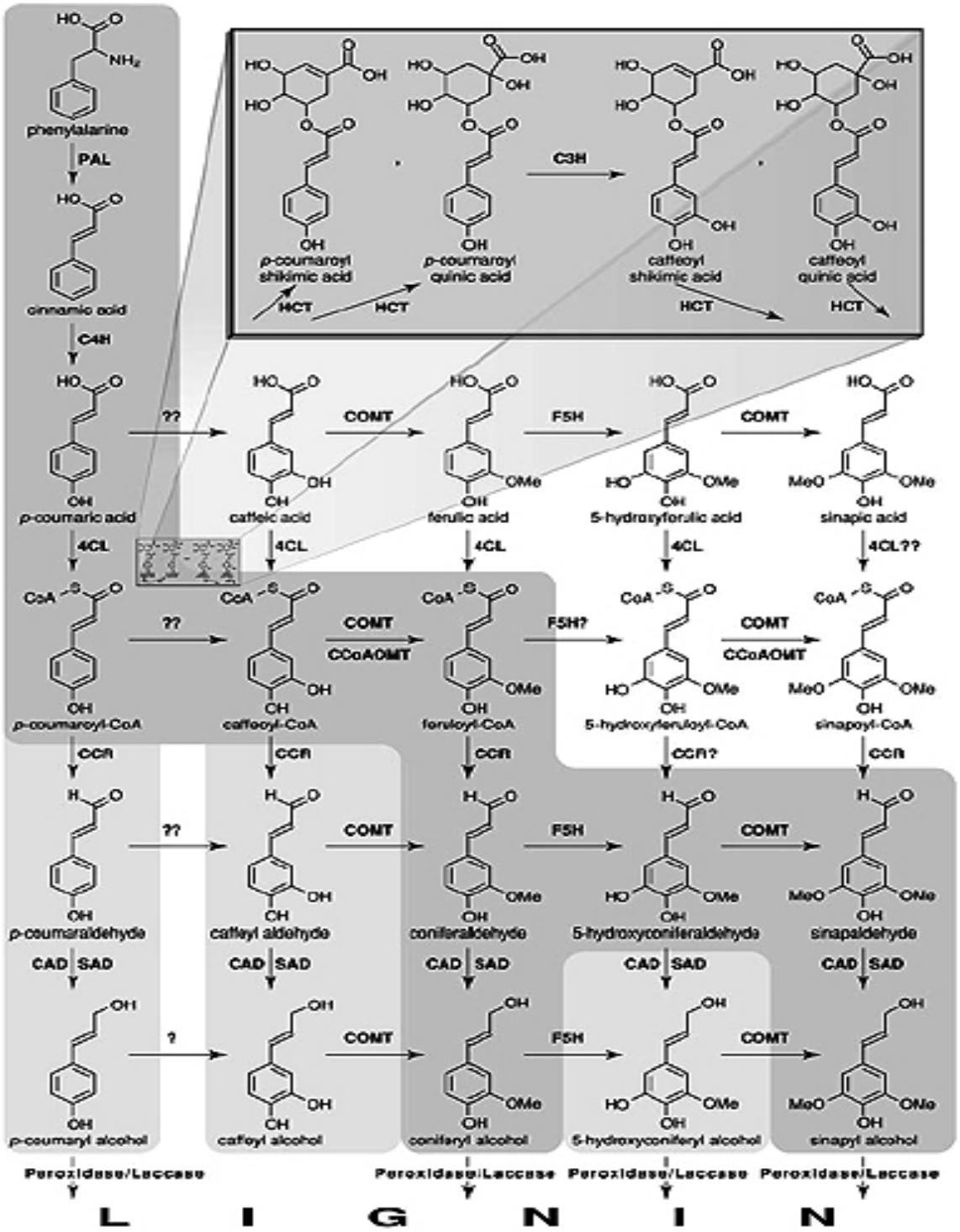


Figura 6. Via biossintética dos monolignóis. A rota cinza escuro representa a via de produção dos monolignóis mais amplamente aceita e descrita para as angiospermas. A rota cinza claro é um caminho alternativo que ocorre dependendo das condições ambientais e da espécie. A parte branca é descrita por alguns autores mas não apresenta significante papel na biossíntese dos monolignóis. Extraído e modificado de Boerjan *et al*., 2003. As siglas dos nomes encontram-se na lista de abreviações.

Os álcoois *p*-coumaril, coniferil e sinapil, quando incorporados ao polímero da lignina, são denominados, respectivamente, de unidades *p*-hidroxifenil (H), guaiacil (G) e siringil (S). Esses monômeros são sintetizados intracelularmente, exportados para a parede celular e, subseqüentemente, polimerizados. A lignina é formada através da polimerização desidrogenativa dos monolignóis, realizada por diferentes classes de proteínas, tais como peroxidases e oxidases (CHRISTENSEN *et al.*, 2000).

Além desses três principais monolignóis, o polímero lignina contém traços de unidades de monolignóis biossintetizados incompletamente e incorpora várias outras unidades de fenilpropanóides tais como hidroxicinamaldeídos, acetatos, *p*-coumaratos, *p*-hidroxibenzatos e ferrulatos tiamínicos (SEDEROFF *et al.*, 1999; BOERJAN *et al*., 2003).

Estudos mais recentes vêm tornando cada vez mais claro o fato das ligninas serem derivadas de vários monômeros e não somente de três monolignóis. Muitas plantas normais contêm uma fração substancial de ligninas derivadas de outros monômeros, além de traços de unidades incompletas (RALPH *et al.*, 2001). Muitas destas unidades têm sido identificadas pela sua maior incorporação no polímero de lignina em plantas transgênicas e mutantes com perturbações na via de biossíntese de monolignóis.

Ralph (1997) sugere que as plantas simplesmente necessitam de um polímero com propriedades mecânicas específicas e que a atual composição da lignina não é particularmente importante para a planta. Este mesmo autor afirma que plantas normais são produzidas quando a produção do monômero tradicional é inibida e as plantas fazem lignina a partir de outros precursores que não os três mais comumente utilizados (ANTEROLA & LEWIS, 2002). Isto sugere que existe um molde de formação da lignina, mas as suas variações muitas vezes também são aceitáveis na constituição das paredes celulares dos vegetais.

Para a maioria das enzimas descritas na via adotada para o nosso estudo, ocorrem múltiplas isoformas que são diferencialmente expressas durante as fases de desenvolvimento

da planta e dependentes do ambiente na qual se encontra (CHEN *et al.*, 2000; HU *et al*., 1998; LAUVERGEAT *et al.*, 2001; LINDERMAYR *et al.*, 2002), apresentando diferentes atividades cinéticas e preferências de substrato (EHLTING *et al.*, 1999; HARDING *et al.*, 2002; ZUBIETA *et al.*, 2002). Certos caminhos na via são mais favorecidos cineticamente do que outros, ocorrem em tipos específicos de células ou condições ambientais ideais, o que permite uma flexibilidade metabólica (Fig. 6).

Como exemplo, existem as isoformas da enzima 4CL (4-Coumarato CoA Ligase) que na maioria das plantas analisadas utilizam os ácidos *p*-coumárico, caféico e o ferúlico como substratos, mas não o sinápico. Isoformas de algumas plantas são capazes de converter ácido sinápico em sinapoil-CoA (LINDERMAYR *et al.*, 2002), criando a possibilidade em determinadas plantas, que monolignóis possam ser sintetizados pela via dos ácidos (VOGEL & JUNG, 2001).

Um outro nível de complexidade diz respeito aos intermediários da via que podem afetar a síntese ou a atividade de certas enzimas. O ácido cinâmico, por exemplo, inibe a expressão do gene *pal* (Fenilalanina amônia liase) em nível transcricional e pós transcricional (BLOUNT *et al.*, 2000; JONES *et al*., 2001) e induz a atividade de HCT (Hidroxicinamoil CoA transferase) (LAMB, 1977). Em estudos com fumo transgênico, a baixa expressão de C4H (Cinamato 4-Hidroxilase) reduz a atividade da PAL por retroalimentação (BLOUNT *et al.*, 2000).

Concentrações de fenilalanina também afetam de forma substancial o fluxo da via. Anterola *et al.* (2002) demonstraram em experimentos com cultura de células em suspensão enriquecidas com o aminoácido fenilalanina que ocorre um aumento nos níveis dos álcoois *p*-coumaril e coniferil e dos transcritos dos genes *pal*, *4cl*, *ccoaomt* (Cafeoil-CoA *o*-metiltransferase) e *ccr* (Cinamoil-CoA redutase). A atividade das enzimas C4H e C3H (4-Coumarato 3-hidroxilase) mostraram-se pouco elevadas em relação às enzimas anteriormente

citadas.

Dessa forma, o quão importante é uma ou outra enzima constituinte da via de biossíntese de lignina é, de fato, difícil avaliar, mas certamente haverá alguma em que alterações na sua expressão ou na sua constituição resultarão em alterações fenotípicas de maior impacto na estrutura, composição e conteúdo de lignina na planta.

1.4 Transgenia

Ensaios *in vitro* com enzimas de uma determinada via e experimentos com marcação radioativa têm sido instrutivos para o entendimento da via de biossíntese dos monolignóis, mas são insuficientes para a compreensão da complexidade da via *in vivo*.

Plantas transgênicas conseguem revelar uma situação mais próxima da realidade *in vivo*. Grandes alterações na quantidade, composição e estrutura primária de algum produto e os efeitos fenotípicos causados por alterações na expressão de um único gene são palpáveis e permitem uma visão global e real de como se comportam essas alterações no próprio organismo.

Nesse contexto, vários genes de múltiplos passos da via biossintética de formação dos precursores da lignina têm sido finamente regulados para a diminuição ou aumento de atividade por estratégias de supressão por *anti-sense* ou superexpressão por homólogos *sense*. Os efeitos dessas manipulações biotecnológicas no processo de lignificação e, em alguns casos específicos, a performance de plantas transgênicas e mutantes naturais durante a etapa de polpagem, têm sido avaliados (JOUANIN *et al.*, 2000).

Como um sistema modelo para o desenvolvimento de plantas, o organismo *Arabidopsis* tem se apresentado como uma ferramenta genética poderosa com a qual é possível dissecar processos complexos como a lignificação (DEAN, 2005; ANDERSON & ROBERTS, 1998). No mutante (*fah*1), experimentos com marcação radioativa indicaram que

a mutação bloqueava a via de biossíntese na etapa de formação do ácido sinápico (CHAPPLE *et al.*, 1992). Este bloqueio resultou na produção de um mutante em que o polímero de lignina não continha resíduos de siringil. Para surpresa da maioria dos pesquisadores, trabalhos recentes têm demonstrado que o substrato *in vivo* desta enzima é, de fato, coniferil aldeído, e não ferulato (HUMPHREYS *et al*., 1999; LI *et al.*, 2000; OSAKABE *et al.*, 1999). Esses achados, em conjunto com outros estudos em plantas geneticamente modificadas, têm trazido sugestões para revisões dos últimos passos da via de monolignóis.

Mutantes de milho, os mais antigos descritos na literatura - há mais de 75 anos -, despertaram um significativo interesse agronômico. Alterações nos seus tecidos vegetais apresentaram diferenças que permitem a digestão mais fácil pelos ruminantes do que as plantas normais, promovendo melhor nutrição para os animais (CHENEY *et al*., 1991). Porém, mesmo com essa grande vantagem, as variedades conhecidas dessas plantas não são largamente cultivadas, por apresentarem crescimento lento, serem mais susceptíveis a pestes e possuírem reduzida produção. Kuc e Nelson (1964) foram os primeiros a demonstrar uma variedade mutante, especificamente os mutantes *bm*1, identificados por uma produção anormal de lignina. Esses mutantes são afetados na expressão da atividade em CAD (Cinamil álcool deidrogenase), enquanto os mutantes *bm*3, são alterados na atividade de COMT (Ácido Caféico *o*-Metiltransferase) (VIGNOLS *et al.*, 1995).

Os resultados obtidos pela análise de plantas transgênicas modificadas na via de biossíntese de monolignóis têm demonstrado que (1) plantas podem tolerar grandes variações no conteúdo e composição da lignina; (2) que outros monômeros que não os álcoois *p*-coumaril, coniferil e sinapil são incorporados nos polímeros de lignina e que (3) a copolimerização desses monômeros incomuns podem resultar em estruturas viáveis de lignina. Esses dados mostram que o polímero de lignina é extremamente flexível na sua composição (BOERJAN *et al*., 2003).

Ao mesmo tempo em que contribuíram para o entendimento de passos importantes da biossíntese da lignina e sua estrutura, as mesmas linhagens transgênicas têm demonstrado que a lignina é importante para a integridade estrutural da parede celular. Plantas que tiveram a sua atividade reduzida em C3H, CCoAOMT e CCR, todas tiveram reduções no conteúdo de ligninas, associadas com o colapso dos vasos e alterações no crescimento, fenótipos alterados estes que podem variar significativamente de acordo com as condições de desenvolvimento e ambiente (PINÇON *et al.*, 2001a; PINÇON *et al.*, 2001b). Estudos mais detalhados com tabaco em condições de atividade reduzida da enzima CCR e de mutantes naturais *irx4* de *Arabidopsis* avaliaram as características e as diferenças nos vasos que sofreram o colapso, revelando uma expansão da parede secundária e a individualização das microfibrilas de celulose (LAMB, 1977).

Entretanto, em algumas situações, a redução do conteúdo de lignina não está associada com o aparecimento de anormalidades. Lee *et al.* (1997) comprovaram que a redução da atividade de 4CL em álamo (*Populus sp.*) resulta no aumento de crescimento da planta, corroborando os resultados de Hu *et al.* (1999), que demonstraram ainda que o conteúdo de lignina *per se* não é essencial para a integridade da estrutura da parede e a redução do conteúdo pode ser compensada por outros constituintes da parede celular. Outro resultado interessante obtido com a transgenia foi em *Populus*, onde observou-se um incremento na quantidade de celulose acompanhada de um decréscimo na lignina decorrente da redução na atividade em 4CL. Contudo, esse não parece ser um fenômeno comum, já que em mutantes de Arabidopsis (*Irx4*) defeituosos em CCR, tiveram menos lignina mas nenhum acréscimo na celulose (JONES *et al*., 2001). Isto sugere que a quantidade e a composição das ligninas variam de acordo com a espécie do organismo, tipos celulares e parede celular, e são influenciadas por uma série de fatores como o estágio de desenvolvimento e pelo ambiente em que se encontram, demonstrando a necessidade de estudos mais aprofundados em

organismos geneticamente modificados.

1.4.1 Transgenia envolvendo o gene 4cl

Experimentos com transgenia realizados por Kajita *et al*. (1996) utilizaram a expressão *anti-sense* para reduzir a atividade de 4CL em fumo transgênico, resultando em tecidos de xilema do caule com coloração marrom e reduzidos níveis de lignina. A taxa de siringil/guaiacil da lignina foi reduzida, assim como o conteúdo total da lignina, ocasionando um aumento no número de vasos que colapsaram nos tecidos de xilema (KAJITA *et al.*, 1997a). Em algumas plantas transformadas com construções *sense* de *4cl*, as atividades de co-supressão não foram uniformes por todo o xilema, mas demonstraram padrões setorizados (KAJITA *et al.*, 1997b).

Em *Arabidopsis*, a repressão *anti-sense* da atividade da enzima 4CL reduziu o conteúdo de lignina das plantas, levando a um aumento da razão siringil/guaiacil, com a diminuição significativa das unidades guaiacil e o aumento das unidades siringil (LEE *et al.*, 1997). O resultado mais atrativo obtido a partir da redução de expressão gênica de 4CL via *anti-sense* foi reportado por Hu *et al.* (1999) em álamos transgênicos. Esta espécie possui pelo menos dois genes diferentes codificando para a enzima 4CL, um que é expresso na epiderme da folha e aparentemente está envolvido na via de flavonóides e o segundo é expresso exclusivamente em tecidos de xilema lignificados. Em árvores que expressavam a construção *anti-sense* para o segundo gene, os níveis de lignina foram reduzidos em mais da 45%, com nenhuma aparente alteração na composição de lignina. Inesperadamente, essas reduções na quantidade foram acompanhadas em folha, caule e raiz, indicando um possível mecanismo para a redução na constituição de lignina em madeiras como um todo.

Num contexto geral, estes estudos indicam que apesar das significativas reduções nos níveis de lignina, a composição do biopolímero permanece próxima do original, alterando em

maiores proporções na quantidade final de lignina mas mantendo a sua qualidade.

1.4.2 Transgenia envolvendo o gene ccoaomt

Zhong *et al.* (1998) estudaram a atividade de CCoAOMT na via de lignificação. A repressão *anti-sense* da expressão de CCoAOMT em fumo demonstrou um decréscimo em quantidade de lignina e alterou a sua composição pelo aumento na proporção de resíduos siringil. No mesmo estudo, reduções na atividade de COMT não afetaram a quantidade de lignina, mas resultaram na produção de lignina com altas concentrações de guaiacil. A simultânea redução de COMT e CCoAOMT em fumo demonstrou uma redução na quantidade de lignina além da obtida pelo bloqueio da atividade de CCoAOMT sozinho. Inibição *anti-sense* da expressão de CCoAOMT em *Populus* demonstrou um modesto decréscimo no conteúdo de lignina, um aumento não significativo nas taxas de siringil para guaiacil e um dramático aumento nos níveis de um produto derivado do ácido sinápico, que não unidades siringil (MEYERMANS *et al.*, 2000).

1.5 As enzimas

A via de biossíntese dos precursores da lignina é a mais bem conhecida no processo de formação da madeira e tem sido o foco da experimentação em biologia molecular florestal. A maioria dos genes que codificam para as enzimas conhecidas desta via, bem como fatores de transcrição e proteínas de parede, foram clonados e caracterizados particularmente em *Pinus taeda* (ALLONA *et al.*, 1998). Além disso, árvores transgênicas de álamo super expressando ou sub-expressando algumas destas enzimas apresentaram alterações importantes na quantidade e qualidade final de lignina (LI *et al.*, 2003). Utilizando *Agrobacterium* como sistema de transformação, construções *antisense* de *4cl* e sense de *cald5h* (Coniferaldeído 5-hidroxilase) foram introduzidas em *Populus tremuloides* gerando árvores que expressavam

cada um e ambos transgenes. Alterações de um pouco mais de 40% de redução no conteúdo de lignina com o incremento de 14% na quantidade de celulose foi obtido em plantas transformadas com a construção *antisense* de *4cl*. Já para as plantas transgênicas transformadas com a construção sense do gene *cald5h* houve aumentos de até 3 vezes na razão siringil/guaiacil sem alterações na quantidade de lignina. Em plantas expressando ambas as construções, os efeitos foram aditivos apresentando 52% menos lignina, um aumento de 64% na razão siringil/guaiacil e 30% mais celulose. Embora genes já tenham sido identificados e expressados em sistemas transgênicos, pouco se sabe sobre a relação entre a diversidade nucleotídica dos variantes alélicos destas enzimas presentes na natureza e as características finais da lignina. Potenciais dificuldades para a utilização em escala de árvores transgênicas e a ampla variabilidade existente no gênero *Eucalyptus* para lignina, sugerem que o entendimento preciso e exploração refinada da diversidade gênica natural para estas enzimas poderá ser uma estratégia de grande impacto no melhoramento florestal.

Para promover uma alteração da lignificação, modificações na via de biossíntese de monolignóis em determinados pontos poderão proporcionar o efeito desejado. Tendo em vista os resultados de experimentos de transgenia, a seguinte hipótese de trabalho pode ser lançada: existe correlação entre a variabilidade de seqüência de genes da via de lignificação e a variabilidade fenotípica nas propriedades químicas da madeira. Esta variabilidade de seqüência pode ser tanto na região codificadora quanto em regiões reguladoras em *cis* que por sua vez afetam a forma como fatores *trans* atuam na regulação da expressão destes genes. Neste trabalho as atenções foram focalizadas em dois genes específicos da via, *4cl* (4-Coumarato:Coenzima A ligase) e *ccoaomt* (Cafeoil-CoA *o*-metiltransferase).

1.5.1 Enzima 4CL

A enzima 4-coumarato:CoA ligase (4CL) faz parte da via de formação da lignina

mediando o último passo do metabolismo geral de fenilpropanóides. A enzima 4CL é membro de uma superfamília de enzimas formadoras de adenilatos que compartilham mecanismos de reação comum com a formação de um intermediário do substrato adenilato na presença de ATP e magnésio, seguido de uma esterificação com coenzima A (CoA ligase), 4´-fosfopanteteína ou oxidação por molécula de oxigênio. Os produtos da reação, ésteres de hidroxicinamoil CoA, servem como substratos para a via específica de formação de fenilpropanóides.

Como sinônimos para 4CL estão a 4-coumaroil-CoA sintase, 4-Coumaril-CoA sintetase, *p*-coumaroil CoA ligase, hidroxicinamoil CoA sintetase, *p*-hidroxicinamoil coenzima A sintetase. Possui EC (*Enzime Commission*) 6.2.1.12 que corresponde a uma ligase, formando ligações carbono-enxofre, uma ligase ácido-tiol, com reação principal: ATP + 4-coumarato + CoA = AMP + difosfato + 4-coumaroil-CoA (Figura 7).

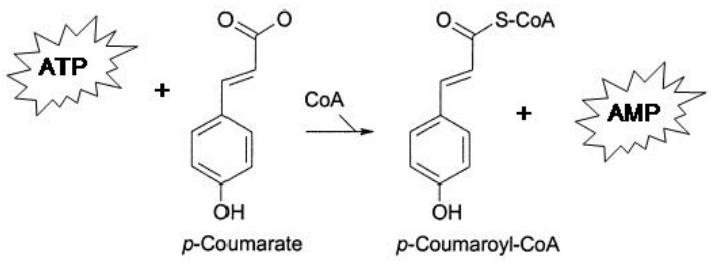


Figura 7. Ilustração esquemática de reação realizada pela 4-coumarato:CoA ligase. Adaptação de Dean, 2005.

A grande importância dos fenilpropanóides para as plantas diz respeito à proteção contra luz ultravioleta (UV) que lhe é conferida e a formação de lignina.

1.5.2 Enzima CCoAOMT

A enzima CCoAOMT, de grande importância na via de formação da lignina, tem papel principal na síntese de polissacarídeos derivados do feruloil. Na via de biossíntese da

lignina, tem como função a metilação de cafeoil-CoA para feruloil-CoA e 5-hidroxiferuloil-CoA para sinapoil-CoA. Parvathi *et al.* (2001) demonstraram a preferência cinética, *in vitro*, de CCoAOMT por Cafeoil CoA.

A enzima *ccoaomt* tem sido isolada e caracterizada em um grande número de plantas, como a alfafa (*Medicago sativa*, INOUE *et al.*, 1998), cenoura (*Daucus carota*, KÜHNL *et al.*, 1989), salsa (*Petroselium crispum*; PAKUSCH *et al.*, 1989), álamo (*Populus* spp.; MEYERMANS *et al.*, 2000), pinheiro (*Pinus taeda*; LI *et al.*, 1999) e tabaco (*Nicotiana tabacum*; MARTZ *et al.*, 1998). Esta enzima está envolvida no reforço das paredes celulares vegetais além de responder a ferimentos e a ataques de patógenos (desafio à patógenos) aumentando a formação de polímeros de ácido fenólico, ligantes de parede celular (GASTEIGER *et al.*, 2003; UNIPROTKB/SWISS-PROT , 2003).

Possui EC 2.1.1.104 que corresponde a uma transferase de grupos de um carbono, da classe das metiltransferases, com reação: S-adenosil-L-metionina + cafeoil-CoA = S-adenosil-L-homocisteína + feruloil-CoA. Como sinônimos, Trans-cafeoil-CoA 3-*o*-metiltransferase, CCOMT e CCoAMT (Figura 8).

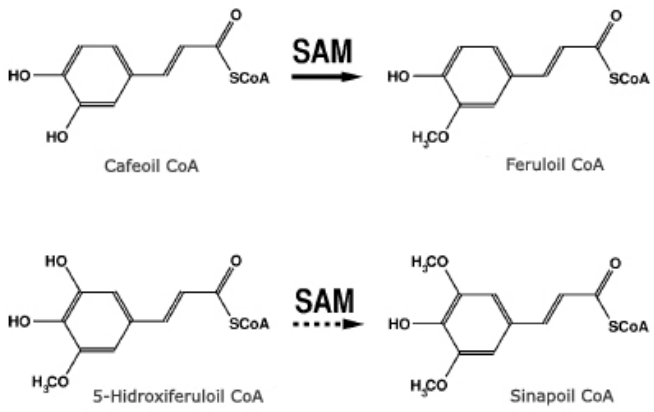


Figura 8. Reação de metilação catalisada por CCoAOMT. A enzima CCoAOMT metila os substratos Cafeoil CoA e 5-Hidroxiferuloil CoA, com preferência, em reações *in vitro*, por Cafeoil CoA. O pontilhado da seta refere-se à possibilidade da reação. SAM (S-adenosil-L-metionina) é o doador do metil em ambas as reações. Ilustração modificada de Ferrer *et al.*, 2005.

1.6 Estudo da seqüência nucleotídica

Análises moleculares nos últimos 20 anos têm demonstrado que existe uma grande variação genética entre indivíduos de uma mesma espécie e entre espécies do mesmo gênero. Esta variação se manifesta em nível morfológico, cromossômico e principalmente nucleotídico.

De maneira geral existem três principais fontes de geração e ampliação de variação genética: mutação, recombinação e fluxo gênico/hibridação. O processo de mutação envolve alterações de base nucleotídica, a maioria delas silenciosas, com eventuais reflexos em troca de aminoácido e geração de nova seqüência protéica. A taxa de mutação é definida como a probabilidade de uma cópia de um alelo mudar para alguma outra forma alélica em uma geração. As taxas de mutação são tão baixas que apenas a mutação não pode responder pela rápida evolução das populações e espécies. Se um novo variante alélico não for perdido, processos genéticos, demográficos e evolucionários, em adição ao processo de deriva, determinam a sua freqüência populacional e sua associação não aleatória com locos adjacentes (desequilíbrio de ligação; DL) (BROWN *et al.*, 2004), que pode ser quebrado pelo processo de recombinação.

No passado, a variação genética em muitas espécies de plantas era mensurada por eletroforese. Hamrick e Godt (1996) sumarizaram dados de aloenzimas para diversas espécies florestais, incluindo informações sobre as características das espécies tais como ciclo de vida, sistema de polinização, dispersão de sementes e sistema de cruzamento. Eles demonstraram que essas características estavam intimamente relacionadas com as quantidades e padrões de variação genética para as espécies em estudo.

Com o avanço da tecnologia molecular, progressos recentes nas técnicas de seqüenciamento permitem hoje estudar a variação genética em detalhe, em complemento às descobertas inicialmente realizadas por eletroforese.

Em plantas, a variação nucleotídica em genes nucleares tem sido bem estudada em plantas anuais, especialmente em *Arabidopsis*. Esses estudos têm proporcionado informações interessantes sobre a estrutura populacional e seleção natural que, por inferência, pode ser esperada para as diversas espécies de plantas, até mesmo árvores. Entretanto, por usualmente possuírem um longo tempo de geração, os efeitos de mutação e seleção nas árvores podem ser diferentes das plantas herbáceas (DVORNYK *et al.*, 2002), exigindo-se assim, estudos específicos.

Desta forma, vários experimentos têm sido conduzidos a fim de esclarecer esta possível diferença. Stephan e Langley (1998) demonstraram em *Lycopersicum* que em 36 loci de oito espécies diferentes ocorrem polimorfismos naturalmente e esse polimorfismo é correlacionado positivamente com a densidade de *crossing-over* ao longo da região do genoma estudada. Brown *et al.* (2004) demonstraram a diversidade nucleotídica em *Pinus* de aproximadamente 18 Kb de DNA, abrangendo 19 loci diferentes. Já em milho (*Zea mays*), o estudo da seqüência de DNA foi mais aprofundado, na tentativa de associar fenótipos com as alterações observadas na seqüência de DNA do gene responsável pelo tempo de florescimento da espécie, *dwarf8* (*d8*) (THORNSBERRY *et al.*, 2001). Outro estudo com milho revelou a seqüência nucleotídica e a variabilidade em 18 segmentos de genes (CHING *et al.*, 2002) e comprovou os altos níveis de diversidade nucleotídica em regiões próximas ao gene *tb1* (CLARK *et al.*, 2004).

Em arroz, a seqüência do gene *xa5* foi descrita, além de 45 Kb de seqüência próxima ao gene. As diferenças encontradas na região acima e abaixo do gene foram analisadas e correlacionadas com o gene (GARRIS *et al.*, 2003). Em *Pinus*, um grande investimento em estudos de seqüência nucleotídica tem trazido várias informações importantes, tanto para o gênero, quanto para agregar informações para o entendimento geral de árvores. Estudo em diversos genes do gênero *Pinus* (DVORNYK *et al.*, 2002; NEALE e SAVOLAINEN, 2004;

GARCIA-GIL *et al.*, 2003) e em genes específicos para a formação de madeira foram estudados (GARNIER-GÉRÉ *et al.*, 2003).

Da mesma forma, em *Eucalyptus,* estudos quanto à seqüência nucleotídica foram realizados em dois genes importantes da via de biossíntese de lignina, o *ccr* e o *cad* (POKE *et al.*, 2003; HAWKINS *et al.*, 1994). Validando a importância de alguns genes presentes em espécies vegetais, estudos em outros gêneros abrangendo estes dois genes foram realizados em fumo (PIQUEMAL *et al.*, 1998; BOUDET *et al.*, 1998), milho (BAUCHER *et al.*, 1998) e em *Pinus taeda* (RALPH *et al.*, 1997), permitindo a comparação da seqüência nucleotídica entre as espécies e o estudo evolucionário das espécies.

1.7 Mutações

Os estudos de variação em nível de pares de bases pelo seqüenciamento de DNA podem fornecer informações de dois tipos. Primeiro, traduzindo as seqüências de regiões codificantes obtidas de diferentes indivíduos em uma mesma população ou de espécies diferentes, podem ser determinadas diferenças de seqüência de aminoácidos. Os estudos eletroforéticos mostram apenas que há variação nas seqüências de aminoácidos, mas não podem identificar quantos ou quais aminoácidos diferem entre os indivíduos.

Segundo, a variação de pares de bases também pode ser estudada para aqueles que não determinam ou mudam a seqüência da proteína. Isto inclui o DNA nos íntrons, nas seqüências flanqueadoras de 5' que podem ser regulatórias, no DNA não-transcrito em 3' para o gene, e nas posições dos nucleotídeos dentro de códons, usualmente nas terceiras posições, cuja variação em sua maioria não resulta em substituições de aminoácidos. Dentro de seqüências codificantes, estes chamados polimorfismos de par de base silenciosos são mais comuns que as mudanças que resultam em polimorfismos de aminoácidos, supostamente porque muitas mudanças de aminoácidos interferem no funcionamento normal da proteína e são eliminadas

por seleção natural.

Existem também restrições em seqüências não-codificantes 5’ e 3’ e em íntrons. Tanto o DNA não-codificante 5’ quanto o 3’ contém sinais de transcrição e os íntrons podem conter acentuadores de transcrição (GRIFFITHS *et al*., 2001).

As alterações na constituição da seqüência do DNA, resultantes em polimorfismos, são caracterizadas por eventos de substituição de bases e adição ou deleção de bases.

A substituição de base consiste na alteração da base existente por uma outra, durante o processo de duplicação do DNA ou por efeitos externos que provoquem essa alteração, como radiação, por exemplo. A substituição de base pode ser classificada em dois tipos: as transições, que constituem a substituição de base por outra de mesma categoria química (purina:purina; pirimidina:pirimidina) e as transversões, que consistem na substituição de base por outra de categoria química diferente, como de uma purina por uma pirimidina ou vice e versa.

As mutações em regiões codificadoras podem ser classificadas em três tipos: mutações silenciosas, na qual o códon é modificado mas o aminoácido resultante é o mesmo. A segunda forma de mutação é a de sentido trocado, na qual o códon modifica o aminoácido. A terceira forma de mutação consiste na sem sentido, em que o códon que determina um aminoácido é substituído por um códon de terminação.

Essas substituições podem ser de forma sinônima, onde o aminoácido resultante é alterado, mas a sua característica química é semelhante ao aminoácido original, ou de forma não sinônima, onde a substituição é por um aminoácido de características químicas diferentes da que continha o aminoácido original. Neste caso, pode haver modificações estruturais e funcionais na proteína. Embora SNPs sinônimos não alterem a seqüência protéica, eles podem modificar a estrutura e a estabilidade do RNA mensageiro e, conseqüentemente, afetar a quantidade de proteína produzida. Além disso, SNPs podem afetar o processamento de introns

(“splicing” alternativo), causar alterações no padrão de expressão de genes (como no caso de alterações em seqüências de promotores) em um determinado tempo ou em resposta a condições ambientais ou em determinados tecidos, gerar ou suprimir códons de iniciação e terminação ou sítios de poliadenilação na molécula de RNA mensageiro (GRIFFITHS *et al*., 2001).

Atrelado ao conhecimento das mutações, o aprofundamento dos estudos das seqüências nucleotídicas e a busca incessante por correlações das diversas variações existentes, surgiram uma série de conceitos que agregam as diferentes áreas de pesquisa, como a estatística, a informática e a biologia.

1.8 SNPs

Single Nucleotide Polymorphism, ou seja, polimorfismo de base individual ou única (SNP) é o acrônimo amplamente utilizado hoje para definir a variação de seqüência observada em uma posição individual da seqüência de DNA, na qual alternativas de alelos existem em indivíduos normais em algumas populações, com freqüência de pelo menos 1%.

Os SNPs podem se apresentar na forma de polimorfismos bi-, tri- ou tetra-alélicos, mas na maioria das aparições é na forma bi-alélica. Além das variações múltiplas de base, os INDELs também são um grupo específico de SNPs. O conceito de SNP está intimamente associado a vários outros conceitos como atuação recessiva, baixo poder de penetrância, loci de características quantitativas (QTL) ou possibilidade de alelos associados, desde que todas estas características ocorram em alguns indivíduos normais (BROOKES, 1999). Entretanto, mesmo a maioria das variações sendo determinada como SNPs pela freqüência mínima de observação de 1%, alguns alelos de baixa freqüência são observados e recebem a denominação de variantes raros na população. A taxa de diferenças nucleotídicas entre duas seqüências escolhidas aleatoriamente é um parâmetro denominado diversidade nucleotídica

(NEI & LI, 1979), o qual será detalhado mais adiante.

Com a grande capacidade de seqüenciamento disponível hoje, o descobrimento, a validação e a utilização de SNPs têm ganhado particular atenção como uma ferramenta para a geração de um grande número de marcadores para mapeamento de todo o genoma. A alta freqüência com a qual os SNPs são encontrados ao longo do genoma definem a utilidade dos SNPs nos programas de mapeamento do cromossomo. Os SNPs podem ser utilizados para estudos de mapeamento com altas concentrações de marcadores, importantes para o estudo de genes candidatos e clonagem posicional ou, mais importante, como polimorfismos relacionados diretamente a algum gene de interesse.

Os SNPs são distribuídos por todo o genoma como na região de introns, exons, regiões intergênicas, promotores ou enhancers, etc. Entretanto, a localização dos SNPs pode ser de relevância funcional e fisiológica para o organismo. Um SNP localizado na região codificadora por exemplo, pode ter impactos profundos na formação da proteína. Um SNP intrônico pode influenciar na região de splicing do mRNA (KRAWEZAK *et al.*, 1992) assim como um SNP no promotor pode influenciar na expressão gênica (DRAZEN *et al.*, 1999). O grau com que o SNP afeta no fenótipo ou é associado com a variação fenotípica é o objetivo maior do estudo de SNPs.

Os polimorfismos podem ser utilizados como simples marcadores genéticos, com os quais podem ser identificados pontualmente os genes. Mas há também um grande potencial para o uso dos SNPs na detecção de associações entre forma alélicas de um gene e fenótipos, especialmente doenças ou características comuns que possuem genética multifatorial (RAFALSKI, 2002).

Atrelado ao conceito de SNP e à sua utilização, outros conceitos surgem para facilitar a utilização e compreensão da função dos SNPs nos diversos estudos. Um conceito amplamente utilizado é o de desequilíbrio de ligação (DL). Desequilíbrio de ligação é a não

independência dos alelos ao longo do genoma.

Se um sítio polimórfico é identificado em um segmento de DNA, ele pode servir como um ponto de referência física ou como um marcador molecular que pode ser seguido ao longo das gerações. Se o SNP está próximo a um gene de interesse, ele pode servir de identificador no genoma para o acompanhamento de outros polimorfismos naquele gene. Entretanto, a distância entre o marcador e o polimorfismo alvo pode ser tal que eventos de recombinação geram associações ao acaso entre alelos do marcador e do polimorfismo alvo, ou seja, os dois estão em equilíbrio de ligação. Este fenômeno, existente com diferentes extensões ao longo do genoma, pode impossibilitar existência de correlação entre marcador e gene alvo.

Baixos índices de recombinação e a associação entre alelos de dois locos diferentes caracterizam a situação de desequilíbrio de ligação. Se dois alelos são encontrados juntos em um gameta mais freqüentemente do que seria esperado, pelo produto de suas freqüências, os alelos estão em desequilíbrio de ligação (D), que é calculado a partir das freqüências genotípicas e alélicas observadas para os dois locos (INGVARSSON, 2005). O valor de D varia de acordo com a freqüência encontrada do genótipo na população e quanto mais distante de zero, maior a evidência de desequilíbrio dos alelos, ou seja, associação.

A fim de facilitar o entendimento, consideremos dois grupos de dados de seqüências que contém dois sítios de SNPs. No primeiro grupo, a primeira posição variante, para todos os indivíduos, o alelo A foi observado. Respectivamente, foi observado o alelo G na segunda posição também para todos os indivíduos, ou seja, associado ao alelo A. Esta associação dar-se o nome de desequilíbrio de ligação. Contrariamente, no segundo grupo analisado, os alelos As e Gs estão aleatoriamente associados nos indivíduos e não há uma relação de aparecimento de alelos no primeiro e segundo sitio. Desta forma, é dito que os sítios estão em equilíbrio de ligação (GAUT & LONG, 2003) (Figura 9).

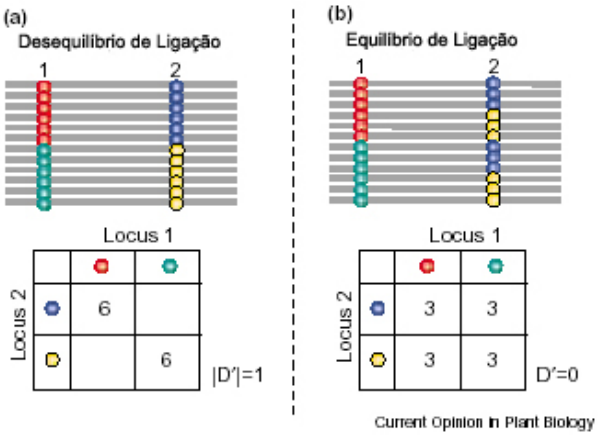


Figura 9. Representação esquemática da situação de desequilíbrio de ligação e equilíbrio de ligação entre dois loci. A. Quando o desequilíbrio de ligação está presente, todos os indivíduos que possuem o alelo vermelho no locus 1 possuem o alelo azul no locus 2. B. Quando há o equilíbrio de ligação, indivíduos com o alelo vermelho no locus 1 podem apresentar qualquer alelo no lócus 2. A tabela de contingência correspondente e os valores de D' estão também apresentados. Figura adaptada de Rafalski (2002).

Entretanto, fórmulas matemáticas foram desenvolvidas para calcular o valor de desequilíbrio de ligação entre dois sítios adjacentes. O "D" é a medida quantitativa de associação alélica. Acompanhando o raciocínio do exemplo acima citado, podemos assim calcular o desequilíbrio de ligação:

$$D_{12} = P_{1A2G} - (P_{1A})\,(P_{2G})$$

Onde D_{12} é a medida de desequilíbrio entre os sítios 1 e 2, P_{1A2G} é a freqüência de seqüências que contém o alelo A no sitio 1 e o alelo G no sitio 2, P_{1A} é a freqüência do alelo A no sitio 1 e P_{2G} é a freqüência do alelo G no sitio 2. Utilizando esta fórmula, é de praxe a obter o valor absoluto de D, não importando quais alelos são mensurados e associados.

De acordo com a fórmula acima citada, o valor de "D" agrega informações sobre associação e freqüências alélicas. Quando da análise de um grupo de seqüências onde algumas das seqüências não apresentam associação entre os alelos de dois sítios, ainda assim

o valor D é calculado. Devido à dependência das freqüências alélicas, é esperado que os valores de D sejam amplamente variáveis entre pares de SNPs analisados, mesmo quando os sítios estão em completo desequilíbrio de ligação, exatamente por dependerem também da freqüência observada dos alelos.

Existem duas formas de direcionar as freqüências alélicas. A primeira é ignorar os variantes de baixa freqüência que contribuem desordenadamente para a variação em D. Estudos mais rigorosos desprezam alelos que apresentam freqüências abaixo de 5%. A segunda solução é a de utilizar medidas de associação alélica que são normalizadas com freqüências alélicas. A mais comum medida de normalização é o D' (LEWONTIN, 1964).

Há outros diferentes tipos de medidas para a estimativa do desequilíbrio de ligação entre quaisquer dois loci bialélicos. Além de D e D', existem o r^2, R, D^2, D* entre outros (HEDRICK, 1987). De todas as medidas acima citadas para desequilíbrio de ligação, D'e r^2 são as medidas mais utilizadas para tal. D' é muito bem utilizado porque ele é diretamente proporcional à fração de recombinação. Entre as duas medidas preferidas, enquanto D' mede somente diferenças de recombinação, r^2 sumariza a história de recombinação e mutação, admitindo um intervalo de confiança não superior a 10%. Os testes de desequilíbrio de ligação são definidos como significantes se os valores de p estiverem dentro da faixa permitida de aceitação. Do contrário, as conclusões podem não ser confiáveis e aceitas.

Mesmo em algumas situações em que a recombinação não explica o padrão de variação entre dois sítios de SNPs, os níveis esperados de desequilíbrio de ligação são em função da recombinação. Quanto maior o nível de recombinação, mais permutado ficará os sítios, acarretando em baixo desequilíbrio de ligação.

O desequilíbrio de ligação é afetado por vários fatores biológicos. Subdivisão populacional e combinação de populações geram o aumento do desequilíbrio de ligação, mas seus efeitos dependem do número de populações, a taxa de troca entre as populações e a taxa

de recombinação (PRITCHARD & PRZEWORSKI, 2001). Similarmente, gargalos populacionais e seleção direcionada para certos alelos aumentam o desequilíbrio, mas na ausência de outro fator de alteração de populações, como subdivisão populacional, seus efeitos são de curta duração (PRZEWORSKI, 2002). Do contrário, quando ocorre cruzamentos aleatórios, mutação, recombinação e ausência de seleção natural, é favorecido o baixo desequilíbrio de ligação.

O estudo dos níveis de desequilíbrio de ligação em plantas é crescente e a forma de utilização mais amplamente desenvolvida e utilizada é a associação marcador-característica seguido de seleção marcador-assistido, empregada em programas de melhoramento. A seleção é realizada tanto ao nível de gene individual quanto de todo o genoma e incluem (1) estimativas de desequilíbrio de ligação em diferentes genomas de plantas ou em diferentes partes do genoma de uma única espécie; (2) medida da diversidade nucleotídica e estrutura haplotípica; (3) avaliação dos efeitos de seleção e domesticação; (4) identificação de associações marcador-característica; etc. (GUPTA *et al.*, 2005).

Em plantas, há poucos experimentos voltados para a estimativa de extensão do desequilíbrio de ligação. Em *Arabidopsis*, como esperado, o desequilíbrio de ligação estende-se por longas distâncias devido ao fato da espécie ser altamente autopolinizadora (NORDBORG, 2000), variando em torno de 250 Kb para o locus *FRIGIDA* (HAGENBLAD & NORDBORG, 2002) e 65 Kb para o locus *CRY2* (OLSEN *et al.*, 2004), ambos relacionados ao tempo de florescimento em *Arabidopsis*. Em milho, o desequilíbrio de ligação decai rapidamente, em torno de 1500 pb para quatro genes estudados e 7000 pb para o gene *su*1, devido principalmente à forma de cruzamento aberto e ao ativo sistema de retrotransposons (REMINGTON *et al.*, 2001). Em *Pinus*, o desequilíbrio de ligação decai, em média, a 1500 pb para vários genes estudados (NEALE & SAVOLAINEN, 2004).

Algumas espécies de plantas apresentam forte desequilíbrio de ligação entre os SNPs

distribuídos por longas distâncias do genoma. Entretanto, por existirem grandes variações de desequilíbrio de ligação ao longo do genoma, ao invés da utilização de um grande número de SNPs distribuídos, utiliza-se o conjunto destes, denominado haplótipos (STEPHENS *et al.*, 2001). O conceito de haplótipo está intimamente relacionado tanto à diversidade nucleotídica quanto ao desequilíbrio de ligação e significa o padrão de associação de SNPs que são herdados conjuntamente e caracterizam um grupo, espécie ou população. Normalmente, em estudos que determinam o nível de desequilíbrio de ligação ao longo de um segmento do genoma, também são inferidos os haplótipos e sua diversidade para a determinada espécie e região do genoma. Haplótipos são utilizados em estudos de associação e/ou mapeamento por desequilíbrio de ligação e já são descritos em algumas plantas, como milho (CHING *et al.*, 2002), tomate (SIMKO *et al.*, 2004), *Pinus* (DVORNYK *et al.*, 2002; GONZÁLEZ-MARTÍNEZ *et al.*, 2004), cevada (RUSSEL *et al.*, 2004), soja (CREGAN *et al.*, 2002; ZHU *et al.*, 2003) entre outros.

1.9 Quantificação da diversidade nucleotídica

Várias estatísticas descritivas são comumente utilizadas para resumir os dados de polimorfismo e estimar os parâmetros populacionais. Assumindo a teoria neutra proposta por Kimura (1968), a variação genética em nível molecular é considerada amplamente neutra, sem influência de forças seletivas, e a extensão da variação é determinada primeiramente pela taxa de mutação e o tamanho efetivo da população (KIMURA & CROW, 1964; NEI, 1987). Entretanto, é possível testar a hipótese de evolução neutra pela comparação das quantidades de variação genética observada e esperada. Se a discrepância entre as quantidades observadas e esperadas for grande, algum tipo de seleção foi invocado (NEI & KUMAR, 2000).

A partir de um modelo neutro é possível utilizar as medidas da diversidade nucleotídica para calcular a densidade esperada de sítios polimórficos para as diversas

freqüências alélicas. A disponibilidade de marcadores SNPs com freqüências alélicas apropriadas é importante para que estes possam ser efetivamente usados para o mapeamento genético, seja por análises de ligação ou baseadas em técnicas de mapeamento de associação por desequilíbrio de ligação (KRUGLYAK, 1997; KRUGLYAK, 1999).

A extensão do polimorfismo de DNA pode ser mensurada de várias maneiras diferentes mas a medida mais comumente utilizada é (1) o número de sítios segregantes por nucleotídeos e (2) a diversidade nucleotídica.

Ao considerar uma região do DNA, ou seja, um locus, e assumir que *m* cópias do segmento são aleatoriamente amostradas de uma população, contendo *n* nucleotídeos, e qualquer sítio nucleotídico que apresente dois ou mais nucleotídeos diferentes nas *m* seqüências, é denominado de sítio segregante. O número total de sítios segregantes (S) observados no grupo de dados dividido pelo número total de nucleotídeos examinados é designado número de sítios segregantes por nucleotídeos (p_s). Considerando o valor esperado de p_s sob condições de ausência de recombinação e a presença de novas mutações sempre ocorrendo em sítios não segregantes, tem-se o modelo genético chamado de modelo sítios-infinitos. Da mesma forma que p_s, θ é definido como um parâmetro de mensuração da variação genética, mais simples que p_s, exatamente pelo fato de o segundo ser proporcional à taxa de mutação e ao tamanho efetivo populacional, independente do tamanho da amostra (NEI & KUMAR, 2000).

O parâmetro de aplicação em modelos simples de genética de populações que determina a quantidade e a distribuição da diversidade nucleotídica é denominado parâmetro de mutação populacional,

$$\theta = 4N_e\mu$$

onde N_e é o tamanho efetivo da população e μ é a taxa de mutação por geração.

Estimativas do parâmetro de mutação populacional podem ser rapidamente calculadas a partir do número de sítios polimórficos presentes em uma amostra de seqüências obtida de uma população aleatória. Este parâmetro é denominado θ_w (WATTERSON, 1975).

Uma segunda forma de estimativa leva em conta a diversidade nucleotídica, também denominada de π (NEI & LI, 1979),

$$\pi = \sum_{ij}^{q} x_i x_j d_{ij}$$

onde q é o número total de alelos, x_i é a freqüência na população (em seqüências) do alelo i, x_j é a freqüência na população (em seqüências) do alelo j e d_{ij} é o número de diferenças nucleotídicas ou substituições por sitio entre os alelos i e j. Resumindo, é o somatório das diferenças de pares de nucleotídeos entre seqüências de uma amostra e depende tanto das freqüências quanto do número de sítios polimórficos, diferentemente de θ_w, que é independente de freqüências.

Para um melhor entendimento dos diferentes parâmetros de diversidade estimados consideremos um conjunto de 10 seqüências de 11 bases, conforme está indicado abaixo:

AGCTTAATTAG
AGCTTAATTTG
AGCTTAATTAG
AGTTTAATTAG
AGCTTAATTAG
AGCTTAATTAG
CGCTCAATTAG
CGCTCAATTAG
CGCTCAATTAG
AGCGCATTTAG

Observando estas seqüências temos que:

n = número de nucleotídeos , n = 11

S = número total de sítios segregantes = 6

p_s = número total de sítios segregantes por nucleotídeo = 6/11 = 0,545

Considerando apenas as primeiras quatro seqüências:

AGCTTAATTAG
AGCTTAATTTG
AGCTTAATTAG
AGTTTAATTAG

e as diferenças entre seqüências:

	Seqüência 1	Seqüência 2	Seqüência 3	Seqüência 4
Seqüência 1	0	1	0	1
Seqüência 2	1	0	1	2
Seqüência 3	0	1	0	1
Seqüência 4	1	2	1	0

A diversidade nucleotídica π é igual a:

$$\pi = (1 / a) \times \sum \text{diferenças} / (a \times n) =$$

$$\pi = ¼ \times (12/(4x11)) = 0{,}06818$$

onde

n = número de nucleotídeos , n = 11

a = tamanho da amostra (número de seqüências) = 4

sob neutralidade $\pi = \theta_\pi = 4N_e\mu$

Entretanto, mesmo diante das estimativas para a diversidade nucleotídica de um determinado segmento de DNA ou para todo o genoma, não é de total certeza que os polimorfismos encontrados são de origem exclusivamente neutra. Desta forma, foram idealizados testes estatísticos capazes de inferir a neutralidade dos alelos presentes na população. Tajima (1989) desenvolveu um teste estatístico, o teste D, para testar a hipótese de que todas as mutações são seletivamente neutras (KIMURA, 1985). O teste D é baseado nas diferenças entre o número de sítios segregantes e a diversidade nucleotídica e é assim representado:

$$D = \frac{d}{\sqrt{\hat{V}(d)}} = \frac{\hat{k} - \frac{S}{a_1}}{\sqrt{e_1 S + e_2 S(S-1)}},$$

assumindo que:

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}, \qquad e_1 = \frac{c_1}{a_1}, \qquad e_2 = \frac{c_2}{a_1^2 + a_2},$$

onde:

S é o número total de sítios segregantes, n é o número de seqüências nucleotídicas e k é a média do número de diferenças nucleotídicas entre um par de seqüências.

O valor de D está atrelado a uma significância estatística, p, onde os limites de confiança são obtidos assumindo que D segue a distribuição beta. O intervalo de confiança sobre o valor de D é importante, pois é ele que indica se a seleção para o segmento analisado está ocorrendo. Valores de D iguais a zero sugere aderência total à teoria neutra; D < 0 sugere diversidade reduzida, ou seja, seleção, e D > 0 sugere seleção balanceada.

Com o advento da genômica e a busca incessante para a descoberta de novos genes, grandes projetos incluem como ferramentas a construção de banco de dados de seqüências para o organismo em estudo. A disponibilidade crescente de seqüências parciais de genes a partir dos projetos genômicos e funcionais permite hoje a identificação de polimorfismos de base individual (SNP) responsáveis ou associados a QTN (*Quantitative Trait Nucleotide*). Por exemplo, em milho, SNPs no gene *dwarf8* foram associados com florescimento (THORNSBERRY *et al*., 2001). Estas associações por sua vez podem resultar no desenvolvimento de marcadores para seleção assistida em plantas baseados na variabilidade de seqüência de genes e não apenas em marcadores microssatélites a eles ligados (MORGANTE & SALAMINI, 2003). O uso do polimorfismo de genes como marcadores para seleção de árvores permite o estabelecimento de relações diretas entre a variabilidade de

seqüência destes genes e a variabilidade fenotípica observada. A grande vantagem desta abordagem em árvores é que a questão potencialmente limitante de equilíbrio de ligação gamética entre alelos de marcadores moleculares e alelos de genes ligados passa a não ser relevante. Além disso, esta abordagem permite a análise direta de bancos de germoplasma e coleções de clones elite detalhadamente caracterizados (GRATTAPAGLIA, 2003).

Explorando uma ampla base de dados de EST (Expressed Sequence Tag) de eucalipto geradas em 2002/2005, no âmbito do projeto Genolyptus, e o seqüenciamento de amplicons gerados de indivíduos tomados ao acaso de três espécies do gênero *Eucalyptus*, esta dissertação é centrada na análise da diversidade nucleotídica inter e intra-específica de dois genes que codificam para duas enzimas consideradas chaves na via de biossíntese de lignina, *4cl* e *ccoaomt*.

2. Hipótese

Existe ampla variação nucleotídica inter e intra específicas para o gênero *Eucalyptus spp.* e para as espécies *E. grandis*, *E. globulus* e *E. urophylla* nos genes que codificam para as enzimas CCoAOMT e 4CL.

3. Objetivos

3.1 Objetivo geral

Caracterização da diversidade nucleotídica existente nas seqüências de dois genes de lignificação, *4cl* (4-coumarato coenzima A ligase) e *ccoaomt* (Cafeoil-CoA *o*-metiltransferase), em nível intraespecífico em *Eucalyptus grandis*, *E. urophylla* e *E. globulus* e em nível interespecífico entre espécies comercias de *Eucalyptus.*

3.2 Objetivos específicos

1. Mineração eletrônica do banco de dados de EST (Expressed Sequence Tag) do projeto Genolyptus em busca de todas as seqüências disponíveis para os genes que codificam para as enzimas 4CL e CCoAOMT, enzimas chaves da via de biossíntese de monolignóis;

2. Seqüenciamento completo de clones de cDNA presentes no banco de dados do Genolyptus, e análises de comparação eletrônica com seqüências do GenBank;

3. Alinhamento de seqüências, identificação de clones potencialmente *full length* e detecção de SNP *in silico*;

4. Desenho de iniciadores para amplificação de seqüências exônicas e intrônicas destes genes cobrindo regiões de ocorrência de SNPs e buscando cobrir a maior amplitude possível do gene;

5. Amplificação de seqüências dos dois genes em uma coleção de 50 indivíduos de cada uma de 3 espécies de *Eucalyptus* (*E. grandis*, *E. globulus* e *E. urophylla*) comercialmente utilizadas no mundo;

6. Seqüenciamento e montagem de um banco de dados de todas as seqüências destes genes e análise detalhada da diversidade nucleotídica em regiões de exons e introns;

7. Comparação das posições de SNPs entre análises *in silico* e por resseqüenciamento em relação a estimativas de diversidade nucleotídica, heterozigosidade esperada por nucleotídeo (SNPs) e de haplótipos de SNPs, estimativa da extensão do desequilíbrio de ligação, bem como a ocorrência de indels;

8. Identificação, isolamento e seqüenciamento do gene completo que codifica para a enzima 4CL pela construção de uma biblioteca "shotgun" de um clone BAC de *E. grandis* visando gerar informação de seqüência para futuros estudos detalhados da variabilidade do gene.

4. Material e Métodos

4.1 Banco de dados de EST

No âmbito do projeto Genolyptus foram construídas 13 bibliotecas de cDNA de diferentes espécies de *Eucalyptus* a partir de amostras de RNA obtidas de diferentes estágios de desenvolvimento e tecidos, como plântula, folha adulta e xilema. As bibliotecas foram construídas pelos vários laboratórios colaboradores do Projeto, utilizando kit *Superscript Plasmid System with Gateway Tecnology for cDNA Synthesis and Cloning* da Invitrogen Life Technology, seguindo as recomendações do fabricante.

Pouco mais de 110.000 seqüências foram geradas a partir do terminal 5' do inserto, baseando no método de Sanger *et al.* (1977). Todas as seqüências foram analisadas quanto a sua qualidade para depois então fazer parte do banco de dados. Remoção de seqüências de vetor, tamanho mínimo de 250 pares de base seqüenciadas com valor de qualidade (QV , Quality value) maior ou igual a 20 avaliada pelo software Phred (EWING *et al.*, 1998; EWING & GREEN, 1998) são alguns dos fatores para a validação da seqüência e inclusão no banco de dados.

O banco de dados foi acessado diretamente pela Universidade Católica de Brasília – UCB pelo endereço eletrônico http://genoma.ucb.br/SistemaGenoma/, onde todas as seqüências geradas foram depositadas e analisadas quanto aos critérios anteriormente citados.

Na central de bioinformática foi realizada uma busca de seqüências que continham identidade e similaridade com as seqüências dos genes *ccoaomt* e *4cl*. Essa identidade foi buscada por meio do *software* BLAST (ALTSCHUL *et al*., 1990; http://www.ncbi.nlm.nih.gov/blast), de utilização pública, e permitia a comparação das seqüências do banco de dados com as seqüências dos genes, relacionando quais seqüências eram afins, e, conseqüentemente, quais clones deveriam ser pinçados para posterior análise. Essa comparação abrangia tanto a seqüência nucleotídica, realizada pelo banco de dados do GenBank, quanto a sua tradução, realizada pelo banco de dados SwissProt

(http://www.expasy.ch), o que trazia maior confiabilidade aos resultados. Uma relação de potenciais cDNA completos (*full-length*), isto é, contendo integralmente a região codificadora dos mRNAs foi organizada e os clones de cDNA foram identificados por possuírem, potencialmente, as seqüências completas. Utilizando ferramentas de bioinformática, os representantes destes clones *full-length* foram selecionados para o seqüenciamento integral. A identificação dos clones potencialmente *full-length* foi realizada utilizando um script desenvolvido no Sistema Genoma que buscava o códon de iniciação de transcrição na seqüência de *Eucalyptus* pela comparação com o códon de iniciação do gene em *Arabidopsis*.

4.2 Reação de PCR para verificar tamanho aproximado dos clones de cDNA selecionados e seqüenciamento dos potenciais clones *full-length*

Clones foram selecionados por meio de análise de bioinformática para os dois genes em estudo, *ccoaomt* e *4cl*, dentre seqüências de várias bibliotecas de cDNA de espécies e tecidos diferentes de *Eucalyptus*. A reação foi realizada em volume final de 25 µl contendo 1,5 µl de DNA (aprox. 20ng/µl), 2,5 µl de Tampão IB 10X (Phoneutria – pht), 4 µl de dNTP 2,5 mM, 1 µl de $MgCl_2$ 50 mM, 100 ng de BSA, 2,5 µl de cada um dos primers a 3,2 pmoles.μl^{-1}, 0,3 µl de Taq DNA polimerase 5U/µl (Phoneutria – pht) e 9,7 µl de água destilada. As condições para a reação de PCR foram: desnaturação inicial a 96 ºC por 2 min, 25 ciclos de desnaturação a 96 ºC por 45 s, anelamento dos primers a 50 ºC por 30s, extensão a 60 ºC por 3 min e extensão final a 60 ºC por 4 min. Os produtos amplificados foram visualizados em gel de agarose 1% corado com brometo de etídio, utilizando marcador molecular *1 Kb DNA ladder* como padrão de comparação.

Para o seqüenciamento dos clones, foi utilizado o kit *Big Dye™ Terminator v3.0 Ready Reaction Cycle Sequencing* e as reações foram preparadas da seguinte forma: 2 µl de DNA plasmidial (aproximadamente de 100-400 ng.μl^{-1}), 1 µl do primer (T7 e GLP) a 3,2 µM,

3 µl de *Save Money*, 1 µl de Big Dye™ e 3 µl de água *Milli-Q* autoclavada, para um volume final de 10 µl. As reações foram amplificadas em termociclador com o seguinte programa: 96 °C por 2 min, 25 ciclos de 96 °C por 45 s, 50 °C por 30 s e 60 °C por 3 min. Após os 25 ciclos, incubar a 60 °C por 4 min.

Após a reação de seqüenciamento, os produtos foram purificados, desnaturados e submetidos à eletroforese em seqüenciador 3700 ou 377 da Applied Biosystems.

4.3 Desenho de iniciadores

Oligonucleotídeos de iniciação foram desenhados com o intuito de amplificar diversas regiões do gene de interesse. Dois grupos de primers foram construídos, de forma que os diferentes pares formassem fragmentos com regiões de sobreposição para a formação de um contíguo, gerando assim a seqüência do gene completo.

O processo de desenho dos primers foi realizado a partir de seqüências de domínio público disponibilizadas no banco de dados GenBank e com o auxílio de ESTs gerados pelo Projeto Genolyptus. Selecionadas as seqüências que indicaram homologia com o gene de interesse, as seqüências foram então alinhadas utilizando o software *ClustalW* (HIGGINS *et al.*, 1994) e, considerando as regiões de maior identidade entre as diversas seqüências, primers foram desenhados a fim de que fosse possível o anelamento e extensão do segmento esperado em todos os indivíduos das três espécies de eucalipto analisadas. Com o auxílio do programa *Primer3* (ROZEN & SKALETSKY, 2000), disponível na Internet pelo endereço eletrônico http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi, vários primers foram gerados e escolheu-se o par que gerasse um fragmento de aproximadamente 400 pares de base. Esse tamanho estimado foi escolhido pelo fato de que as seqüências utilizadas para a construção dos primers eram ESTs, ou seja, só continham regiões codificantes. Como o nosso objetivo era amplificação em DNA genômico, tínhamos ainda que supor a existência de regiões

intrônicas, de diversos tamanhos, que afetariam na extensão final do fragmento amplificado. Conseqüentemente, com um fragmento de maior dimensão, o risco seria muito grande de não se obter, por fim, o seqüenciamento completo do segmento, impedindo a formação das sobreposições e construção dos contíguos do gene.

4.4 Material Genético

Para o presente estudo foram selecionadas três espécies de eucalipto entre as mais plantadas no mundo e com ampla variabilidade fenotípica para características da madeira: *Eucalyptus grandis*, *Eucalyptus globulus* e *Eucalyptus urophylla*. Foram selecionadas 50 árvores de cada espécie, tomadas ao acaso a partir de progênies de polinização aberta, inicialmente estabelecidas por sementes coletadas na Austrália, de uma mesma procedência, e plantadas no Brasil pelas empresas Suzano Celulose e Papel e Klabin S.A.. Para *E.grandis*, foram tomadas plantas a partir de duas procedências distintas, a *E. grandis* Pine Creek (sudeste da Austrália) e a *E. grandis* Atherton (nordeste da Austrália), as quais são as mais utilizadas em programas de melhoramento no Brasil.

As amostras foram enviadas na forma de folhas, devidamente embaladas e identificadas. Após o recebimento, as folhas foram estocadas a -20 °C seguida de extração do DNA, conforme protocolo a seguir.

4.5 Extração e quantificação do DNA genômico total de *Eucalyptus*

A extração do DNA das várias espécies de *Eucalyptus* foi realizada a partir de tecido foliar fresco e congelado. A extração foi realizada por meio de beeds de cerâmica de acordo com o protocolo de Doyle & Doyle (1991), com adaptações de Ferreira & Grattapaglia (1998), utilizando 10 a 20 mg de tecido por amostra. O tecido foi condicionado em tubos (2 ml) com tampa de rosca contendo a esfera menor das *beeds* no fundo do tubo e a esfera maior

colocada após o tecido. Para o rompimento das paredes e membranas celulares, foram adicionados 700 µl do Tampão de Extração de DNA previamente aquecido a 65 ºC e o tubo submetido a uma agitação constante no agitador FastPrep por 20 s. Após maceração por agitação física, foi iniciada a etapa de incubação em banho-maria a 65 ºC por 1 h, com leve agitação de 1 min a cada 10 min de incubação. Terminado o tempo de incubação, os tubos foram resfriados em temperatura ambiente e, em capela de exaustão, foi realizado o processo de extração da fase orgânica das amostras com a adição de 600 µl de CIA (Clorofórmio:Álcool isoamílico 24:1) em cada tubo, seguidos de agitação leve por 5 min e centrifugação a 12000 rpm por 5 min. A fase sobrenadante das amostras foi recuperada e transferida para um novo tubo. Os ácidos nucléicos foram precipitados com a adição de 360 µl de isopropanol absoluto gelado, seguido de agitação leve e incubação *overnight* a -20 ºC. Os tubos foram centrifugados a 12000 rpm por 15 min e o sobrenadante descartado. É possível a formação de um *pellet* visível após a centrifugação. O processo de lavagem do pellet foi realizado duas vezes com 1 ml de etanol 70%, deixando-o imerso por 10 min. Em seguida, o pellet foi ressuspendido em 500 µl de Solução NaCl 2 M e os tubos incubados a 4 ºC por 40 min. As amostras foram centrifugadas a 12000 rpm por 10 min e o sobrenadante foi transferido para novos tubos. Foram acrescentados ao sobrenadante 330 µl de isopropanol absoluto gelado, precipitando novamente os ácidos nucléicos. Os tubos foram incubados a -20 ºC overnight e centrifugados a 12000 rpm por 15 min. Ao pellet formado, foram realizadas duas lavagens com 1 ml de etanol 70% por 10 min e uma lavagem com etanol absoluto gelado por 10 min. Após o descarte do sobrenadante, os tubos foram centrifugados no concentrador Speed Vacum por 15 minutos a 30 ºC para a completa secagem do pellet. Após secos, foram ressuspendidos em 100 µl de Tampão TE acrescido de RNAse (1 mg/ml) e incubados a 37 ºC por 40 min. A quantificação do DNA extraído foi realizada por eletroforese em gel de agarose 1% corado com brometo de etídio, utilizando como padrão de comparação, DNA do fago

lambda em concentração conhecida.

4.6 Amplificação de segmentos dos genes

Utilizando iniciadores desenhados por Gion *et al.* (2000), fragmentos do gene *ccoaomt* e *4cl* foram amplificados utilizando DNA genômico dos 50 indivíduos de cada uma das três espécies de eucalipto. As reações de amplificação seguiram o seguinte protocolo: aproximadamente 15 ng de DNA genômico, 1,675 µl de Tris-HCl 1 M, 1 µl de BSA a 1 ng/µl, 1 µl de β-mercaptoetanol a 5%, 1 µl de $MgCl_2$ a 50 mM, 4 µl de Acetato de amônio a 100 mM, 2 µl de dNTP a 2,5 mM cada, 0,2 µl de Taq DNA polimerase (Phoneutria – pht) a 5 U/µl, 1 µl de cada um dos primers a 5 µM e água *Milli-Q* autoclavada a q.s.p 25 µl.

As condições de PCR, para ambos os genes, foram as seguintes: 95 °C por 5 min, 30 ciclos de 94 °C por 30 s, 64 °C para *4cl* e 68 °C para *ccoaomt* por 45 s, 72 °C por 1 min 30 s, extensão final de 72 °C por 20 min, seguida de 12 °C por tempo indeterminado.

Para os iniciadores desenhados neste trabalho para o gene *4cl*, o protocolo foi o seguinte: 20 ng de DNA genômico, 2,5 µl de tampão 10 X IIB (Phoneutria – pht), 0,25 µl de cada um dos primers a 10 µM, 2 µl de dNTP a 2,5 µM cada, 1 µl de $MgCl_2$ a 50 mM, 0,2 µl de Taq DNA polimerase (Phoneutria – pht) e água *Milli-Q* autoclavada a q.s.p 25 µl.

As condições de PCR foram: 94 °C por 5 min, 30 ciclos de 94 °C por 30 s, 62 °C por 45 s, 72 °C por 2 min 30 s, extensão final de 72 °C por 25 min seguida de 12 °C por tempo indeterminado.

Os produtos foram analisados por eletroforese em gel de agarose a 1% corado com brometo de etídeo, e comparados com DNA de fago *lambda* em concentração e tamanhos conhecidos.

4.7 Seqüenciamento de produto de PCR e análise

Fragmentos amplificados com primers específicos para cada gene foram gerados e seqüenciados tanto a partir do terminal 3' quanto do 5'. Quando não eram obtidas seqüências de qualidade esperada, foi realizado o resseqüenciamento. As reações de seqüenciamento foram realizadas com dois kits: o kit *DYEnamic™ ET Terminator Cycle Sequencing* (Pharmacia Biotech, EUA) e o kit *Big Dye™ Terminator v3.0 Ready Reaction Cycle Sequencing*. As reações foram realizadas da seguinte forma: Para o kit Big Dye™: 1 µl do produto de PCR sem purificação, 2 µl de 5X *Sequencing Buffer*, também conhecido como *Save Money*, 2 µl de *Big Dye™ Terminator v3.0,* 1 µl do primer utilizado na amplificação na concentração de 3,2 µM e água *Milli-Q* autoclavada em q.s.p para 10 µl. As reações foram realizadas em termociclador *Applied Biosystems GeneAmp* PCR *System* 9700 (Applied Biosystems, EUA) sob o seguinte programa: 96 °C por 1 min, 30 ciclos de 96 °C por 10 s, 50 °C por 5 s e 60 °C por 4 min, seguido de incubação a 4 °C por tempo indeterminado.

Para o kit *DYEnamic™ ET Terminator*, as reações foram preparadas da seguinte forma: 1 a 3 µl do produto de PCR sem purificação, 1 µl do primer a 3,2 µM, 2 µl de *DYEnamic™ ET Terminator* e água Milli-Q autoclavada em q.s.p 10 µl. As reações foram realizadas em termociclador *Applied Biosystems GeneAmp* PCR *System* 9700 (Applied Biosystems, EUA) sob o seguinte programa: 30 ciclos de 95 °C por 20 s, 50 °C por 15 s e 60 °C por 1 min, seguido de incubação a 4 °C por tempo indeterminado.

A purificação dos 10 µl de reação, para ambos os kits, envolveu a lavagem das amostras com 2,5 µl de EDTA a 125 mM seguido do acréscimo de 25 µl de etanol absoluto a temperatura ambiente. Misturados os volumes com o auxilio de uma ponteira, o material foi incubado por 30min à temperatura ambiente. Aguardado o tempo de incubação, o material foi centrifugado a 12000 rpm por 30 min. O sobrenadante foi descartado e 70 µl de etanol 70% foram adicionados para a lavagem do pellet. O material foi novamente centrifugado a 12000

rpm por 15 min, e, após o descarte do sobrenadante, o pellet foi submetido ao Speed Vaccum (Eppendorf concentrador 5301) por 10 min para secar o pellet. A ressuspensão do pellet foi diferenciada de acordo com o equipamento a ser utilizado para o seqüenciamento. 7 µl de L*oading buffer* foram adicionados ao pellet para seqüenciamento em seqüenciador automático 377 DNA Sequencer (Applied Biosystems ABI Prism). No caso do seqüenciador automático 3100 DNA Sequencer (Applied Biosystems ABI Prism), o pellet foi ressuspendido em 10 µl de formamida (Hi-Di) e antes de serem acoplados e injetados no seqüenciador, o material foi submetido a um processo de desnaturação a 95 °C por 5 min, seguido de 3 min de incubação em gelo.

As seqüências geradas foram analisadas no software SeqScape 2.1 (Applied Biosystems), com o seguinte protocolo de análise: mistura de bases a partir de 30% de altura do pico e no máximo 20% do total de bases, aceitação de seqüências com no máximo de 10% de bases identificadas com N, tamanho mínimo de 50 pb e score de base de no mínimo PHRED 15. As seqüências também foram avaliadas visualmente quanto à qualidade e polimorfismo, corrigindo possíveis erros de denominação de base e certificando as bases denominadas com qualidade baixa.

As seqüências foram comparadas com uma seqüência de referência (Apêndice A) gerada a partir do consenso de segmentos amplificados com primers específicos para o gene de interesse, bem como de seqüências depositadas no banco de dados do Genolyptus e no GenBank. Na formação da seqüência de referência, as posições que apresentavam polimorfismo foram analisadas quanto à freqüência dos alelos (base) e aquele que apresentou maior freqüência foi o escolhido para compor a seqüência de referência. Esse consenso foi também comparado com banco de dados de proteínas (SwissProt) e, em conjunto com dados de seqüência de aminoácidos e por comparação com seqüência completa do gene em *Arabidopsis*, foram identificadas regiões de exon e intron na seqüência de referência.

Todos os segmentos gerados foram comparados com essa seqüência de referência com o objetivo de se obter a seqüência completa para cada um dos indivíduos analisados.

Os seqüenciamentos de cada indivíduo se estenderam ao número de vezes suficientes para obter o segmento compreendido pela seqüência de referência.

4.8 Análise de diversidade de seqüência

As seqüências geradas foram analisadas quanto à qualidade das bases, alinhamento e edição pelo programa SeqScape da Applied Biosystems. Nesse mesmo programa, foram detectados os polimorfismos e sítios potencialmente em heterozigose, os quais foram confirmados por inspeção visual. Todos os *singletons* foram confirmados mas não fizeram parte dos cálculos estatísticos.

Utilizando o software DnaSP 4.10 (www.ub.es/dnasp) (ROZAS *et al.*, 2003) foram obtidas as seguintes estimativas:

a) Número total de sítios segregantes (S);

b) Diversidade nucleotídica π (NEI & LI, 1979; NEI, 1987), o número médio de diferenças nucleotídicas (SNPs, polimorfismos de bases individuais) por sítio entre duas seqüências;

c) Diversidade nucleotídica sob a hipótese de neutralidade do polimorfismo $\pi = \theta_\pi = 4N_e\mu$, onde N_e é o tamanho efetivo populacional e μ a taxa de mutação por nucleotídeo por geração (NEI, 1987). Ou seja, pela teoria *neutra* da evolução molecular, o nível de polimorfismo θ é proporcional ao tamanho efetivo populacional e à taxa de mutação.

DnaSP 4.10 também foi utilizado para calcular:

d) o teste estatístico D de Tajima (TAJIMA, 1989), para testar a aderência dos dados observados à teoria neutra da evolução molecular (KIMURA, 1985). Este teste é baseado no fato de que, sob o modelo neutro, as estimativas do número de sítios segregantes e do número

médio de diferenças nucleotídicas são correlacionados. O software DnaSP calcula o intervalo de confiança de D (teste bilateral).

O desequilíbrio de ligação foi estimado por r^2 e D', e sua significância foi calculada pelo teste exato de Fisher, aplicando a correção de Bonferroni para múltiplas comparações, utilizando o software DnaSP 4.10. As análises incluíram somente SNPs com freqüência alélica maior ou igual a 5% e que tinham, portanto, poder suficiente para detectar desequilíbrio de ligação.

Haplótipos foram determinados utilizando o software PHASE versão 2.1.1 (STEPHENS *et al.*, 2001; STEPHENS & DONNELLY, 2003) e reconstruídas as seqüências com o número total de haplótipos esperados.

4.9 Identificação de clone BAC com o gene alvo

No âmbito do projeto Genolyptus foi construída uma biblioteca genômica de BAC de um indivíduo da espécie *Eucalyptus grandis* (S. Brommonschenkel com. pess.). Essa biblioteca composta por 20160 clones com tamanho médio de 120 a 150 Kb fornece uma cobertura estimada de 4 vezes do genoma do eucalipto. Esta biblioteca foi submetida a uma triagem via PCR com o objetivo de identificar um ou mais clones BAC contendo o gene alvo para posteriores estudos da estrutura gênica completa (promotor, região codificadora e introns).

Uma estratégia hierárquica de "pools" de clones BAC (reuniões de clones) foi utilizada para rapidamente alcançar o gene alvo. Os 20.160 clones BAC foram organizados em 210 microplacas de 96 poços. Mini-preparações de DNA de BACs foram realizadas em "pools" de 96 clones de uma mesma placa. As 210 amostras de DNA foram então organizadas em 35 "superpools" de 6 placas ("pools") de 96 totalizando 576 clones por "superpool" (Figura 10). As extrações de DNA dos clones BAC foram realizadas para cada "pool" de 96

clones resultando em 210 amostras de DNA a serem utilizadas na triagem para identificação do gene alvo.

A construção dos superpools iniciou-se na fase de seleção dos clones para a formação de uma placa de 96 clones. Após a construção da biblioteca, os clones foram selecionados e alocados em microplacas de polipropileno de 250 µl com fundo "U" contendo meio LB e antibiótico Clorafenicol a 12,5 $mg.ml^{-1}$. Após o período de crescimento, procedimento corriqueiro para a realização da minipreparação de DNA, foi realizado um plaqueamento em placa de petri contendo meio LB sólido acrescido do antibiótico. Esse plaqueamento foi realizado de forma identificada, onde cada inóculo de cada poço da microplaca constituiu um novo clone. Transcorrido o tempo de crescimento em estufa a 37 ºC, com o auxílio de uma lâmina de preparações histológicas e a adição de 500 µl de meio LB líquido, todos os clones foram misturados e a solução serviu de pré-inóculo para um inóculo maior. Esse inóculo foi então submetido a uma miniprepação para a extração do DNA de BAC de todos os clones em conjunto. Dessa mesma forma, todas as 210 placas foram submetidas ao processo. Essa extração resultou na formação de um "pool". Subseqüentemente, os 210 pools foram distribuídos em três microplacas de 96 poços. Esse arranjo em três placas foi novamente organizado para somente uma placa (Figura 10). Cada uma das três placas foi dividida pela metade (1 ao 6 e 7 ao 12). Os seis primeiros "pools" de cada linha constituíram um "superpool", assim como os demais seis "pools". Assim, foram criados 35 "superpools", arranjados em uma microplaca de 96 poços conforme ilustra a figura 10.

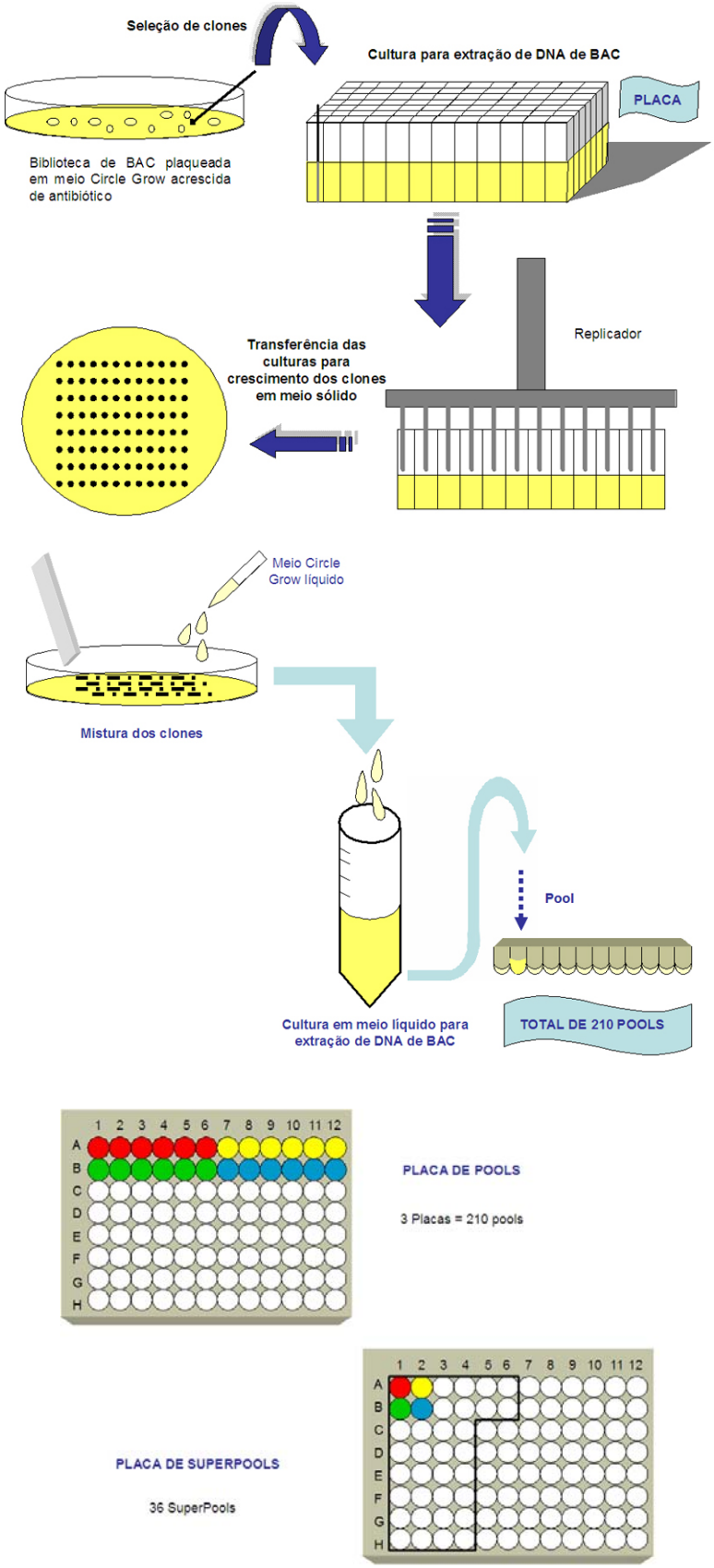


Figura 10. Formação dos pools e superpools para o screening da biblioteca para os genes *ccoaomt* e *4cl*.

4.10 Triagem (*screening*) dos clones de BAC

Inicialmente, foram realizadas 35 reações de PCR e mais os controles positivo e negativo, utilizando os 35 superpools construídos como DNA molde e primers específicos para o gene alvo. Após visualização do produto amplificado por eletroforese em gel de agarose 1 % corado com brometo de etídio, descobriu-se qual o "superpool" que continha o clone desejado. Descoberto o "superpool", esse foi desmembrado em "pools" para continuar a busca pelo clone alvo. Dessa forma, foram realizadas mais 6 reações de PCR, com controle positivo e negativo, cada uma com um dos "pools" correspondentes ao "superpool". Verificado o produto amplificado em gel de agarose, identificou-se em qual "pool" estava o clone alvo. Do mesmo modo, o "pool" foi aberto e a placa correspondente ao "pool" foi analisada. Assim, 96 reações foram realizadas e verificado, pontualmente, qual clone continha o gene alvo.

O protocolo para a reação de PCR para o "screening" dos "superpools", pools e placas de BAC foi o seguinte: 2 µl do DNA do clone BAC, 2,5 µl de tampão *pht* 10X IIB (Phoneutria – pht), 1 µl de cada um dos primers a 10 µM, 4 µl de dNTPs a 2,5 mM, 1 µl de BSA a 2,5 $mg.ml^{-1}$, 1 µl de $MgCl_2$ a 50 mM para o gene de *4cl* e 1,5 µl para o gene *ccoaomt*, 0,5 µl Taq DNA polimerase a 5 $U.\mu l^{-1}$ (Phoneutria – pht) e água Milli-Q autoclavada em q.s.p 25 µl.

As condições para a reação foram as seguintes: 95 °C por 4 min, 30 ciclos de 92 °C por 45 s, 68 °C para *ccoaomt* e 65 °C para *4cl* por 45 s, 72 °C por 1 min 30 s, extensão final de 72 °C por 10 min e incubação a 12 °C por tempo indeterminado. Os produtos da reação foram analisados por eletroforese em gel de agarose 1%, corado com brometo de etídeo e comparados com marcador molecular *1 Kb plus DNA Ladder* (Invitrogen, Carlsbad, CA – EUA).

4.11 Isolamento de DNA de BAC por lise alcalina

Para os clones BAC selecionados foi organizada uma microplaca de 96 colônias em meio LB líquido e antibiótico Clorafenicol na concentração adequada de 12,5 $\mu g.ml^{-1}$. Após o processo de crescimento, foi adicionado glicerol 50% em volumes suficientes para o estoque das células bacterianas a -20 ºC.

Para o isolamento do DNA de BAC, inicialmente foi realizada uma pré-cultura dos clones. Em uma câmara de fluxo laminar, foi depositado em cada poço de uma placa de 96 poços, 100 µl de meio Circle Grow (Bio 101 Inc., Carlsbad, CA – EUA) contendo Clorafenicol na concentração final de 12,5 $\mu g.ml^{-1}$. As culturas foram inoculadas utilizando um replicador de colônias com 96 pinos, a placa foi selada com adesivo e incubada a 37 ºC por 15 h. Depois de crescidas, as bactérias foram inoculadas em placa tipo *deep well* com 1,3 ml de meio Circle Grow e Clorafenicol (12,5 $\mu g.ml^{-1}$), com dois toques do replicador a partir do pré-inóculo. A placa foi selada, o adesivo perfurado com o auxílio de um palito no local relacionado ao poço e incubada por 16 h a 37 ºC sob agitação constante de 250 rpm.

Transcorrido o crescimento bacteriano, a placa foi centrifugada a 3700 rpm por 10 min para sedimentar as células. O sobrenadante foi descartado e a placa invertida em papel absorvente por 5 min para a remoção de resíduos de meio Circle Grow. Foram adicionados 100 µl de Solução I (Tris-HCl 50 mM pH 7,4; EDTA 10 mM pH 8,0) acrescida de RNAse com concentração final de 100 $\mu g.ml^{-1}$ por poço da placa, a qual foi selada e agitada em shaker por 10 min a 380 rpm para a ressuspensão do pellet. O adesivo foi retirado e 80 µl da solução de células foram transferidos para uma placa de 250 µl de polipropileno de fundo redondo (tipo Elisa). 80 µl de Solução II (SDS 1%; NaOH 0,2 M) foram adicionados à solução de células de cada poço, a placa foi novamente selada com adesivo novo, invertida 20 vezes para homogeneizar as soluções, incubada a temperatura ambiente por 5 min e centrifugada a 3700 rpm por 1 min, para remover todo o lisado do adesivo e não causar

problemas de contaminação na hora de sua retirada. Após, foram adicionados à suspensão 80 µl de Solução III (KOAc 3 M) por poço da placa, a qual foi selada, seguida de inversões e incubação do neutralizado por 10 min em gelo. A placa foi centrifugada a 3700 rpm por 30 min a 4 ºC.

Enquanto realizava-se a centrifugação, utilizando ligas elásticas e fita adesiva, uma placa *Millipore* (MAGV N22) era fixada no topo de uma microplaca de 96 poços de fundo "V" de 250 µl de polipropileno, verificando o alinhamento dos poços. Foram transferidos 180 µl do sobrenadante contido na placa de fundo redondo para a placa *Millipore* e o aparato foi centrifugado a 3000 rpm por 6 min. A placa *Millipore* foi removida e ao filtrado resultante na placa fundo "V" acrescentados 180 µl de LiCl 5 M seguido de incubação por 10 min à temperatura ambiente. A placa foi centrifugada por 10 min a 3700 rpm, o sobrenadante transferido para uma nova microplaca fundo "V" e adicionados 110 µl de isopropanol em cada poço. A placa foi selada com um novo adesivo, homogeneizada por inversões (10 vezes), incubada à temperatura ambiente por 15 min e centrifugada a 3700 rpm por 30 min a 4 ºC. Após, o sobrenadante foi descartado por inversão e, aos poços, adicionados 200 µl de etanol 70%, por duas vezes, para a lavagem dos pellets. A placa foi centrifugada a 3700 rpm por 5 min descartando o sobrenadante e, com a placa invertida sobre papel absorvente, esta foi novamente centrifugada a 900 rpm por 1 min para retirar o excesso de etanol. Coberta com papel de filtro, a placa foi incubada em estufa a 37 ºC por 1 h para secar o DNA de BAC. O DNA foi ressuspenso com a adição de 25 µl de água *Milli-Q* autoclavada, seguida de breve agitação e incubação a 4 ºC por 12 h, para depois ser estocada a -20 ºC.

A quantificação do DNA de BAC foi realizada por eletroforese em gel de agarose 0,8% corado com brometo de etídio e utilizando DNA de fago *lambda*, em concentração conhecida, como padrão de comparação.

4.12 Isolamento de DNA de plasmídeo de biblioteca *shotgun* por lise alcalina

Uma biblioteca genômica shotgun do clone BAC, selecionado por conter o gene alvo para a enzima 4CL, foi produzida na UFV pelo Prof. Sergio Brommonschenkel e sua equipe. Subclones do BAC com tamanho médio de 1 Kb foram acondicionados em microplacas de 96 poços de 250 µl de polipropileno com fundo "U" contendo meio LB líquido e antibiótico Ampicilina na concentração adequada de 100 $\mu g.ml^{-1}$. Após o processo de crescimento, foi adicionado glicerol 16% em volumes suficientes para o estoque das células bacterianas a -80 ºC.

Para o isolamento do DNA do plasmídeo, inicialmente foi realizado inóculo em placa tipo *deep well* com 1,5 ml de meio Circle Grow e Ampicilina (100 $\mu g.ml^{-1}$), com dois toques do replicador a partir do pré-inóculo. A placa foi selada, o adesivo perfurado com o auxílio de um palito no local relacionado ao poço e incubada por 16 h a 37 ºC sob agitação constante de 250 rpm. Transcorrido o crescimento bacteriano, a placa foi centrifugada a 4000 rpm por 15 min para sedimentar as células. O sobrenadante foi descartado e a placa invertida em papel absorvente por 5 min para a remoção de resíduos de meio Circle Grow. Foram adicionados 240 µl de Solução I, os sedimentos ressuspendidos em vortex por 2 min e novamente centrifugados a 4000 rpm por 15 min. O sobrenadante foi descartado e a placa invertida em papel de filtro por 5 min para certificação da eliminação de qualquer resíduo de meio Circle Grow. Ao sedimento, foram adicionados 70 µl de Solução I (Glicose 20%; EDTA 500 mM pH 8,0; Tris-HCl 1 M pH 7,4) acrescida de RNAse com concentração final de 100 $\mu g.ml^{-1}$ por poço da placa, a qual foi selada e agitada em vortex por 5 min a 380 rpm para a ressuspensão do pellet. O adesivo foi retirado e 70 µl da solução de células foram transferidos para uma placa de 250 µl de polipropileno de fundo redondo (tipo Elisa). 70 µl de Solução II (SDS 1%; NaOH 0,2 M) preparada na hora foram adicionados à solução de células de cada poço, a placa foi novamente selada com adesivo novo, invertida 40 vezes para homogeneizar as soluções,

incubada a temperatura ambiente por 10 min e centrifugada a 4000 rpm por 1 min, para remover todo o lisado do adesivo e não causar problemas de contaminação na hora de sua retirada. Após, foram adicionados à suspensão 70 µl de Solução III (KOAc 3 M) gelada por poço da placa, a qual foi selada, seguida de 40 inversões e incubado o neutralizado por 10 min à temperatura ambiente. A placa foi então submetida a uma centrifugação a 4000 rpm por 1 min para retirar todo o neutralizado do adesivo. O adesivo foi então removido e a placa incubada a 90 ºC por 15 min. Transcorrido o tempo, a placa foi novamente selada e incubada por 10 min em gelo e centrifugada a 4000 rpm por 15 min à 4 ºC. Durante a centrifugação, o aparato contendo uma placa de filtro *Millipore* acoplada à uma microplaca de 96 poços de 250 µl de polipropileno com fundo "V" foi montado. Após a centrifugação, foi transferido à placa de filtro, 120 µl do sobrenadante e o aparato foi centrifugado a 4000 rpm por 6 min. A placa *Millipore* foi removida e ao filtrado resultante na placa fundo "V" acrescentados 110 µl de isopropanol à temperatura ambiente. A placa foi selada com um novo adesivo, homogeneizada por inversões (20 vezes) e centrifugada a 4000 rpm por 45 min a 4 ºC. Após, o sobrenadante foi descartado por inversão e, aos poços foram adicionados 200 µl de etanol 70% para a lavagem dos pellets. A placa foi centrifugada a 4000 rpm por 15 min descartando o sobrenadante e, com a placa invertida sobre papel absorvente, esta foi novamente centrifugada a 900 rpm por 30 s para retirar traços de etanol. Coberta com papel de filtro, a placa foi incubada em estufa a 37 ºC por 1 h para secar o DNA do plasmídeo. O DNA foi ressuspenso com a adição de 50 µl de TE autoclavado, seguida de breve agitação e incubação à temperatura ambiente por 12 h para depois ser estocada a -20 ºC.

4.13 Seqüenciamento da biblioteca *Shotgun*

As reações de seqüenciamento foram realizadas com kit de seqüenciamento *Big Dye*™ da seguinte forma: 1 a 3 µl de DNA, 1 µl de primer a 3,2 µM (T3 e T7), 2 µl de Tampão 2,5X

(200 mM Tris-HCl pH 9,0; 5 mM $MgCl_2$), 2 µl de *Big Dye*™, em uma reação de 10 µl de volume final. As amplificações foram realizadas em termociclador sob o seguinte programa: 96 °C por 2 min, 25 ciclos de 96 °C por 45 s, 50 °C por 30 s e 60 °C por 3 min. Após os 25 ciclos, incubar a 60 °C por 4 min e a 4 °C por 2 min.

Após a reação de seqüenciamento, os produtos foram purificados, desnaturados e submetidos à eletroforese em seqüenciador 3700 ou 377 da Applied Biosystems.

5. Resultados

5.1 Mineração do banco de dados de EST

Foi realizada uma análise inicial do banco de dados do projeto Genolyptus (Tabela 1). Tanto o tamanho médio dos clones, determinados a partir da amplificação do inserto de cDNA inserido no plasmídeo, bem como a porcentagem que estava sendo seqüenciada foi determinada. Foram observados insertos que variavam de 300 pb a 2000 pb (Figura 11). O banco de dados contém seqüências de aproximadamente 400 pb, sugerindo que ambas as extremidades do inserto poderiam ser seqüenciadas a fim de obter a seqüência completa do inserto.

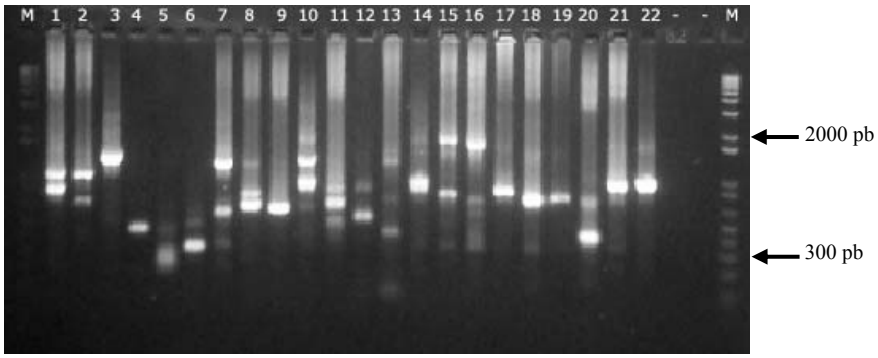


Figura 11. Análise, em gel de agarose, do tamanho médio dos insertos de cDNA clonados em plasmídeo na construção da biblioteca. 1-22, clones selecionados *in silico*; M, marcador 1 Kb Plus DNA ladder (Invitrogen, Carlsbad, CA).

Tabela 1. Relação das bibliotecas constituintes do banco de dados do Projeto Genolyptus.

Espécie	Material vegetal	Material genético	Sigla da Biblioteca	Clones sequenciados	Seqüências válidas*
	Folhas maduras	cDNA	EUGR-ML	8.067	5399 (66,93%)
	Folhas jovens	cDNA	EUGR-YL	1.344	584 (43,45%)
	Xilema	cDNA	EUGR-XY	1038	641 (61,75%)
	Flores abertas	cDNA	EUGR-FL	2.788	285 (10,22%)
E. grandis	Plântulas inteiras	cDNA	EUGR-SE	14.502	10.296 (70,99%)
	Plântulas tratadas	cDNA	EUGR-TS	14.871	10.275 (69,09%)
	Plântulas infectadas *Puccinia psidii*	cDNA	EUGR-PU	5.088	4.133 (81,23%)
	BAC	DNA	EUGR-BC	17.598	13027 (74,03%)
	Folha	DNA	EUGR-GE	9.896	6342 (68,43%)
E. globulus	Xilema	cDNA	EUGL-XY	18.051	13.549 (75,06%)
E. urophylla	Xilema	cDNA	EUUR-XY	9.901	6624 (66,90%)
E. pellita	Xilema	cDNA	EUPE-XY	11.343	8.380 (73,88%)
Eucalyptus sp.	Floema(mistura)	cDNA	EUSP-FX	14.614	10.434 (71,40%)
(mistura)	Xilema(mistura)	cDNA	EUSP-XX	96	13 (13,54%)
	Raízes(mistura)	cDNA	EUSP-RX	2.379	884 (37,15%)
				131.576	90.866

* São consideradas seqüências válidas aquelas que apresentaram no mínimo 250 pares de bases seqüenciados como um valor de qualidade maior ou igual a 20 avaliado pelo software PHRED

(1) Dados gerados pela página BIOFOCO de submissão de seqüências na UCB em maio de 2005.

Para ambos os genes uma busca foi realizada sobre os possíveis clones “full length” (comprimento total) presentes no banco. Foram selecionados do banco 11 clones para o gene *4cl* e 41 clones para *ccoaomt* e todos estes foram resseqüenciados, tanto no sentido 3’ quanto no 5’. As seqüências geradas foram reunidas em uma nova biblioteca no sistema genoma denominada de “full length”. Estas seqüências foram utilizadas e aproveitadas em outros estudos, inclusive para a montagem e complementação dos respectivos genes no banco de dados.

Buscas realizadas no banco de dados do Projeto Genolyptus identificaram várias ESTs (*Expressed Sequence Tags*) homólogas aos genes em estudo. A pesquisa baseou-se em identidade de uma determinada seqüência às seqüências do banco de dados.

Na página do banco de dados onde se realiza a pesquisa, é possível selecionar a forma como será realizada a pesquisa. Em um campo específico, pode-se selecionar que a pesquisa seja realizada em todas as mais de 131.000 seqüências depositadas no banco de dados ou, de forma mais ágil, selecionar a forma de busca por clusters, ou grupos de seqüências.

Cluster é o conjunto de seqüências que, reunidas, forma uma seqüência maior, um contig (Figura 12). A busca por clusters permite que, além de uma forma mais resumida e completa da busca, o resultado seja mais extenso, resultando no somatório das seqüências e não somente as seqüências em separado.

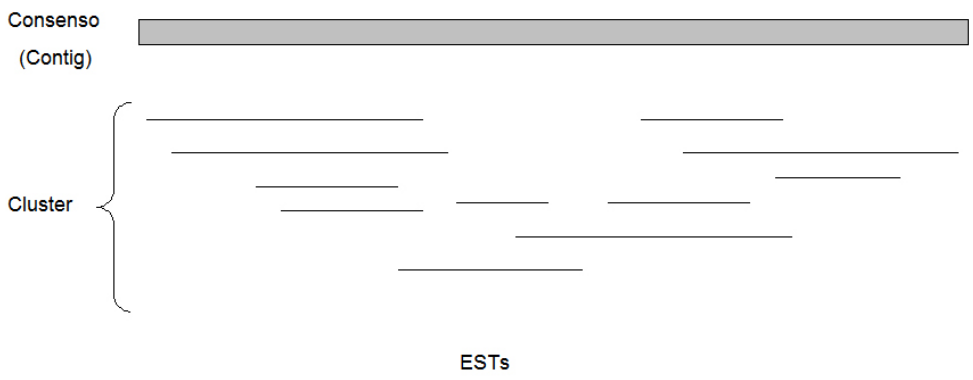


Figura 12. Ilustração esquemática de formação de um cluster.

5.1.1. Mineração de seqüências do gene ccoaomt

A pesquisa para o gene *ccoaomt* foi realizada a partir da seqüência de referência utilizada para as análises de polimorfismo em seqüências genômicas. O resultado revelou a existência de 6 contigs, todos do mesmo cluster e três singletons. As seqüências que compõem os contigs vieram de várias bibliotecas e de diferentes espécies conforme lista a tabela 2 (Apêndice I).

Tabela 2. Resultado da busca por seqüências no banco de dados, pelo método de clusters, para o gene *ccoaomt*.

	Cluster 6						**Singleton**		
	Contig 1	**Contig 2**	**Contig 3**	**Contig 4**	**Contig 5**	**Contig 6**	**EUGR**	**EUSP**	**EUGL**
Bibliotecas	EUGL-XY	EUGR-XY	EUPE-XY	EUGL-XY	EUGL-XY	EUUR-XY	EUGR-BC	EUSP-FX	EUGL-XY
	EUPE-XY	EUGL-XY	EUSP-FX	EUSP-FX	EUPE-XY				
	EUUR-XY	EUPE-XY	EUGR-PU		EUUR-XY				
	EUSP-FX	EUUR-XY							
		EUGR-ML							
		EUGR-PU							
		EUGR-SE							
		EUSP-FX							
Total de seqüências	11	24	10	8	21	3	1	1	1
Tamanho total	867 pb	1087 pb	1083 pb	1068 pb	1012 pb	713 pb	759 pb	739 pb	734 pb

Uma série de comparações foi realizada com os contigs. A primeira delas foi a determinação das regiões de exon e a forma de splicing realizada para cada um dos contigs. Foi utilizado um software de predição de genes por seqüências homólogas, onde um grupo de seqüências que sofreram splicing, como ESTs, cDNA ou mRNA, é comparado com seqüências que não sofreram splicing, ou seja, seqüências de DNA genômico. Esse software, denominado *EST2Genome* (MOTT, 1999; RICE *et al*., 2000; http:\\www.emboss.org), infere regiões de intron e exons com relativa precisão. A partir da seqüência de referência utilizada para a análise de diversidade nucleotídica (vide materiais e métodos), as regiões foram estabelecidas para os contigs em estudo (Figura 13).

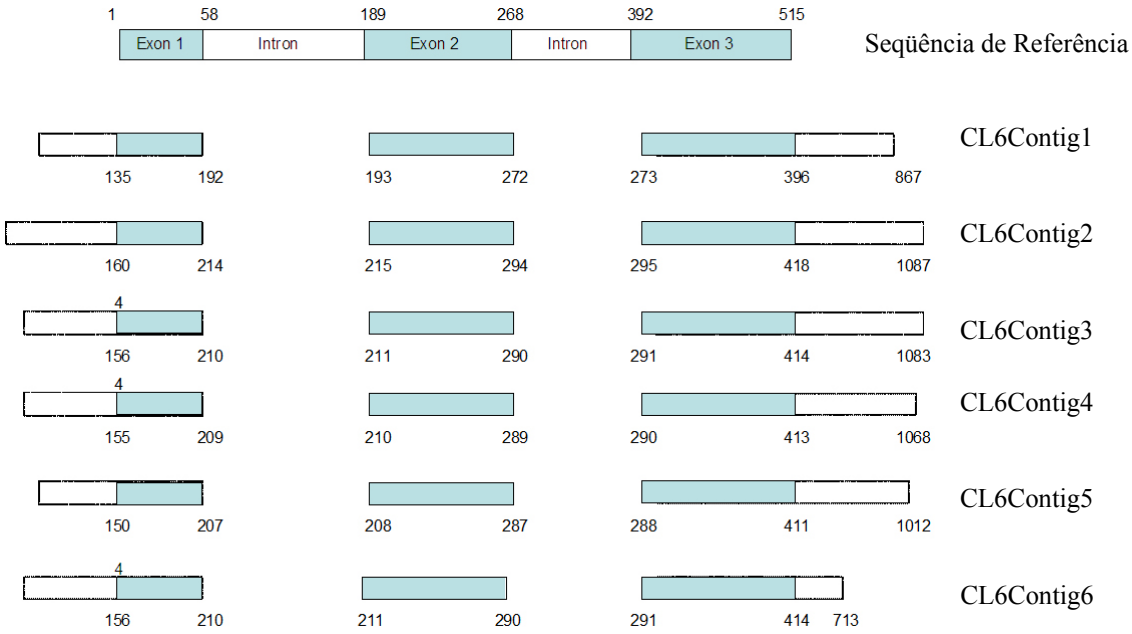


Figura 13. Ilustração esquemática das regiões de exon nos contigs obtidos do resultado da pesquisa no banco de dados do Projeto Genolyptus. Inferências realizadas a partir da seqüência de referência utilizada para os estudos de diversidade nucleotídica pelo programa *EST2Genome*. A região em azul nos contigs identifica identidade com as regiões de exon da seqüência de referência. Já as regiões brancas não resultaram em identidade com regiões de exons da seqüência de referência, o que infere-se que os exons 1 e 3 da seqüência de referência estão incompletos.

Foi observado que todos os contigs mantiveram o mesmo padrão de splicing, e que tanto o exon 1 e o exon 3 da seqüência de referência ainda está muito incompleto, visto que na inferência das regiões, os contigs tiveram identidade na posição 135 para o exon 1 e 418 na região do exon 3.

A segunda comparação realizada foi entre o conjunto de contigs. Todos os contigs foram alinhados conjuntamente pelo software *ClustalW* (HIGGINS *et al.*, 1994) e dos aproximadamente 1000 pb de extensão das seqüências, apenas aproximadamente 450 pares de bases indicaram identidade com score acima de 79 (Apêndice B). Mesmo pertencentes ao mesmo cluster, os vários contigs possuem diferenças entre si, seja por regiões distintas de formação do cluster, seja por diferenças pequenas nas seqüências devido a erros na construção do cDNA, erros de seqüenciamento ou mesmo polimorfismos originais.

Mesmo dentre as diferenças existentes entre os contigs, foi possível observar a

identidade entre dois contigs em especial, o contig 1 e o contig 5. Estes dois contigs foram alinhados pelo programa *Align* (SMITH & WATERMAN, 1981) e avaliadas algumas características.

Primeiramente, o alinhamento demonstrou uma identidade entre as seqüências de 94,7% e similaridade de mesma porcentagem entre 882 pb. A predição do códon ATG inicial, determinante do início da região codificadora do gene, foi determinado com o auxílio do software *ATGprediction* (SALAMOV *et al.*, 1998; http://www.hri.co.jp/atgpr/) e inferido o ATG de posição 100 para o contig 1 e 115 para o contig 5 (Apêndice C), ambos no quadro 1 de leitura. Essa predição corrobora a hipótese da existência de uma região 5'UTR nos contigs, e que no caso, seria de aproximadamente 100 pb para ambos os contigs. Para certificação da inferência correta do ATG inicial, foi realizada a comparação com a seqüência de aminoácidos dos contigs na posição do ATG inferido e de seqüências relacionadas ao gene em *Eucalyptus gunni* e em *Eucalyptus globulus* presentes no Genbank.

A primeira comparação foi realizada entre os contigs 1 e 5 e as seqüências disponíveis no GenBank para as isoformas do gene *ccoaomt*, CCoAOMT 1 (gi|3319278|) e CCoAOMT 2 (gi|5739373|) de *E. globulus* (87% de identidade e 93,9% de similaridade entre as duas seqüências) e a seqüência disponível para *E. gunni* (gi|1934859|). O resultado confirma a inferência correta do "Start Codon" para as seqüências dos contigs, pois confere com o códon inicial nas demais seqüências (Figura 14A). Do mesmo modo, é visível a diferença entre as seqüências dos contigs daquela referente a isoforma *ccoaomt* 2, sugerindo fortemente que os contigs do banco de dados fazem referência a isoforma *ccoaomt* 1 (Figura 14B), assim como a seqüência disponível no GenBank para *E. gunni*. A mesma comparação foi realizada para os demais contigs e o resultado foi o contrário. Ou seja, para os contigs 2, 3, 4, e 6, a similaridade com a isoforma *ccoaomt* 2 é maior do que para a *ccoaomt* 1. A região 3'UTR e 5'UTR apresentaram algumas variações, mas mesmo assim, de identidade maior para a

isoforma número 2. A região codificadora, foi próxima a 100% idêntica (Figura 15).

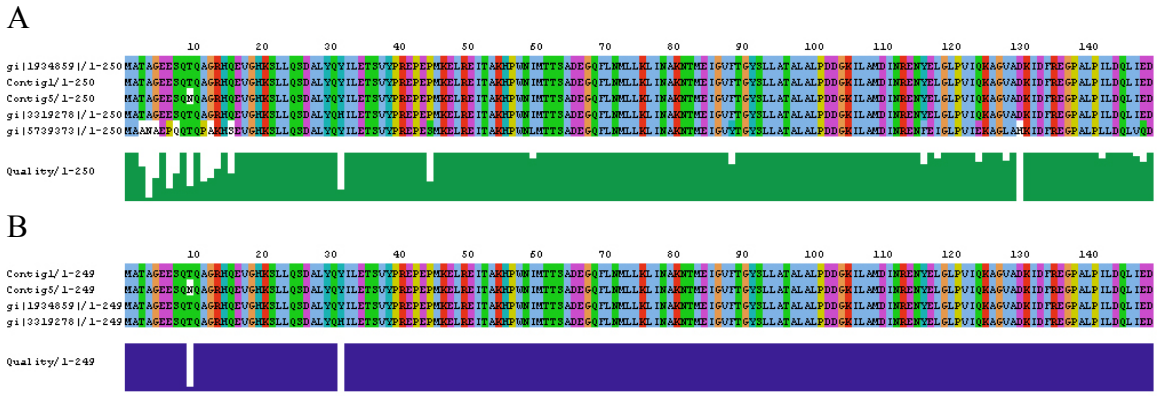


Figura 14. Alinhamento múltiplo realizado pelo programa *ClustalW* e visualizado pelo programa Jalview (CLAMP *et al*. 2004; http://www.jalview.org/). A. O alinhamento revela a correta inferência do códon ATG inicial para os contigs 1 e 5 do estudo. B. O alinhamento demonstra a similaridade de aproximadamente 100% entre os contigs do estudo e a seqüência da isoforma *ccoaomt* 1, inferindo que a seqüência de referência utilizada para a busca no banco de dados, é proveniente dessa isoforma, assim como os contigs e a seqüência obtida da espécie *E. gunni*.

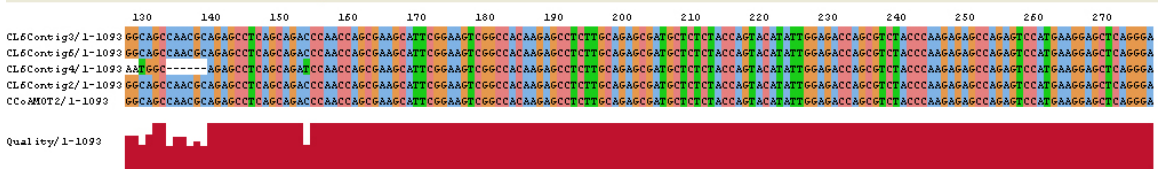


Figura 15. Alinhamento múltiplo dos contigs 2, 3, 4 e 6 do cluster 6 com seqüência nucleotídica da isoforma *ccoaomt* 2, demonstrando a alta similaridade entre as seqüências.

Além da alta similaridade, o alinhamento com as seqüências nucleotídicas revelou duas regiões muito interessantes para futuros estudos, tais como mapeamento do gene e assinaturas de seqüência espécie-específicas. As duas regiões de interesse encontram-se na suposta região 5' UTR das seqüências. Foi observada, na posição 23 do contig 5, uma inserção de 11 pb (Figura 16). A outra região compreende uma seqüência de microssatélite, composto por um dinucleotídeo e que apresenta polimorfismo de tamanho, já observado nos dois contigs analisados (Figura 16). Ambas as regiões fornecem potenciais marcadores úteis para estudos de mapeamento gênico e análises filogenéticas.

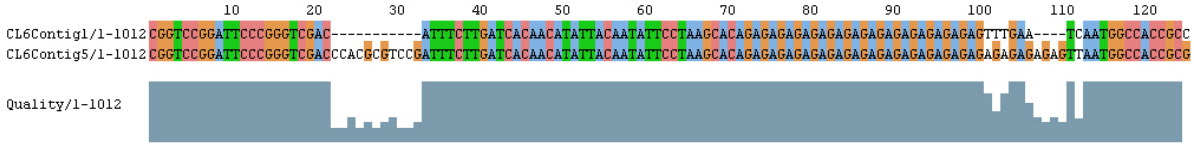


Figura 16. Alinhamento realizado pelo *ClustalW* e visualizado com o recurso Jalview entre os contigs 1 e 5 utilizados no estudo. No detalhe, a inserção de nucleotídeos no contig 5, próximo à posição 20 e o microssatélite dinucleotídico encontrado próximo à posição 70, polimórfico para ambas as seqüências.

Ainda na análise dos contigs 1 e 5, a região 3' UTR revelou ser bastante similar, não apresentando características visíveis que pudessem diferenciá-los em questão de seqüência (Apêndice D).

Para os demais contigs, o alinhamento múltiplo revelou uma identidade de quase 100% para toda a extensão, excetuando a região 5' UTR. Nesta região, foram observadas duas inserções de grupos de nucleotídeos. A primeira inserção, em torno da base 36, consiste no acréscimo de 5 bases, ACACA, visualizada no contig 4 (Figura 17, Apêndice E). A segunda inserção, localizada na posição 124, de três pares de base, supõe-se ser uma repetição, das bases TCA, que no contig 2 e o contig 4, estão presentes (Figura 17). Ademais, tanto as regiões 3' UTR, 5' UTR e exons, apresentam identidade entre os 4 contigs. São eles, o contig 2, contig 3, contig 4 e contig 6.

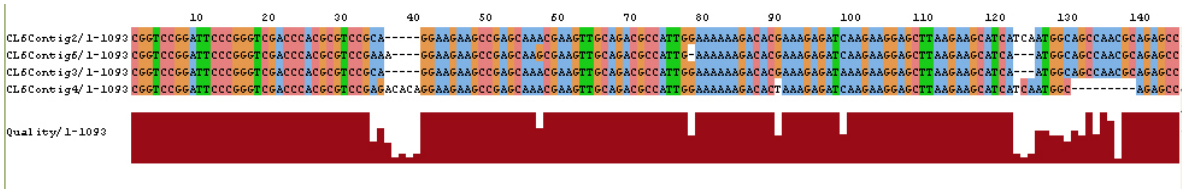


Figura 17. Alinhamento realizado pelo *ClustalW* e visualizado com o recurso Jalview entre os contigs 2, 3, 4 e 6 utilizados no estudo. No detalhe, a inserção de 5 nucleotídeos no contig 4, na posição 36 e a inserção de 3 nucleotídeos na posição 124, para os contigs 2 e 4.

Entre os clusters que foram resultados da pesquisa no banco de dados, três singletons também retornaram na pesquisa. Singletons são clusters compostos de somente uma

seqüência e por diferenças com os demais clusters ou por abranger uma outra região do gene, não é agrupado.

Um dos singletons resultado da pesquisa, de nome EUGR-BC tem a sua origem de uma biblioteca de BAC. Ou seja, esse clone não é uma EST, e sim uma região genômica, composta de exons e introns. Com 759 pares de base, essa seqüência é derivada do seqüenciamento das pontas dos BACs.

Com qualidade de seqüência aceitável, a cópia singular da seqüência gênica foi comparada com os representantes do banco para as isoformas do gene. O Contig 1 do cluster 6, por comparação e identidade comprovada anteriormente, foi determinado com o representante da isoforma *ccoaomt* 1 e o Contig 2 do cluster 6 como representante da isoforma *ccoaomt* 2. O alinhamento da seqüência nucleotídica do BAC com o Contig1, revelou similaridade e alta identidade com a isoforma *ccoaomt* 1. A comparação com a isoforma 2, diferenças no exon 1 e exon 3 determinaram, indubitavelmente, não se tratar desta segunda isoforma (Figura 18 e 19). Foram determinadas também as regiões de exons e introns para a seqüência genômica identificada (Figura 20).

```
Note Best alignment is between forward est and forward genome, and splice  sites
imply forward gene
Exon       170  99.4    64   235 EUGR_BAC        21   192 CL6Contig1
+Intron    -20   0.0   236   365 EUGR_BAC
Exon        80 100.0   366   445 EUGR_BAC       193   272 CL6Contig1
+Intron    -20   0.0   446   568 EUGR_BAC
Exon       143  99.3   569   713 EUGR_BAC       273   417 CL6Contig1

Span       353  99.5    64   713 EUGR_BAC        21   417 CL6Contig1

Segment    170  99.4    64   235 EUGR_BAC        21   192 CL6Contig1
Segment     80 100.0   366   445 EUGR_BAC       193   272 CL6Contig1
Segment    143  99.3   569   713 EUGR_BAC       273   417 CL6Contig1


EUGR_BAC vs CL6Contig1:

  EUGR_BAC     64 ACATTTCTTGATCACAACATATTACAATATTCCTAAGCAGAGAGAGAGAG    113
                  |||||||||||||||||||||||||||||||||||||| ||||||||||
CL6Contig1     21 ACATTTCTTGATCACAACATATTACAATATTCCTAAGCACAGAGAGAGAG     70

  EUGR_BAC    114 AGAGAGAGAGAGAGAGAGAGTTTGAATCAATGGCCACCGCCGGAGAGGAG    163
                  |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig1     71 AGAGAGAGAGAGAGAGAGTTTGAATCAATGGCCACCGCCGGAGAGGAG    120

  EUGR_BAC    164 AGCCAGACCCAAGCCGGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCT    213
                  |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig1    121 AGCCAGACCCAAGCCGGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCT    170

  EUGR_BAC    214 TCAGAGTGATGCTCTTTACCAAgtgag.....tgcagTATATTTTGGAGA    378
                  |||||||||||||||||||||>>>>> 130 >>>>>|||||||||||||
CL6Contig1    171 TCAGAGTGATGCTCTTTACCAA.............TATATTTTGGAGA    205

  EUGR_BAC    379 CCAGCGTGTACCCAAGAGAGCCTGAGCCCATGAAGGAGCTCAGGGAAATA    428
                  |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig1    206 CCAGCGTGTACCCAAGAGAGCCTGAGCCCATGAAGGAGCTCAGGGAAATA    255

  EUGR_BAC    429 ACAGCAAAACATCCATGgtgag.....aatagGAACATAATGACAACATC    586
                  |||||||||||||||>>>>> 123 >>>>>|||||||||||||||||||
CL6Contig1    256 ACAGCAAAACATCCATG.............GAACATAATGACAACATC    290

  EUGR_BAC    587 AGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCA    636
                  |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig1    291 AGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCA    340

  EUGR_BAC    637 AGAACACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTCGCCACC    686
                  |||||||||||||||||||||||||||||||||||||||||| ||||||
CL6Contig1    341 AGAACACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACC    390

  EUGR_BAC    687 GCTCTTGCTCTTCCTGATGACGGAAAG    713
                  ||||||||||||||||||||||||||
CL6Contig1    391 GCTCTTGCTCTTCCTGATGACGGAAAG    417


Alignment Score: 353
```

Figura 18. Alinhamento e determinação das regiões de exon e intron da seqüência gênica de *ccoaomt* obtida para o singleton EUGR_BC, resultado do seqüenciamento das pontas de BAC. A comparação realizada com o contig 1, representante da isoforma 1 do gene *ccoaomt* no presente estudo, revela a sua identidade e similaridade com a referida isoforma. Inferências realizadas pelo programa *EST2Genome.*

```
Note Best alignment is between forward est and forward genome, and splice  sites
imply forward gene
Exon        45  68.8    98   235 EUGR_BAC        79   214 CL6Contig2
+Intron    -20   0.0   236   365 EUGR_BAC
Exon        68  92.5   366   445 EUGR_BAC       215   294 CL6Contig2
+Intron    -20   0.0   446   568 EUGR_BAC
Exon       118  89.9   569   716 EUGR_BAC       295   442 CL6Contig2

Span       191  82.4    98   716 EUGR_BAC        79   442 CL6Contig2

Segment      3 100.0    98   100 EUGR_BAC        79    81 CL6Contig2
Segment      2 100.0   101   102 EUGR_BAC        83    84 CL6Contig2
Segment      6  71.4   103   116 EUGR_BAC        86    99 CL6Contig2
Segment      2 100.0   118   119 EUGR_BAC       100   101 CL6Contig2
Segment      2  60.0   121   130 EUGR_BAC       102   111 CL6Contig2
Segment      1  60.0   132   136 EUGR_BAC       112   116 CL6Contig2
Segment     11  73.9   138   160 EUGR_BAC       117   139 CL6Contig2
Segment     -1  40.0   161   165 EUGR_BAC       141   145 CL6Contig2
Segment     35  75.4   167   235 EUGR_BAC       146   214 CL6Contig2
Segment     68  92.5   366   445 EUGR_BAC       215   294 CL6Contig2
Segment    118  89.9   569   716 EUGR_BAC       295   442 CL6Contig2


EUGR_BAC vs CL6Contig2:

  EUGR_BAC     98 AAG-CA-GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGTTTGAATCAATG    145
                  ||| || || |||||   ||| || |||    ||| ||  |  |||||||
CL6Contig2     79 AAGACACGAAAGAGATCAAGA-AG-GAGCTTAAGA-AGCAT-CATCAATG    124

  EUGR_BAC    146 GCCACCGCCGGAGAG-GAGAGCCAGACCCAAGCCGGGAGGCACCAGGAGG    194
                  ||  ||  || ||||    || |||||||| | | || |||   ||| |
CL6Contig2    125 GCAGCCAACGCAGAGCCTCAG-CAGACCCAACCAGCGAAGCATTCGGAAG    173

  EUGR_BAC    195 TTGGCCACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAAgtgag....    235
                  | |||||||||   ||| | ||||| |||||||| ||||| >>>>> 130
CL6Contig2    174 TCGGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAG.........    214

  EUGR_BAC    235 .tgcagTATATTTTGGAGACCAGCGTGTACCCAAGAGAGCCTGAGCCCAT    409
                   >>>>>|| || |||||||||||||| |||||||||||||| ||| ||||
CL6Contig2    214 ......TACATATTGGAGACCAGCGTCTACCCAAGAGAGCCAGAGTCCAT    258

  EUGR_BAC    410 GAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGgtgag.....aata    445
                  |||||||||||||||||||||||| ||||||||||>>>>> 123 >>>>
CL6Contig2    259 GAAGGAGCTCAGGGAAATAACAGCCAAACATCCATG..............    294

  EUGR_BAC    445 gGAACATAATGACAACATCAGCAGACGAAGGGCAGTTCTTGAACATGCTT    617
                  >|||| | ||||| ||||| || || |||||||||||| ||||||||||
CL6Contig2    294 .GAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGCTC    343

  EUGR_BAC    618 CTCAAGCTCATCAACGCCAAGAACACCATGGAGATTGGTGTCTTCACTGG    667
                  |||||||||||||||||||||||||||||||||| || |||| ||| ||
CL6Contig2    344 CTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACCGG    393

  EUGR_BAC    668 CTACTCTCTCCTCGCCACCGCTCTTGCTCTTCCTGATGACGGAAAGGTC    716
                  |||||||||||||||||||| |||||||||| |||||||||||| ||
CL6Contig2    394 CTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGATC    442
```

Figura 19. Alinhamento e determinação das regiões de exon e intron da seqüência genômica de *ccoaomt* obtida no singleton EUGR_BC, resultado do seqüenciamento de ponta de BAC. A comparação realizada com o contig 2, representante da isoforma 2 do gene *ccoaomt* no presente estudo, revela a sua diferença com a referida isoforma, descartando a possibilidade da seqüência gênica ser um representante da isoforma 2. Inferências realizadas pelo programa *EST2Genome*.

A comparação entre a seqüência genômica do BAC e a seqüência de referência utilizada no estudo revelou identidade e similaridade de 99,2% em 515 pares de base, além de a primeira ser visivelmente maior do que a seqüência de referência. Esse acréscimo no tamanho é útil na medida que as seqüências gênicas disponíveis no banco de dados de acesso público para o gene *ccoaomt* são de outros organismos que não *Eucalyptus* e a descoberta de uma seqüência maior que a obtida pela amplificação com iniciadores, como a obtida no presente estudo, complementa esta análise.

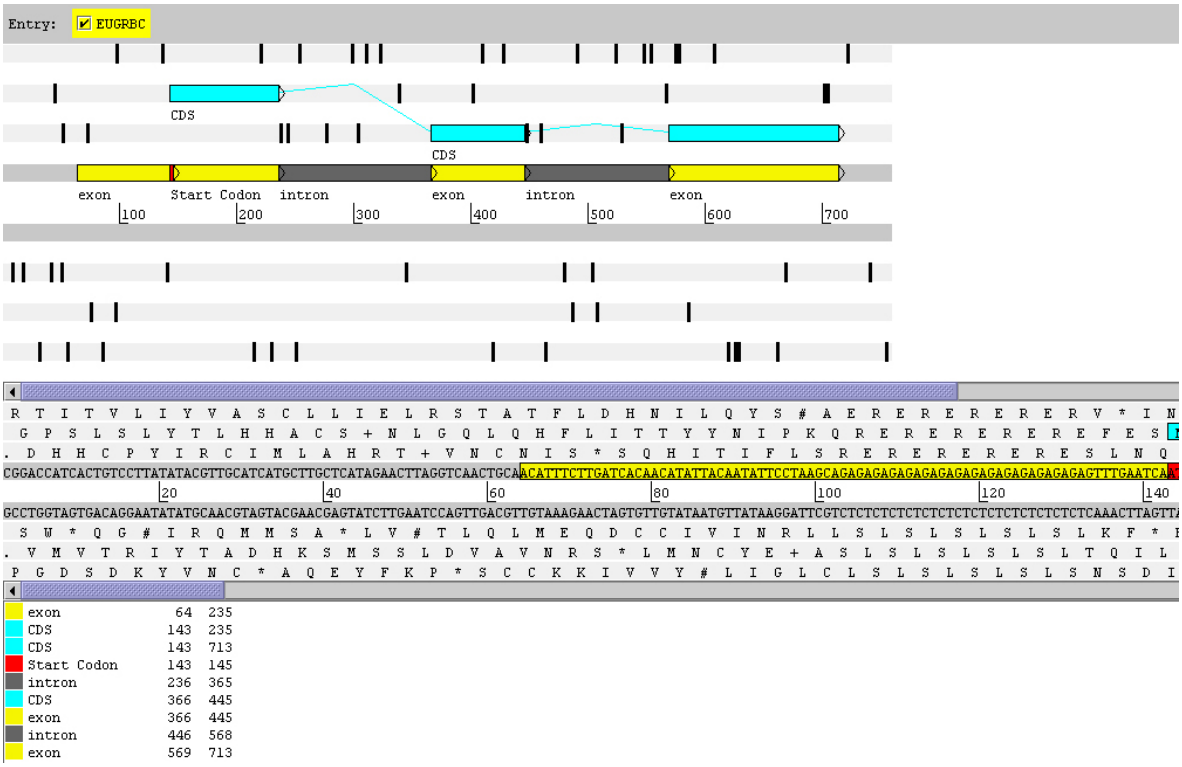


Figura 20. Ilustração esquemática da estrutura gênica da seqüência nucleotídica do singleton EUGR-BC. Regiões de intron e exons inferidas a partir do resultado obtido com a comparação, pelo programa *EST2Genome*, da seqüência do BAC com o contig 1. Figura gerada pelo programa *Artemis* (RUTHERFORD et al., 2000; http://www.sanger.ac.uk/Software/Artemis/).

Entretanto, mesmo com 244 pares de bases além da seqüência de referência, a seqüência do BAC não está completa para o gene *ccoaomt*. Este fato é confirmado quando analisados os alinhamentos com o Contig 1 (Apêndice F). O Contig 1, por ser uma EST, constitui a região codificadora do gene, e possui os exons em sua totalidade. Partindo do

pressuposto de que o Contig 1 apresenta todos os exons necessários para a tradução da enzima CCoAOMT 1 e que a região gênica do BAC é codificante para a mesma isoforma que o Contig 1, restam-lhe completar 460 pares de bases do exon 3, presentes no Contig 1 mas que não tiveram identidade com a seqüência do BAC. Os últimos 46 pares de base do singleton tiveram identidades muito baixas e portanto, nem estes foram determinados como exon 3 (Apêndice F). Provavelmente, a região codificadora da seqüência gênica oriunda do BAC não está completa. Deve existir ainda mais um exon, visto que a seqüência do Contig 1 já determina um exon de 144 nucleotídeos e seria muito pouco provável que mais 450 nucleotídeos fossem adicionados à esse exon, até mesmo pelo padrão de tamanho que é observado, de aproximadamente 200 nucleotídeos, ou ainda a existência de uma região 3'UTR.

Os dois últimos singletons, EUSP-FX e o EUGL-XY foram analisados em conjunto. O alinhamento múltiplo foi satisfatório considerando algumas ressalvas. Apesar de possuírem tamanhos aproximados, a região de alinhamento é deslocada (Apêndice G). O início da seqüência de EUGL-XY não tem identidade com a de EUSP-FX, assim como o final desta não possui identidade com a primeira. Ou seja, a região 3'UTR está representada em EUSP-FX e a 5'UTR em EUGL-XY. A identidade ocorre na região codificadora, e abrange o início do primeiro exon e o final do terceiro (Apêndice H).

No singleton EUSP-FX, é visível a região de microssatélite com o dinucleotídeo AG, assim como nos contigs 1 e 5 (Apêndice G). Diante da possibilidade da existência de maiores identidades que não o microssatélite, foi então realizado um alinhamento múltiplo entre as três seqüências.

A região inicial, de 22 pares de bases, revelou identidade e similaridade de 100%. Após essa região, foi identificada uma duplicação de 12 nucleotídeos no singleton EUSP-FX, presente uma única vez no contig 5 e ausente no contig 1 (Figura 21).

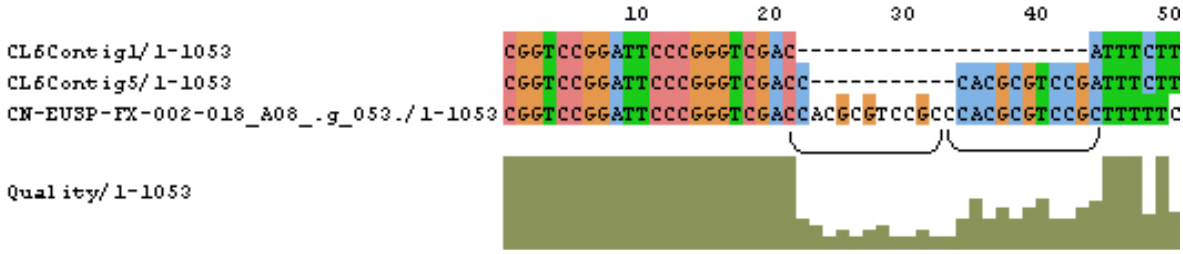


Figura 21. Alinhamento múltiplo de seqüências com similaridade para o gene *ccoaomt*, realizado pelo *ClustalW* e visualizado com o auxilio do recurso Jalview. Detalhe para a região de duplicação de uma seqüência de 12 nucleotídeos entre as seqüências selecionadas, contigs 1 e 5 do cluster 6 e a seqüência CN_EUSP_FX_002_018_A08.

Além da região de duplicação, a região do microssatélite com repetições dinucleotídicas AG também foi detectada, revelando polimorfismo de tamanho entre as três seqüências presentes no banco (Figura 22).

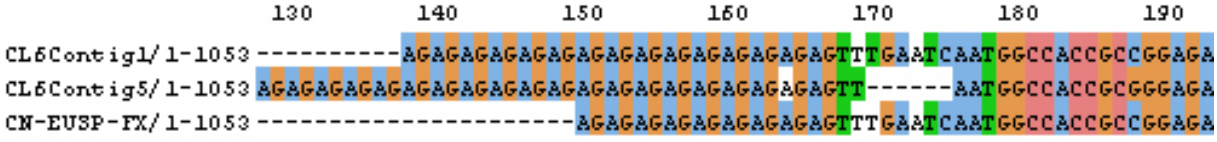


Figura 22. Alinhamento múltiplo de seqüências com similaridade para o gene *ccoaomt*, realizado pelo *ClustalW* e visualizado com o auxilio do recurso Jalview. Detalhe para a região de microssatélite dinucleotídica, AG. Entre as três seqüências, é possível observar o polimorfismo de tamanho, característica fundamental para que a região possa ser realmente considerada como um marcador microssatélite.

Para as demais regiões das seqüências, inclusive para região 3'UTR, foi observada identidade acima de 90% entre as três seqüências.

Uma nova busca ao banco de dados, pelo sistema Genoma, foi realizada para o gene *ccoaomt*. Utilizando seqüências obtidas no Genbank, a busca para duas isoformas do gene foi feita tendo como resultado novos consensos (Tabela 3, Apêndice I).

Tabela 3. Resultado da busca por seqüências no banco de dados, pelo método de clusters para as isoformas de *ccoaomt.*

CCoAOMT 1 gi|3319277|	**CCoAOMT 2** gi|5739372|
Cluster 6: Contig 1, Contig 2, Contig 3, Contig 4, Contig 5, Contig 6	Cluster 6: Contig 1, Contig 2, Contig 3, Contig 4, Contig 5, Contig 6
	Cluster 5089 Contig 1
Singleton: EUGR-BC, EUSP-FX, EUGL-XY_002, EUGR-ML_A04, EUGR-ML_F05	Singleton: EUGR-BC, EUSP-FX, EUGL-XY_001, EUGL-XY_002, EUGR-ML_A04, EUGR-ML_F05

Para a isoforma *ccoaomt* 1, em comparação com a busca realizada com a seqüência de referência, houve o acréscimo de dois singletons, ambos derivados da biblioteca da espécie *Eucalyptus grandis* de folhas maduras. As novas seqüências não acrescentaram informação, já que ambas as seqüências não tiveram identidade significativa com as demais. Provavelmente, a inclusão destas seqüências nos resultados da busca para a isoforma *ccoaomt* 1 deve-se a uma pequena região de identidade e que por esse motivo foram adicionadas ao grupo. Vale ressaltar que a busca no banco de dados está propensa a erros, e que as identidades com a seqüência de busca não necessariamente são de 100%, permitindo identidades limitadas. Diante desta possibilidade, faz-se necessário a análise de cada uma das seqüências do resultado para confirmação. No caso dos singletons EUGR-ML_A04 e EUGR-ML_F05 a identidade, bem como a similaridade, revelaram ser muito baixas e por isso não foram aceitos nas análises.

A isoforma *ccoaomt* 2 foi também confrontada com o banco de dados e o resultado foi a adição de 4 consensos além daqueles já descritos para a seqüência de referência dos estudos de diversidade nucleotídica. Dois dos consensos são os mesmos a pouco descritos para a isoforma *ccoaomt* 1. Os outros dois compreendem um singleton e um contig de um cluster.

O singleton EUGL-XY_001, resultado da busca com a isoforma *ccoaomt* 2, foi alinhado com o Contig 2, representante da isoforma no banco de dados do Genolyptus. A comparação revelou uma pequena região de identidade e, devido a possíveis problemas no seqüenciamento, a seqüência do singleton possuía varias repetições o que impossibilitou a maior identidade com o Contig 2. Da mesma forma, o retorno dessa seqüência na busca no banco de dados foi devido a uma restrita região que teve similaridade, ainda que não 100%, com a seqüência da isoforma utilizada na busca.

O Contig 1 do Cluster 5809 foi comparado por alinhamento global com o contig 2 do cluster 6. Dos 337 pares de base de extensão da seqüência, os primeiros 100 nucleotídeos tiveram baixa identidade com a isoforma, mas identidade total com restante da seqüência, confirmando que o contig 1 do cluster 5809 é um exemplar da isoforma 2 do gene *ccoaomt* (Apêndice J).

5.1.2. Mineração de seqüências do gene 4cl

É bem descrito na literatura que o gene *4cl* pertence a uma família gênica (EHLTING *et al.*, 1999). Por isso, a busca no banco de dados do Projeto Genolyptus baseou-se em seqüências de quatro isoformas do gene, descritas para *Arabidopsis thaliana.* A escolha dessa espécie teve como consideração dois pontos importantes. Primeiro pelo fato de *Arabidopsis* ser o organismo modelo para estudos em plantas e segundo por ser bem estudado em *Arabidopsis* o gene *4cl* e suas variações. As seqüências utilizadas para a pesquisa foram obtidas do banco de dados de domínio público SwissProt (GASTEIGER *et al.*, 2001), de 4 formas diferentes de apresentação da enzima, 4CL1 (sp|Q42524|), 4CL2 (sp|Q9S725|), 4CL3 (sp|Q9S777|) e 4CL4 (sp|Q9LU36|). As 4 seqüências são de origem protéica, ou seja, seqüências de aminoácidos.

O motivo crucial para a escolha das seqüências de aminoácidos e não nucleotídicas

deve-se ao fato de que, como as seqüências da busca de *Arabidopsis thaliana* e as seqüências do banco são de espécies distintas, a comparação entre as seqüências de nucleotídeos torna-se impossibilitada, mesmo codificando para a mesma enzima, por diferenças existentes nas seqüências as quais impossibilitam o alinhamento e a comparação entre elas. Dessa forma, a busca no banco de dados foi realizada pela adição da seqüência protéica em campo apropriado no sistema Genoma e o método de busca escolhido foi o *tblastn*, que se baseia na busca de identidades entre uma seqüência de aminoácidos e as seqüências de nucleotídeos traduzidas do banco de dados.

O resultado da busca para as quatro isoformas revelou a existência de vários clusters e singletons que possuíram identidade significativa para as isoformas. Os contigs e singletons são ranqueados por scores, indicados pelo Bit Score, que significa uma pontuação de identidade entre as duas seqüências, ou seja, a pontuação entre a seqüência utilizada para a busca e a resultante da busca. O Score Bit está relacionado também ao número de seqüências, onde quanto maior o número de seqüências maior a pontuação (Tabela 4).

Tabela 4. Resultado da busca por seqüências no banco de dados, pelo método de clusters para o gene *4cl* . Foram listados apenas os 5 primeiros resultados da busca.

Isoformas	Referências	Resultado do banco de dados	Bit Score	No. Seqüências
4CL1	sp|Q42524| 17*	CL273Contig2	758	11
		CL4405Contig1	303	3
		CL273Contig1	241	2
		CL2848Contig1	239	2
		EUSP-FX-002	214	1
4CL2	sp|Q9S725| 21*	CL273Contig2	766	11
		CL4405Contig1	322	3
		CL273Contig1	266	2
		CL2848Contig1	257	2
		EUSP-FX-002	223	1
4CL3	sp|Q9S777| 19*	CL273Contig2	659	11
		CL4405Contig1	356	3
		CL273Contig1	247	2
		CL2848Contig1	236	2
		EUSP-FX-002	234	1
4CL4	sp|Q9LU36| 17*	CL273Contig2	624	11
		CL4405Contig1	268	3
		CL273Contig1	222	2
		CL2848Contig1	199	2
		EUSP-FX-002	190	1

*Número total de contigs e singletons resultados da pesquisa para a isoforma.

Foram selecionados os 4 consensos do banco que resultaram em maior identidade com as isoformas. Por coincidência, todos os 4 consensos foram idênticos para as 4 isoformas e comparações entre eles foram realizadas (Apêndice K).

Inicialmente, foi inferido o ATG inicial bem como toda a seqüência de aminoácidos para cada um dos 4 contigs mais idênticos com as isoformas, são eles: CL273Contig2, CL4405Contig1, CL273Contig1 e CL2848Contig1 (Apêndice K). As seqüências de proteína das isoformas foram alinhadas com os consensos traduzidos utilizando o programa *Align* (SMITH & WATERMAN, 1981) e verificadas as identidades em busca de relacionar a qual isoforma o consenso era originário (Tabela 5).

Tabela 5. Identificação dos consensos quanto às isoformas por comparação de seqüências.

Consenso	Isoforma	Identidade (%)	Similaridade (%)
CL273Contig2	4CL1	74,1	85,5
CL273Contig1	4CL2	79,6	91,2
CL4405Contig1	4CL3	84,7	92,6
CL273Contig1	4CL4	78,7	89,4

O Contig 1 do Cluster 273 apresentou similaridade significativa tanto para a isoforma 4CL2 quanto 4CL4, mas foi o de maior identidade para a isoforma 4CL4. Comparando ambas isoformas, foi possível observar que elas apresentam uma similaridade significativa de 81,2%, propiciando a dupla identificação para as isoformas. No entanto, é importante deixar bem claro que o contig 1 do cluster 273 possui maior similaridade com a isoforma 2 do que com a 4 (Apêndice L). Essas análises são sugestões iniciais para a representação das isoformas no banco de dados do projeto Genolyptus.

Diante da representação das isoformas no banco de dados, foi realizada uma comparação com a seqüência genômica de *4cl* obtida na montagem do *shotgun* do clone BAC contendo o gene *4cl*. A descrição da montagem e obtenção da seqüência gênica será descrita

mais adiante na seção 5.5.

A seqüência genômica utilizada para a busca, com 5203 pb e regiões de exon e introns determinados, foi submetida ao banco de dados do Projeto Genolyptus. Com o modo de busca em clusters, a pesquisa foi realizada pelo método *blastn*, comparando a seqüência de nucleotídeos com o banco de dados. Com o retorno da pesquisa, foram obtidos 2 contigs, CL273Contig1 e CL273Contig2, ambos também resultantes da pesquisa com as isoformas, e apresentaram bit score de 426 e 1853 respectivamente (Tabela 6).

Para validar o resultado emitido pelo sistema, os Contigs 1 e 2 do cluster 273, com 740 e 2022 nucleotídeos respectivamente, foram alinhados com a seqüência genômica da *4cl* e conferidas as regiões de introns e exons. Foi possível certificar que as seqüências que retornaram pelo sistema conferem com a seqüência de *4cl* e possuem elevada identidade.

Tabela 6. Resultado da busca por seqüências no banco de dados pelo método de clusters para o gene *4cl* .

Banco de dados	Biblioteca	No. Seqüências	Bit Score
CL 273 Contig 1	EUGR-TS	2	426
CL 273 Contig 2	EUGL-XY EUGR-ML EUGR-PU EUGR-SE EUPE-XY EUUR-XY	11	1853

Apesar de ambas as seqüências dos contigs apresentarem alta identidade com as regiões de exon da seqüência genômica, a seqüência do Contig 2 do cluster 273 revelou a maior identidade na comparação além de compreender na totalidade os exons previstos para a seqüência, incluindo parte da região 3'UTR (Figura 23).

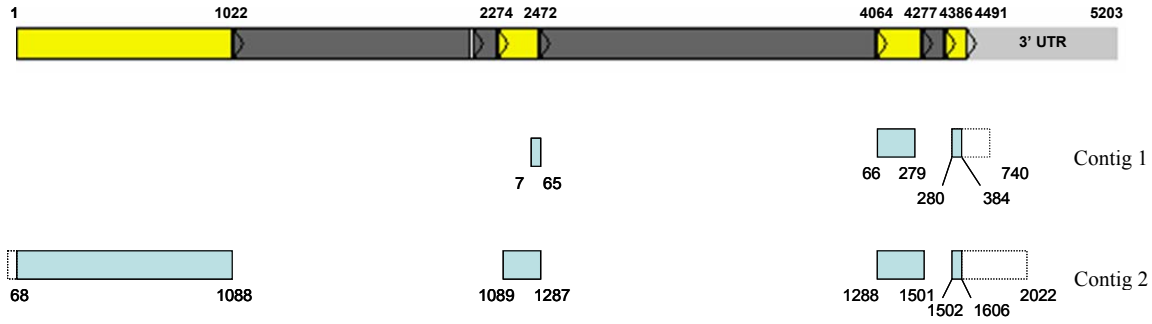


Figura 23. Ilustração esquemática da distribuição dos exons dos contigs, resultantes da busca no banco de dados do Projeto Genolyptus. A seqüência referência é a seqüência montada a partir das seqüências obtidas com o seqüenciamento do BAC. A região em tracejado é a seqüência de continuidade do exon da EST que não revelou identidade com a seqüência genômica. A primeira montagem refere-se ao contig 1 e a segunda montagem ao contig 2 do cluster 273. As regiões em amarelo, cinza escuro, cinza claro e azul são, respectivamente, regiões de exon, intron, 3'UTR e EST.

Esta observação sugere que a seqüência gênica obtida a partir da biblioteca *shotgun* de BAC para a espécie *E. grandis* é um exemplar da isoforma 1 do gene *4cl*, por inferência anteriormente realizada para o contig 2 do cluster 273. Para confirmação, foram inferidas as regiões de exons e introns por comparação do consenso das ESTs com a seqüência gênica pelo programa *EST2Genome*. O alinhamento realizado pelo programa revela a identidade da seqüência advinda do seqüenciamento do BAC com a isoforma 1 do gene *4cl* (Figura 24, Apêndice M).

```
    4CL_EUGR    451 CGCGGAGGAGAACGGGGTGAAGGTCGTGTGCATCGATACCGCGCCGGAGG    500
                    |||||||||||||||||| |||||||||||||||||||||||||||| |
CL273Contig2    518 CGCGGAGGAGAACGGGGTCAAGGTCGTGTGCATCGATACCGCGCCGGA-G    566

    4CL_EUGR    501 GGCTGCCTGCACTTCTCGGAATTGATGCAGGCGGACGAGAACGCCGCCCC    550
                    |||||||||||||||||||||||||||||||||||| ||||||||||||||
CL273Contig2    567 GGCTGCCTGCACTTCTCGGAATTGATGCAGGCGGATGAGAACGCCGCCCC    616

    4CL_EUGR    551 CGCGGCGGACGTCAAGCCGGACGACGTCTTGGCGCTCCCCTATTCGTCGG    600
                    |||||||||||||||||||||||||||||||||| ||||||||||||||||
CL273Contig2    617 CGCGGCGGACGTCAAGCCGGACGACGTCTTGGCACTCCCCTATTCGTCGG    666

    4CL_EUGR    601 GCACGACGGGGCTTCCCAAGGGAGTGATGCTTACGCACAGGGGTCAAGTG    650
                    |||||||||||||| |||||||| |||||||| ||||||||||||||||||
CL273Contig2    667 GCACGACGGGGCTCCCCAAGGGGGTGATGCTCACGCACAGGGGTCAAGTG    716

    4CL_EUGR    651 ACCAGCGTGGCGCAGCAGGTCGACGGAGACAACCCCAACTTGTACTTCCA    700
                    |  ||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    717 AGTAGCGTGGCGCAGCAGGTCGACGGAGACAACCCCAACTTGTACTTCCA    766

    4CL_EUGR    701 CAAGGAGGACGTGATCCTGTGCACGCTCCCGTTGTTCCACATATACTCCC    750
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    767 CAAGGAGGACGTGATCCTGTGCACGCTCCCGTTGTTCCACATATACTCCC    816

    4CL_EUGR    751 TCAACTCGGTGATGTTCTGCGCGCTCCGTGTAGGCGCCGCCATCCTGATC    800
                    |||||||||||||||||||||||||||||| ||||| ||||||||||||
CL273Contig2    817 TCAACTCGGTGATGTTCTGCGCGCTCCGTGTCGGCGCTGCCATCCTGATC    866

    4CL_EUGR    801 ATGCAGAAGTTCGAGATCGTGGCGCTGATGGAGCTCGTGCAGCGGTACCG    850
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    867 ATGCAGAAGTTCGAGATCGTGGCGCTGATGGAGCTCGTGCAGCGGTACCG    916

    4CL_EUGR    851 GGTGACGATCCTGCCCATTGTCCCGCCGATCGTGCTGGAGATCGCCAAGA    900
                    |||||||||||||||||| |||||||| |||||||||| ||||||||||||
CL273Contig2    917 GGTGACGATCCTGCCCATCGTCCCGCCAATCGTGCTGGCGATCGCCAAGA    966

    4CL_EUGR    901 GCGCCGAGGTGGACCGGTACGACCTGTCGTCGATCCGGACCATCATGTCG    950
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    967 GCGCCGAGGTGGACCGGTACGACCTGTCGTCGATCCGGACCATCATGTCG   1016

    4CL_EUGR    951 GGTGCGGCCCCGATGGGGAAGGAGCTCGAGGACACCGTGCGAGCCAAGCT   1000
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2   1017 GGTGCGGCCCCGATGGGGAAGGAGCTCGAGGACACCGTGCGAGCCAAGCT   1066

    4CL_EUGR   1001 GCCCAATGCCAAGCTCGGACAGgtgaa......tgcagGGCTATGGGATG   2285
                    ||| |||||||||||||||>>>>> 1251 >>>>>|||||||||||||
CL273Contig2   1067 GCCGAATGCCAAGCTCGGACAG.................GGCTATGGGATG   1100

    4CL_EUGR   2286 ACGGAGGCGGGCCCGGTGCTGGCAATGTGCCCGGCATTTGCAAAGGAGCC   2335
                    |||||||||||||||||||||||||||||| ||||||||||||||||||
CL273Contig2   1101 ACGGAGGCGGGCCCGGTGCTGGCAATGTGCCTGGCATTTGCAAAGGAGCC   1150

    4CL_EUGR   2336 GTTCGAGATCAAGTCAGGCGCATGCGGGACCGTCGTGAGGAACGCGGAGA   2385
                    |||||||||||||||||||| ||||||||||||||||||||||||||||
CL273Contig2   1151 GTTCGAGATCAAGTCAGGCGCGTGCGGGACCGTCGTGAGGAACGCGGAGA   1200
```

Figura 24. Parte do alinhamento realizado pelo programa *EST2Genome* para a seqüência genômica originária do BAC e o consenso das ESTs disponíveis no banco de dados do Projeto Genolyptus. As identidades refletem a região de exon e a parte pontilhada, indicando 1251 pb é a região de intron, presente na seqüência gênica e ausente no consenso de ESTs.

A partir da seqüência do contig, que se refere à seqüência do consenso das ESTs montadas para a isoforma 1 do gene *4cl*, surgiu o questionamento se havia obtido a seqüência completa da região codificadora. A predição do ATG "Start Codon" foi a estratégia escolhida para resolver tal questionamento. Com o auxilio do programa *ATGprediction* (SALAMOV *et al.*, 1998; http://www.hri.co.jp/atgpr/) foi gerada uma lista com 5 opções de posição de ATG's e seqüências de aminoácidos resultantes dessa predição. A posição de ATG 336 foi o primeiro a ser encontrado dentre toda a seqüência nucleotídica do contig 2 do cluster 273, gerando uma ORF com 462 aminoácidos. Entretanto, essa predição não pode ser a correta, visto que o contig analisado trata-se de um consenso de ESTs onde admite-se como codificadora toda a sua extensão o quê de fato não ocorrerá admitindo o ATG predito como o inicial (Apêndice N).

Contudo, a fim de obter maior certeza na inferência do "Start Codon", a seqüência de aminoácidos gerada com o ATG inicial na posição 336 foi comparada com o banco de dados de domínio público através do programa *Blastp* (ALTSCHUL *et al*., 1990; GISH & STATES, 1993; http://www.ncbi.nlm.nih.gov/blast/). O resultado revelou alinhamentos com seqüências de varias espécies vegetais mas todas na posição 80 aproximadamente. Ou seja, mesmo selecionando a maior seqüência de 4CL de *Eucalyptus* gerada pela predição da posição ATG 336, ainda faltam aminoácidos anteriores para completar a região codificadora da enzima.

A seqüência de aminoácidos obtida com o ATG inicial em 336 também foi comparada com as seqüências de aminoácido das isoformas da enzima 4CL para a espécie *Arabidopsis*. As comparações revelaram que o ATG predito para o contig 2 do cluster 273 não condiz com o ATG inicial das seqüências, mas sim vários aminoácidos após. Diante desse resultado é possível propor duas hipóteses: a primeira é que realmente há um problema na predição do ATG correto, podendo existir um outro anterior a este que não está presente na seqüência

nucleotídica. A segunda hipótese é de que como as seqüências comparadas são de uma outra espécie, é de se esperar que haja uma variação na posição do ATG inicial entre as espécies.

Entretanto, analisando com mais cuidado a posição 336 do ATG inicial descrita para a seqüência e comparando com as inferências realizadas para a seqüência obtida pelo seqüenciamento do BAC, esse ATG estaria na região central do exon inferido, diferentemente do habitual. O codon inicial, na maioria das vezes, está contido na região inicial do exon. A partir dessa dedução, foram verificadas visualmente as possíveis posições do códon ATG e admitiu-se um encontrado na posição 90, o primeiro encontrado e não inferido pelo programa *ATGpr*. Selecionada a seqüência do consenso a partir da posição 90 e comparando-a de forma traduzida no programa *BlastX* (<http://www.ncbi.nlm.nih.gov/blast/>) no banco de dados de domínio público, NCBI, foi observado como resultado em *Populus* a identidade com o seu códon inicial e a alta similaridade entre as seqüências de 89% (Figura 25). Com base nesse resultado, pode-se sugerir que a seqüência codificadora para a enzima 4CL encontra-se completa com o contig 2 do cluster 273, visto que a espécie *Populus* é passível de comparação com *Eucalyptus*. Vale ressaltar ainda que a inferência sobre a posição do ATG inicial é aceitável visto que a seqüência analisada de *Populus* é uma seqüência validada, ou seja, confirmada por outros seqüenciamentos e/ou por experimentos.

Com esse resultado, pode-se inferir, por comparação com o contig 2 do cluster 273, que há uma região 5'UTR presente na seqüência gênica obtida com o seqüenciamento do BAC, já que o ATG "Start Codon" inicia-se na posição 90 (Figura 26).

```
>gi|2911799|gb|AAC39366.1|  4-coumarate:CoA ligase 1 [Populus balsamifera subsp. trichocarpa
          x Populus deltoides]
gi|7437854|pir||T07909  4-coumarate-CoA ligase (EC 6.2.1.12) 1 [validated] - western
          balsam poplar x cottonwood
        Length=557

 Score =  814 bits (2102),  Expect = 0.0
 Identities = 437/555 (78%), Positives = 494/555 (89%), Gaps = 5/555 (0%)
 Frame = +1

Query  1     MEAKPSEQPREFIFRSKLPDIYIPDNLSLHAYCFENISEFADRPCVINGATGRTYTYAEV  180
             MEAK ++Q +EFIFRSKLPDI+IP++L LH YCFEN+S F D PC+ING TG  +TYAEV
Sbjct  1     MEAK-NDQAQEFIFRSKLPDIHIPNHLPLHTYCFENLSRFKDNPCLINGPTGEIHTYAEV  59

Query  181   ELISRRVSAGLNGLGVGQGDVIMLLLQNCPEFVFAFLGASYRGAISTTANPFYTPGEXXX  360
             EL SR+V++GLN LG+ QGDVI+LLLQN PEFVFAFLGAS  GAISTTANPFYTP E+AK
Sbjct  60    ELTSRKVASGLNKLGIKQGDVILLLLQNSPEFVFAFLGASIIGAISTTANPFYTPAEVAK  119

Query  361   XXXXXXXXXVITQAAYADKVRPFAEEN-GVKVVCIDTAPEGCLHFSELMQADENXXXXXX  537
             QA+A++AK++ITQA YA+KV+ F +EN  VK+V +D+ PE  LHFSEL  +DE+  PA +
Sbjct  120   QATASKAKLIITQAVYAEKVQEFVKENVHVKIVTVDSPPENYLHFSELTNSDEDDIPAVE  179

Query  538   XXXXXXXXXPYSSGTTGLPKGVMLTHRGQVSSVAQQVDGDNPNLYFHKEDVILCTLPLFH  717
             + PDDV+ALPYSSGTTGLPKGVMLTH+G V+SVAQQVDG+NPNLYFH++DVILC LPLFH
Sbjct  180   INPDDVVALPYSSGTTGLPKGVMLTHKGLVTSVAQQVDGENPNLYFHEKDVILCVLPLFH  239

Query  718   IYSLNSVMFCALRVGAAILIMQKFEIVALMELVQRYRVTXXXXXXXXXXXXXXKSAEVDRY  897
             IYSLNSV+ C LRVG+AIL+MQKFEIV LMELVQ+Y+VTI P VPP+VLA+AK   VD+Y
Sbjct  240   IYSLNSVLLCGLRVGSAILLMQKFEIVTLMELVQKYKVTIAPFVPPVVLAVAKCPVVDKY  299

Query  898   DLSSIRTIMSGAAPMGKELEDTVRAKLPNAKLGQGYGMTEAGPVLAMCLAFAKEPFEIKS  1077
             DLSSIRT+MSGAAPMGKELEDTVRAKLPNAKLGQGYGMTEAGPVL+MCLAFAKEPFEIKS
Sbjct  300   DLSSIRTVMSGAAPMGKELEDTVRAKLPNAKLGQGYGMTEAGPVLSMCLAFAKEPFEIKS  359
```

Figura 25. Resultado do Blastx realizado na página do NCBI indicando a similaridade da seqüência do contig 2 do cluster 273 com a seqüência de um híbrido *Populus*, indicando a identidade do ATG inicial.

Além disso, a identidade com o contig é admitida a partir da posição 68, não observando qualquer identidade na região anterior à essa posição. Dessa forma, pode-se ainda inferir que essa região que antecede a identidade no contig, pode ser uma putativa região 5'UTR, que não está representada no BAC (Figura 26).

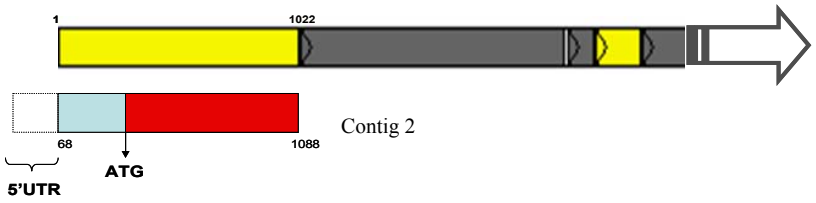


Figura 26. Ilustração esquemática da região inferida de 5'UTR no contig 2 do cluster 273 em comparação com a seqüência gênica do gene *4cl*.

5.2. Desenho de oligonucleotídeos iniciadores

Com o intuito de abranger uma região maior do gene *4cl*, iniciadores foram desenhados para a região codificadora do gene, além do par de primer descrito por Gion *et al*. (2000). A partir de seqüências EST disponíveis no banco de dados do Projeto Genolyptus, um consenso para a região codificadora de *4cl* foi gerado utilizando o sistema Genoma. Com 740 pb, o consenso de nome Cluster225_Contig1, serviu de molde para que os iniciadores fossem desenhados, por meio do programa *Primer3* (ROZEN & SKALETSKY, 2000), levando em consideração o tamanho aproximado de amplificação e as temperaturas de anelamento variando em 62° C ± 2° C (Tabela 7). Para inferir o tamanho esperado, foram realizadas comparações com o banco de dados de *Arabidopsis*, de disponibilidade pública, para determinar os locais de exons e introns (Tabela 8).

Para o gene *ccoaomt* foram utilizados iniciadores desenhados por Gion *et al.* (2000) os quais foram testados nas três espécies em estudo, gerando fragmentos de tamanhos diferentes daqueles esperados (Tabela 8).

Tabela 7. Oligonucleotídeos iniciadores utilizados para a amplificação de regiões dos genes *4cl* e *ccoaomt* .

Gene	Nome do Iniciador	Seqüência
4cl	**4CL1A-F**	5´ GGG TCA CCA GAT CAT GAA AG 3´
	4CL1B-R	5´AGC CAC CCT TCT TTG TCT ATG 3´
	4CL1C-R	5´GTC CTC GGT GAT TAC GGA AC 3´
	4CL1D-F	5´GAG GTT CCT GTT GCA TTC G 3´
	4CL1E-R	5´GCA TGT GAA ATC AAA TCA TCG 3´
	G-4CL - F *	5' GGG ACC GTC GTG AGG AAC GC 3'
	G-4CL - R *	5' CGC CTC TAA TTC GGC CGG AG 3'
ccoaomt	**G-CCoAOMT - F***	5' CGG GAG GCA CCA GGA GGT T 3'
	G-CCoAOMT - R*	5' GCA AGA GCG GTG GCA AGG A 3'

* descritos por Gion *et al.* , 2000

Tabela 8. Pares de iniciadores utilizados para a amplificação dos genes *ccoaomt* e *4cl* e os respectivos tamanhos esperados e observados nas três espécies de eucalipto: *E. grandis* , *E. globulus* e *E. urophylla* .

	Iniciadores	Tm (°C)	Tamanho esperado	Tamanho observado
A	4CL-A – F 4CL-B – R	63,5	78 pb	1600 pb
B	4CL-A – F 4CL-C – R	63,5	647 pb	1800 pb
C	4CL-D – F 4CL-E – R	63,5	439 pb	1100 pb
D	G-4CL – F G-4CL – R	64	281 pb *	2000 pb
E	G-CCoAOMT - F G-CCoAOMT - R	68	266 pb *	600 pb

* segundo dados de Gion *et al.* , 2000.

5.3. Resseqüenciamento e análise dos segmentos genômicos

5.3.1. Análise dos segmentos do gene 4cl

Dois grupos de iniciadores foram utilizados para a amplificação do fragmento da região genômica do gene *4cl*. O primeiro grupo de iniciadores foi descrito por Gion *et al.* (2000) e abrange uma região de aproximadamente 1800 pb. (Figura 27)

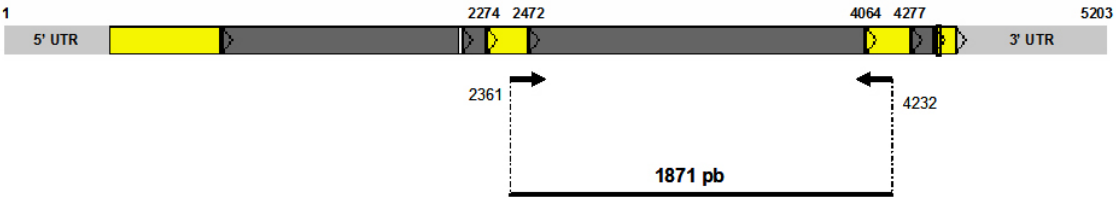


Figura 27. Representação esquemática da região de anelamento dos oligonucleotídeos iniciadores para o gene *4cl*, descritos por Gion *et al.* (2000). As regiões em cinza claro, escuro e amarelo correspondem, respectivamente, à região UTR, intron e exon.

Esse par de iniciador demandou um trabalhoso processo de otimização da reação de PCR para se chegar a uma única banda de amplificação (Figura 28). As condições descritas por Gion *et al.* (2000) incluíam, além da amplificação normal do fragmento, a eluição da banda amplificada e a re-amplificação da mesma. Nessa etapa, os autores não deixam claro se,

no momento da primeira amplificação, há a geração de apenas um fragmento. Entretanto, mesmo com a reação de PCR otimizada, após o seqüenciamento direto do produto de PCR de alguns indivíduos, foi detectada a presença de mais de uma seqüência, o que impossibilitou a utilização desse par de iniciadores para a obtenção de seqüências úteis para análise. Como era inviável a eluição da banda amplificada para todos os indivíduos das três espécies, e por abranger uma região de 2000 pb, sem iniciadores internos para facilitar o seqüenciamento do fragmento como um todo, optou-se por desenhar e utilizar novos iniciadores que abrangessem uma região de tamanho maior e que houvesse iniciadores internos que facilitassem o seqüenciamento.

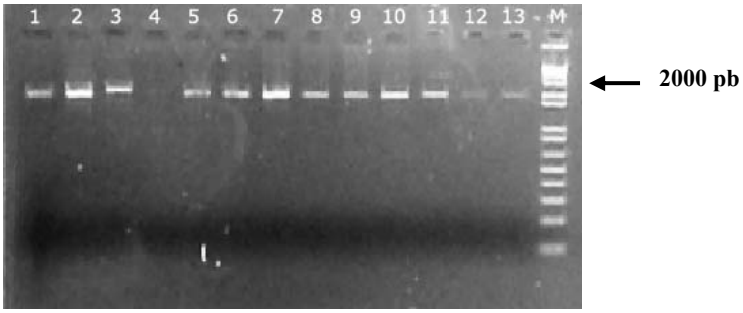


Figura 28. Perfil em gel de agarose dos fragmentos amplificados pelos iniciadores G-4CL. 1-13, indivíduos da espécie *E. grandis*; M, marcador molecular 1 Kb Plus DNA ladder (Invitrogen, Carlsbad, CA).

Foi desenhado um grupo de iniciadores identificados com o nome de 4CL A a E, composto por 5 oligonucleotídeos, agrupados de tal forma a gerar fragmentos genômicos de *4cl*. Os fragmentos gerados possuem regiões de sobreposição que facilitam a montagem de um fragmento maior, de aproximadamente 2600 pb, que engloba três exons, dois introns e parte da região 3' não traduzida (Figura 29).

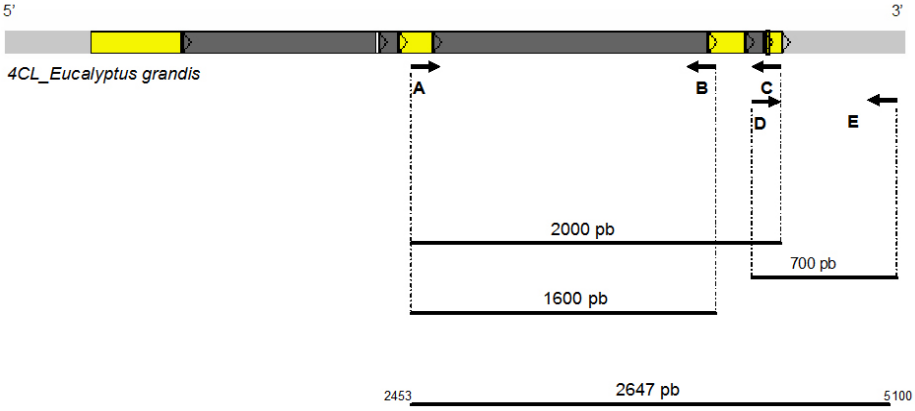


Figura 29. Representação esquemática da região de anelamento dos iniciadores 4CL (A-E) e o tamanho esperado dos fragmentos amplificados. As regiões em cinza claro, escuro e amarelo correspondem, respectivamente, à região UTR, intron e exon.

A reação de PCR foi de fácil otimização para as três espécies (Figura 30). No entanto, o seqüenciamento, assim como para os iniciadores descritos por Gion *et al.* (2000), também foi infrutífero, apresentando o seqüenciamento de duas seqüências para um único indivíduo. A solução neste caso seria a clonagem dos fragmentos amplificados para posterior seqüenciamento.

A amplificação dúbia dos fragmentos de *4cl* sugere fortemente que há mais de uma cópia do gene em estudo, corroborando a hipótese de que o gene *4cl* em *Eucalyptus* é pertencente a uma família gênica.

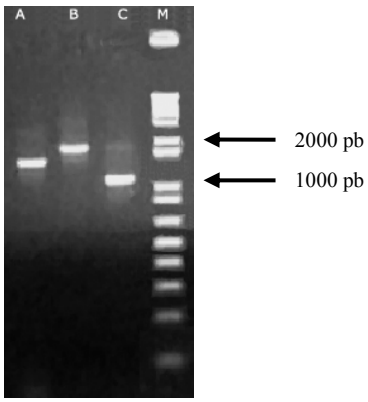


Figura 30. Perfil em gel de agarose dos fragmentos gerados pela combinação dos 5 iniciadores construídos para o gene 4CL. A, par 4CL-A e 4CL-B; B, par 4CL-A e 4CL-C; C, par 4CL-D e 4CL-E; M, marcador 1 Kb Plus DNA ladder (Invitrogen, Carlsbad, CA).

5.3.2. Análise dos segmentos do gene ccoaomt

Para o gene *ccoaomt* foi possível gerar um amplicon robusto com tamanho aproximado de 600 pb (Figura 31), abrangendo 1 exon em sua totalidade e 2 parciais, além de dois introns (Figura 32). Para identificar e caracterizar a diversidade nucleotídica intra e interespecífica neste gene, amostras de DNA de 76 indivíduos de *E. grandis*, 48 de *E. globulus* e 43 de *E. urophylla* foram submetidas à amplificação deste segmento do gene *ccoaomt*.

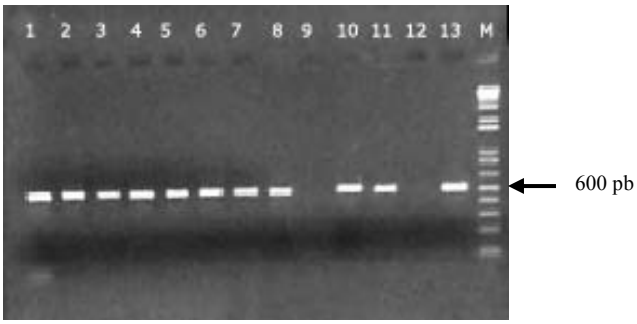


Figura 31. Perfil em gel de agarose dos fragmentos amplificados com os oligonucleotídeos descritos por Gion *et al.* (2000). 1-13, indivíduos da espécie *E. grandis*; M, marcador 1 Kb Plus DNA ladder (Invitrogen, Carlsbad, CA).

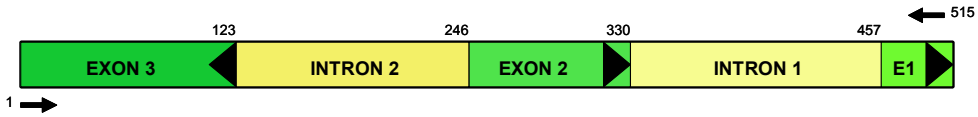


Figura 32. Esquema ilustrativo do segmento do gene *ccoaomt* analisado. As setas indicam a região de anelamento dos iniciadores G-CCoAOMT. Para a determinação das regiões de introns e exons foram realizadas comparações com o banco de dados de *Arabidopsis* do amplicon gerado.

Para alguns indivíduos não foi possível obter amplificação adequada por problemas na qualidade do DNA extraído.

Na análise das seqüências dos fragmentos genômicos amplificados nas espécies de

eucalipto: *E. grandis*, *E. urophylla* e *E. globulus*, somente foram consideradas 440 bases centrais do segmento. Esta porção foi escolhida por apresentar satisfatória resolução de bases na maioria das leituras indicadas pelo programa *SeqScape* e por inspeção visual dos cromatogramas. De acordo com a seqüência de referência, a região analisada se estende da base 31 ao 470.

Para a espécie *E. urophylla*, 147 seqüências foram aceitas para compor 43 consensos distintos de indivíduos desta espécie (Apêndice A). A comparação das seqüências revelou a existência de 7 posições polimórficas (SNPs), indicando uma freqüência média de um polimorfismo a cada 62,8 pares de base nesta espécie (Figura 33).

Para a espécie *E. grandis*, foram 169 leituras aceitas para 67 indivíduos (Apêndice A). A comparação das seqüências revelou a existência de 8 SNPs, resultando numa freqüência de um polimorfismo a cada 55 pares de base, em média (Fig. 33). Em *E. globulus*, 73 leituras para 37 indivíduos (Apêndice A), resultando em 4 SNPs, sendo que dois deles eram polimórficos em relação à seqüência de referência e demonstraram estar fixados para esta espécie. Estes sítios não foram considerados nas análises estatísticas. Dessa forma, a freqüência de polimorfismo para *E. globulus* é de um a cada 220 pares de base, a menor das freqüências entre as espécies em estudo (Fig. 33). A distribuição dos polimorfismos estendeu-se por todo o fragmento, tanto em regiões de exons quanto de introns (Figura 34).

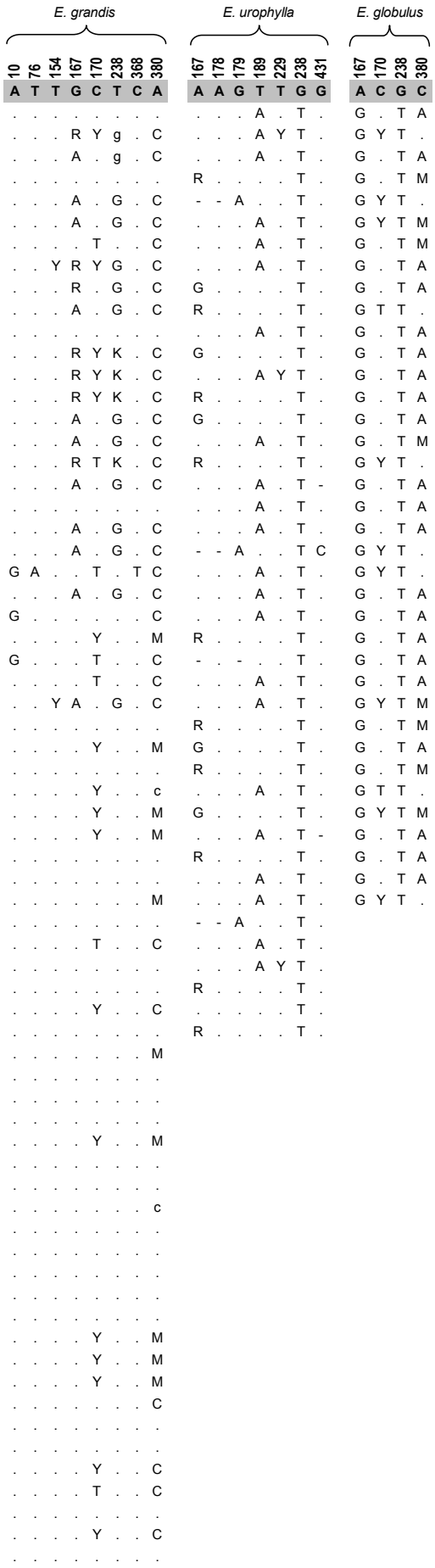


Figura 33. Relação de todos os SNPs por posição na seqüência de referência para as três espécies de eucalipto no segmento analisado do gene *ccoaomt*. No alinhamento, as variações são definidas pela sua respectiva base e a igualdade com a seqüência de referência é demonstrada por um ponto. Estão listados todos os polimorfismos para todos os indivíduos nas três espécies.

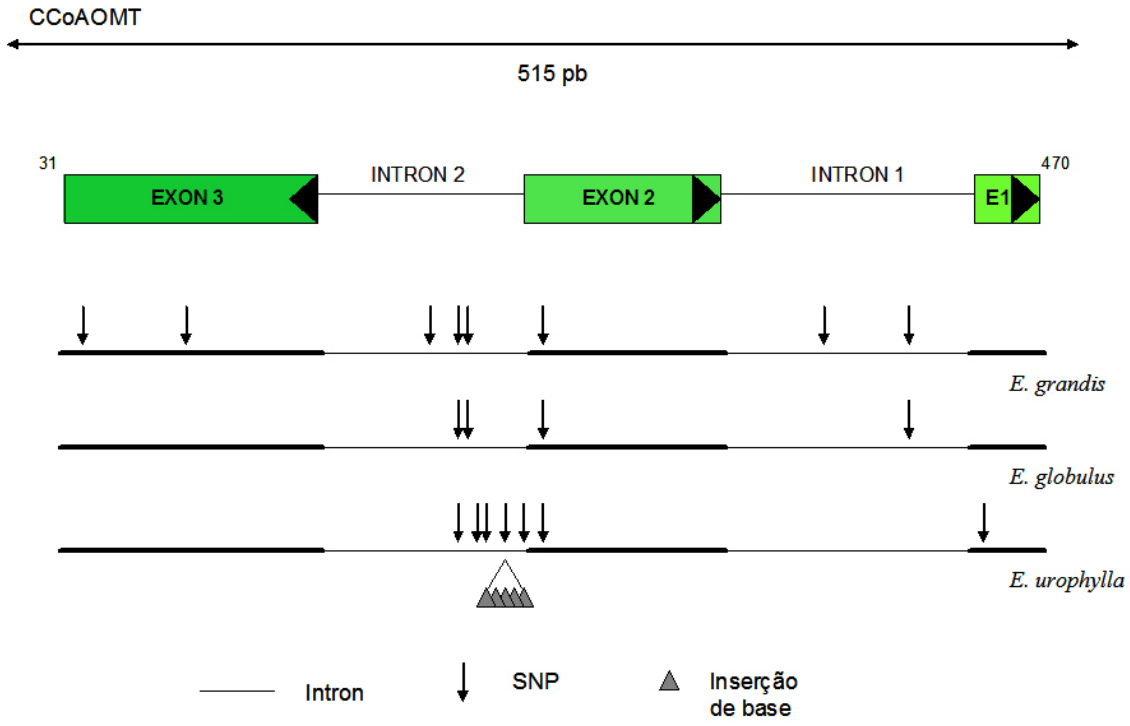


Figura 34. Estrutura genômica da região do gene *ccoaomt* amplificada e a distribuição dos SNPs identificados neste estudo. As setas indicam a posição e a quantidade de sítios variantes encontrados. No caso de *E. urophylla*, uma inserção de 5 nucleotídeos foi detectada em alguns indivíduos.

Com relação aos polimorfismos restritos à região codificadora, pode-se verificar, na espécie *E. grandis*, a existência de duas variações (posições 10 e 76) localizadas no Exon 3, resultando em 4 mutações silenciosas e uma variação na posição 238, localizada no Exon 2, onde ocorre uma troca, em alguns indivíduos, dos nucleotídeos G→T. Esse tipo de mutação, de sentido trocado, provoca a alteração do aminoácido valina por fenilalanina, gerando uma mutação sinônima (Fig. 33 e 34).

Para a espécie *E. urophylla*, uma variação na posição 229, localizada no Exon 2, gerou 3 mutações de sentido trocado resultando na alteração do aminoácido cisteína por arginina, gerando uma mutação não sinônima. Além disso, outra variação na posição 431, localizada no Exon 1, gerou 3 mutações. No entanto, essa posição apresentou dois tipos de alteração. Uma

delas foi a troca do nucleotídeo G→C, ocorrido uma única vez, que gerou uma mutação de sentido trocado, alterando o aminoácido original valina por uma leucina, ocasionando uma alteração sinônima. A outra alteração, constituiu-se de uma deleção, onde a mutação gerada foi sem sentido, ocasionando o término truncado da tradução desse exon (Fig. 33 e 34). Esse truncamento precipitado na tradução acarreta a formação parcial da proteína, e pelo fato de estar localizada no exon 1, somente alguns poucos aminoácidos formarão a proteína, tornando-a inviável. Essa deleção deverá ser mais detalhadamente estudada para possibilitar a inferência dos reais prejuízos que ela poderá causar.

Para a espécie *E. globulus*, observou-se apenas um polimorfismo em região codificadora comparando com a seqüência de referência, mas este por sua vez estava fixado para todos os indivíduos da espécie. Da mesma forma, o polimorfismo na posição 167, região de intron. Esses dados com fixação do polimorfismo, não entraram nos cálculos das análises estatísticas. (Fig. 34).

5.4. Análises estatísticas

Estimativas da diversidade nucleotídica intra-específica no segmento estudado do gene *ccoaomt* para as três diferentes espécies em estudo estão sumarizadas na tabela 9. Na tabela 10 a diversidade é relatada separadamente em regiões de exons e introns.

Entre as três espécies estudadas, *E. grandis* apresentou o maior nível de diversidade nucleotídica comparada com as outras duas espécies. Embora *E. grandis* tenha apresentado um θ inferior ao de *E. urophylla* os valores são próximos e portanto não é possível afirmar que de fato sejam diferentes. O número de haplótipos, bem como a diversidade haplotípica foi maior em *E. grandis*, mesmo apresentando igual número de sítios segregantes em *E. urophylla*. Observou-se ainda que a diversidade nucleotídica é, de uma forma geral, maior para *E. grandis* do que para as outras duas espécies (Figura 35). Vale lembrar que foram

considerados sítios segregantes (polimórficos) aqueles que apresentaram a freqüência do alelo mais raro maior ou igual a 5%.

Apesar do segmento amplificado abranger três exons, estes apresentaram menor diversidade nucleotídica do que os introns. Este resultado é esperado tendo em vista a menor pressão de seleção em regiões não traduzidas. No caso da espécie *E. grandis*, a região do exon 2 apresentou maior diversidade nucleotídica do que o intron 1, mas não excedeu à observada no intron 2. Essa situação é observada pelo fato de no exon 2, existir uma variante muito freqüente em uma região de pequena extensão, apenas 84 pb, aumentando o índice de diversidade nucleotídica.

A estatística D de Tajima visando testar a aderência dos dados observados à teoria neutra da evolução molecular apresentou valores absolutos maiores do que zero para todas as três espécies. O teste de significância sobre estas estimativas indicou que somente para a espécie E. *grandis* o valor de D é significativamente diferente de zero com valor positivo. Para as outras duas espécies o valor de D não é significativamente diferente de zero. Valores não diferentes de zero indicam aderência à neutralidade. Valores diferentes de zero indicam que a diversidade nucleotídica observada é mais dependente de alelos de alta freqüência, um indicativo da ocorrência de seleção balanceada ou subdivisão populacional e existência de um número de alelos de freqüência intermediária maior do que o esperado.

Tabela 9. Diversidade nucleotídica e teste de neutralidade para o gene *ccoaomt*.

	E. grandis	*E. globulus*	*E. urophylla*
No. seqüências	134	74	86
No. sítios	440	440	445
Índice de diversidade			
S	4	2	4
No. haplótipos	7	3	6
Diversidade haplotípica	0,695	0,497	0,671
π	0,00356	0,00168	0,00254
θ	0,00166	0,00093	0,00179
Teste de neutralidade		ns	ns
Tajima's D	2,16069*	1,30359	0,85413

S, número de sítios segregantes; π, diversidade nucleotídica.

*P<0.05

ns, não significativo; P>0,10

Tabela 10. Diversidade nucleotídica e número de polimorfismos nas regiões de íntron e exon do segmento amplificado, com 440 pb do gene *ccoaomt* .

Espécies	Análise de diversidade	Exon 1 (13pb)	Intron 1 (127pb)	Exon 2 (84pb)	Intron 2 (123pb)	Exon 3 (93pb)
E. grandis	π	0	0,00394	0,00436	0,00569	0
	θ	0	0,00144	0,00218	0,00297	0
	S	0	1	1	2	0
	si	0	-	0	-	4
	ns	0	-	0	-	0
	sn	0	-	53	-	0
E. globulus	π	0	0,0035	0	0,00239	0
	θ	0	0,00162	0	0,00167	0
	S	0	1	0	1	0
	si	0	-	0	-	0
	ns	0	0	0	0	0
	sn	0	-	0	-	0
E. urophylla	π	0	0	0,00081	0,00829	0
	θ	0	0	0,00237	0,00466	0
	S	0	0	1	3	0
	si	0	-	0	-	0
	ns	0	-	3	-	0
	sn	3	-	0	-	0

S, número de sítios segregantes; si, número de polimorfismos silenciosos; ns, número de polimorfismos não sinônimos; sn, número de polimorfismos sinônimos

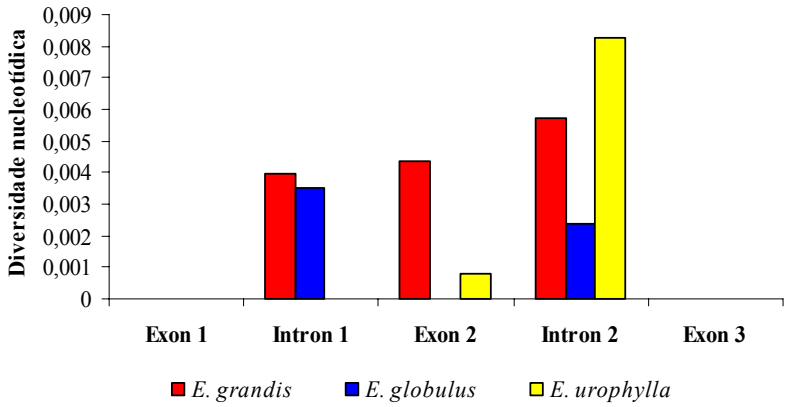


Figura 35. Diversidade nucleotídica em um segmento de 440 pb do gene *ccoaomt* em três espécies de eucalipto.

Embora a extensão de seqüência estudada seja bastante reduzida, apenas 440 pb, e muito poucos sítios polimórficos disponíveis para análise, foi tentativamente estimada a extensão do desequilíbrio de ligação para a região do gene *ccoaomt*, calculado pelo software DnaSP (ROZAS *et al.*, 2003).

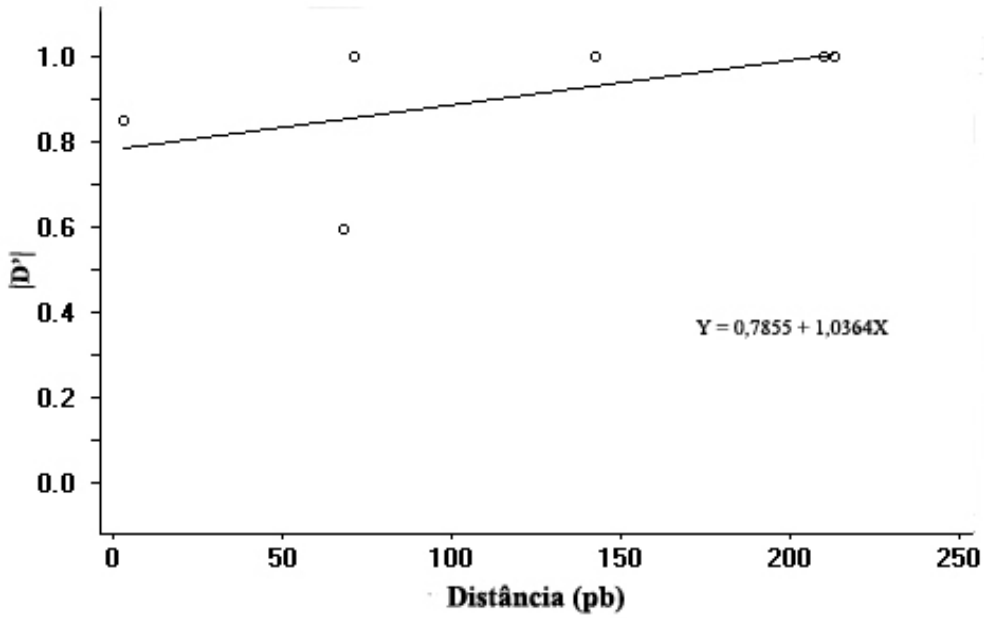
Dentre as inúmeras formas de cálculo da extensão do desequilíbrio de ligação, para o presente estudo, foram escolhidas duas estimativas, D' e r^2. De uma forma geral, o desequilíbrio de ligação significativo foi estimado para todos os alelos analisados. Para a espécie *E. grandis*, em ambos os parâmetros, D' e r^2, não se observa redução de DL com o aumento de distância, embora os valores médios absolutos de D' e r^2 sejam diferentes, com $|D'| = 0,57$ e $r^2 = 0,29$

A forma mais aceita de demonstração das medidas de desequilíbrio de ligação é realizada em forma de gráficos com a plotagem dos valores calculados par a par para os sítios polimórficos (Figura 36 e 37). Desta maneira, o método permite a percepção mais fácil do aumento ou diminuição do desequilíbrio de ligação com o aumento da distância física entre os sítios polimórficos. No caso de *E. grandis*, o gráfico define claramente a manutenção do

desequilíbrio de ligação por toda a distância compreendida pelos sítios polimórficos encontrados para esta espécie (Fig. 36). Em *E. urophylla* o comportamento foi semelhante à *E. grandis,* com os pontos variando de forma mais ampla em torno da reta de tendência ajustada (Fig. 37). Para a espécie *E. globulus*, como os sítios polimórficos eram reduzidos a dois, a inferência na forma de gráfico não foi possível, mas o valor |D'| para a espécie foi de 1,00 e o de r^2 foi de 0,44, o que significa que os polimorfismos estão em desequilíbrio de ligação.

Esta análise, embora limitada a apenas um gene e com base em poucos sítios polimórficos sugere que dentro de um gene e a distâncias menores do que aproximadamente 250 pb para a espécie *E. grandis* e aproximadamente 70 pb para a espécie *E. urophylla*, SNPs tendem a se encontrar em desequilíbrio de ligação. Os dados levantados não permitem inferências sobre a extensão do DL e muito menos sobre a tendência do comportamento do valor de DL com o aumento da distância além de 400 pb tendo em vista o número muito limitado de pontos para se ajustar uma reta de tendência.

A



B

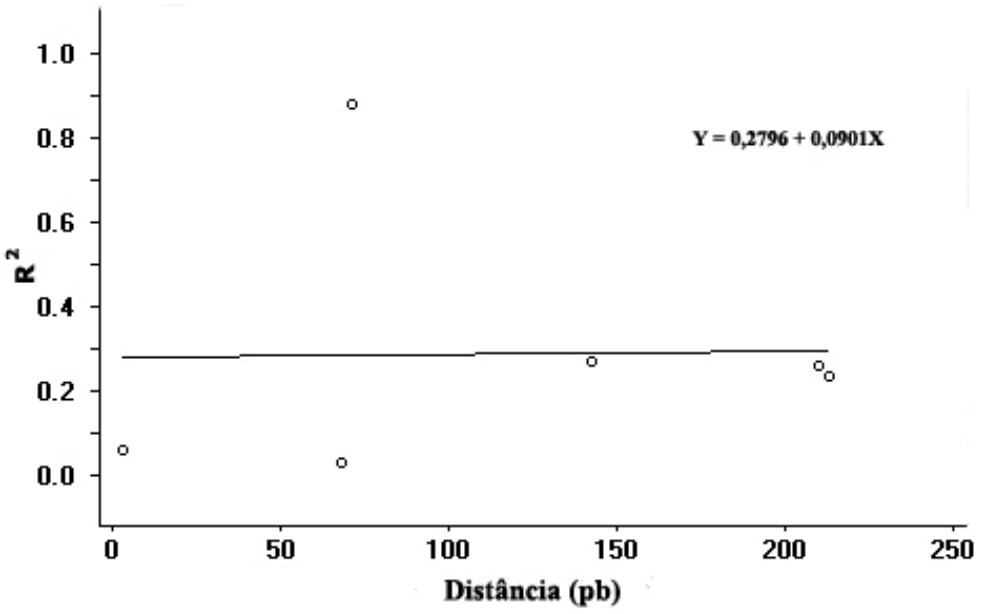
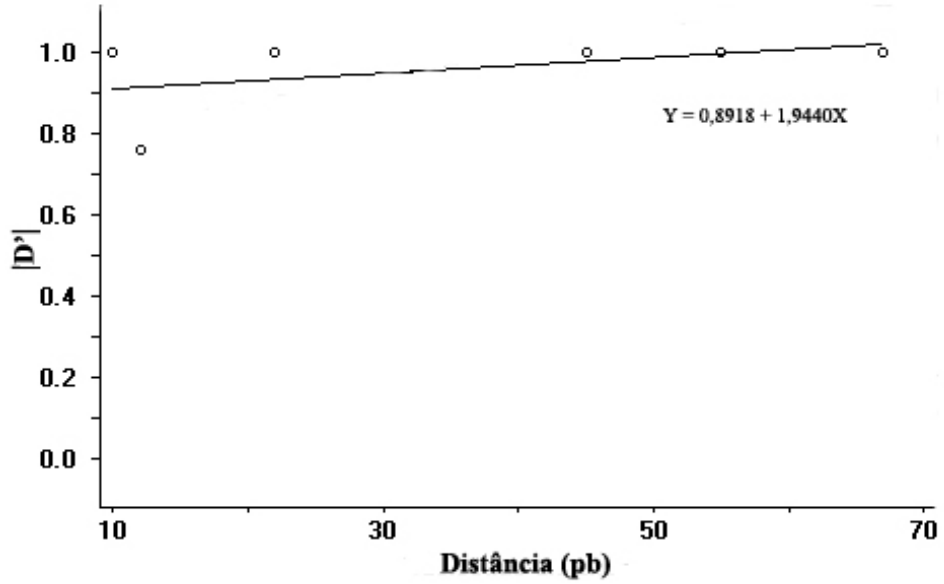


Figura 36. Desequilíbrio de ligação versus distancia para a espécie *E. grandis*. A, o gráfico utiliza o valor |D'| como medida de associação para um par de sítios variantes. B, utilização do valor R^2 como medida de associação para dois sítios variantes. Na área do gráfico encontra-se a equação da regressão calculada para a construção da linha de tendência.

A



B

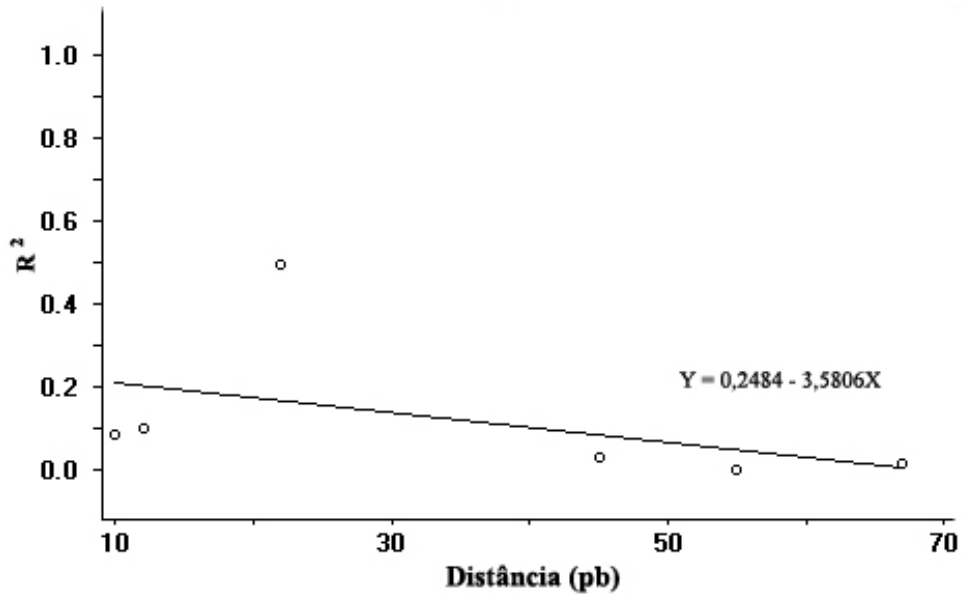


Figura 37. Desequilíbrio de ligação versus distancia para a espécie *E. urophylla*. A, o gráfico utiliza o valor |D'| como medida de associação para um par de sítios variantes. B, utilização do valor R^2 como medida de associação para dois sítios variantes. Na área do gráfico encontra-se a equação da regressão calculada para a construção da linha de tendência.

O número de haplótipos esperados para cada uma das três espécies, bem como a estrutura haplotípica estimada pelo software PHASE (STEPHENS *et al.*, 2001; STEPHENS & DONNELLY, 2003) estão descritos na tabela 11. As freqüências dos haplótipos estão ilustradas pela figura 38.

Tabela 11. Reconstrução haplotípica com as respectivas freqüências para cada uma espécies de eucalipto em estudo.

Espécies	H	n	Posições nucleotídicas																						
			167	168	169	170	171	173	176	179	180	183	186	189	190	200	210	220	234	238	250	300	350	380
		SR	G	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	C
E. grandis	H1	62	G	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	A
	H2	12	G	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	C
	H3	1	G	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	G	C	G	A	C
	H4	28	G	C	A	T	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	C
	H5	2	G	C	A	T	T	A	G	G	T	G	C	T	T	A	G	T	T	G	C	G	A	C
	H6	28	A	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	G	C	G	A	C
	H7	1	A	C	A	T	T	A	G	G	T	G	C	T	T	A	G	T	T	G	C	G	A	C
E. globulus	H1	50	G	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	A
	H2	11	G	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	C
	H3	13	G	C	A	T	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	C
E. urophylla	H1	43	A	C	A	C	T	A	G	G	T	G	C	A	T	A	G	T	T	T	C	G	A	C
	H2	3	A	C	A	C	T	A	G	G	T	G	C	A	T	A	G	T	C	T	C	G	A	C
	H3	13	A	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	C
	H4	1	A	C	A	C	T	A	G	A	T	G	C	T	T	A	G	T	T	T	C	G	A	C
	H5	21	G	C	A	C	T	A	G	G	T	G	C	T	T	A	G	T	T	T	C	G	A	C
	H6	5	G	C	A	C	T	A	G	A	T	G	C	T	T	A	G	T	T	T	C	G	A	C

n, número de seqüências com os haplótipos esperados; H, haplótipos; Sr, seqüência de referência.

H, haplótipos

SR, seqüência de referência.

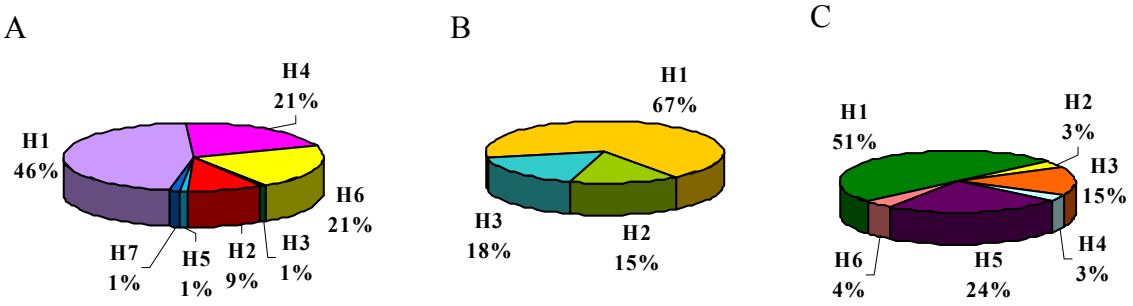


Figura 38. Freqüência haplotípica para cada uma das três espécies em estudo para uma região de 440 pb do gene *ccoaomt*. A, freqüência haplotípica para a espécie *E. grandis*; B, freqüência haplotípica para a espécie *E. globulus*; C, freqüência haplotípica para a espécie E. *urophylla*. H, haplótipos que fazem referência à tabela 5.

No total, foi estimada a existência de 7 haplótipos para a espécie *E. grandis*, 6 para *E. urophylla* e 3 para *E. globulus*. Houve o compartilhamento de três haplótipos entre as espécies, sendo que somente um era compartilhado pelas três espécies (Figura 39). Os demais

haplótipos foram específicos para as espécies, ou seja, era possível identificar a espécie por haplótipo-específico. Entre as espécies *E. grandis* e *E. globulus* ocorreu compartilhamento de sítios variantes.

Hgr2	G	C	G	T	T	T	C
Hur5	G	C	G	T	T	T	C
Hgl2	G	C	G	T	T	T	C
Hgr1	G	C	G	T	T	T	A
Hgl1	G	C	G	T	T	T	A
Hgr4	G	T	G	T	T	T	C
Hgl3	G	T	G	T	T	T	C
Hgr5	G	T	G	T	T	G	C
Hgr3	G	C	G	T	T	G	C
Hur6	G	C	A	T	T	T	C
Hur4	A	C	A	T	T	T	C
Hur3	A	C	G	T	T	T	C
Hgr7	A	T	G	T	T	G	C
Hgr6	A	C	G	T	T	G	C
Hur1	A	C	G	A	T	T	C
Hur2	A	C	G	A	C	T	C

Figura 39. Estrutura dos haplótipos para cada uma das espécies. Hgr, Hgl e Hur determinam, respectivamente, haplótipo da espécie *E. grandis, E. globulus e E. urophylla*.

Uma análise comparativa da diversidade nucleotídica entre seqüências obtidas *in silico* e seqüências obtidas através de resseqüenciamento foi realizada. A partir do contig 2 do cluster 6, o qual teve identidade significativa com o fragmento gênico amplificado para *ccoaomt,* foi selecionada uma região de 400 pb dentre as seqüências que compunham o contig. As seqüências derivadas do resseqüenciamento foram selecionadas aleatoriamente em igual número de seqüências tomadas do banco de dados, abrangendo as três espécies, com 8 seqüências escolhidas ao acaso para *E. grandis*, 7 para *E. globulus* e 7 para *E. urophylla*, totalizando 22 seqüências.

As estimativas dos parâmetros para o grupo das seqüências obtidas do banco de dados

(*in silico*) foram condizentes com o esperado, demonstrando valores menores do que o observado para o grupo das seqüências de resseqüenciamento, principalmente pela região analisada, somente exons por serem ESTs, e pelo reduzido tamanho da seqüência (Tabela 12).

Tabela 12. Comparação da diversidade nucleotídica e teste de neutralidade para o gene *ccoaomt* entre seqüências de cDNA (*in silico*) e obtidas de resseqüenciamento.

	in silico	*Ressequenciamento*
No. seqüências	22	22
No. sítios	400	440
Índice de diversidade		
S	4	5
No. haplótipos	6	8
Diversidade haplotípica	0,68	0,87
π	0,00234	0,00475
θ	0,00282	0,00312
Teste de neutralidade		
Tajima's D	- 0,48165*	1,55871*

S, número de sítios segregantes; π, diversidade nucleotídica.

* não significativo, P>0,10

Outra análise foi realizada de forma a obter inferências sobre a diversidade nucleotídica no gênero *Eucalyptus* como um todo. Para tanto, foram reunidas todas as seqüências obtidas com o seqüenciamento das três espécies em um único arquivo para gerar as informações. Os valores ficaram mais próximos ao observado para a espécie *E. grandis*, indicando S igual a 7, 13 haplótipos distintos, diversidade haplotípica de 0,864, π de 0,00457, θ de 0,00254, com D de Tajima não significativo estimado em 1,62011. Estas informações são relevantes para a determinação do potencial de se observar polimorfismo em SNPs em cruzamentos intra e inter específicos visando o mapeamento de genes bem como a utilização de SNPs em genes para estudos de mapeamento de associação.

Para *Eucalyptus grandis*, uma análise mais refinada dentro da espécie foi realizada.

Como foram utilizadas amostras de DNA de duas procedências distintas, Pine Creek (sudeste da Austrália) e Atherton (nordeste da Austrália), uma análise em separado das duas procedências foi realizada revelando reduzidos sítios polimórficos para a procedência de Atherton em comparação ao de Pine Creek (Tabela 13).

Tabela 13. Comparação da diversidade nucleotídica e teste de neutralidade para o segmento do gene *ccoaomt* entre duas procedências de *E. grandis* .

	Pine Creek	*Atherton*
No. seqüências	46	88
No. sítios	440	440
Índice de diversidade		
S	5	2
No. haplótipos	8	3
Diversidade haplotípica	0,65	0,463
π	0,00381	0,00164
θ	0,00259	0,0009
Teste de neutralidade		
Tajima's D	1,16588*	1,29667*

S, número de sítios segregantes; π, diversidade nucleotídica.

* não significativo, P>0,10

5.5. Triagem dos clones BAC para identificação de genes completos

A triagem da biblioteca de BACs e as subseqüentes confirmações com amplificação dos clones candidatos, bem como a extração de DNA em larga escala dos supostos clones, resultou no isolamento de um único clone para o gene *4cl* (Figura 40, Tabela 14) e 4 clones para o gene *ccoaomt* (Figura 41, Tabela 14).

Tabela 14. Resultado da triagem da biblioteca de BAC para os genes *ccoaomt* e *4cl* .

Gene	Superpool	Pool	Placa	Clone
4cl	F4	163-168	164	H11
	A1	1 - 6	3	E2, F3
ccoaomt	E1	49-54	50	G2
	E2	55-60	57	B5

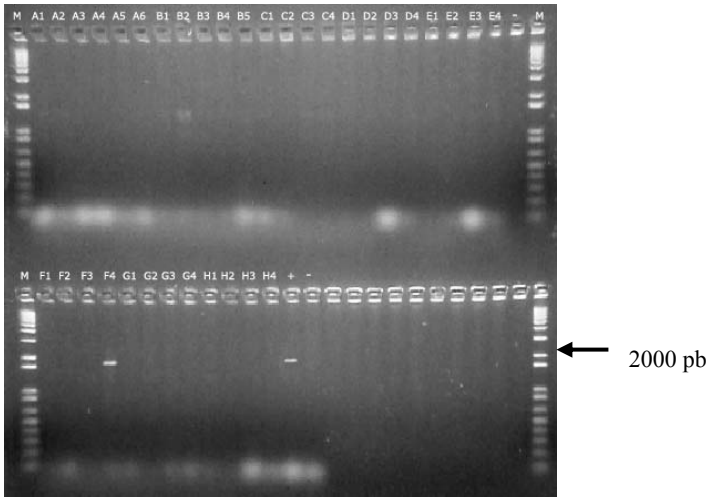
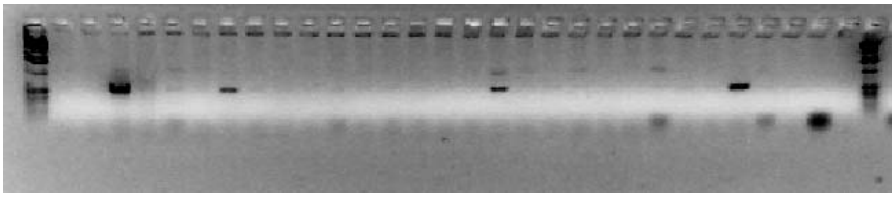


Figura 40. Triagem de biblioteca de BAC. Gel de agarose dos produtos amplificados com iniciadores para *4cl* nos 35 superpools de BAC. A-H, superpools; M, marcador 1 Kb Plus DNA ladder (Invitrogen, Carlsbad, CA).

A



B

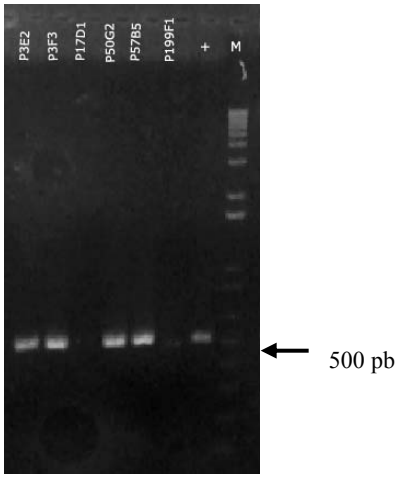


Figura 41. Identificação dos clones BAC contendo o gene *ccoaomt*. A. Gel de agarose definindo alguns superpools que continham supostos clones. B. Análise de 6 supostos clones para *ccoaomt*, onde os clones P17D1 e P199F1 são falsos positivos nos superpools, provavelmente por contaminação. M, marcador 1 Kb Plus DNA ladder (Invitrogen, Carlsbad, CA).

O clone contendo o gene *4cl* foi em seguida utilizado para a construção de uma biblioteca *shotgun*. Cerca de 960 clones desta nova biblioteca foram seqüenciados. As seqüências foram analisadas quanto à qualidade e redundância, gerando no total 1052 seqüências válidas, com aproximadamente 400 pb cada. A cobertura obtida do clone BAC alvo foi de cerca de 3 vezes do tamanho estimado para o clone que era de 130000 pb.

Com o auxílio do programa CAP3 (HUANG & MADAN, 1999) as seqüências foram montadas e um contig foi obtido com a seqüência completa do gene *4cl*.

A seqüência montada cobre um segmento contínuo de 5203 pb, regiões de exons, introns e região 3'UTR, apresentando uma região putativa de poliadenilação. As inferências quanto às regiões foram realizadas por comparação com banco de dados de *Arabidopsis* e visualizadas com o auxílio da montagem pelo programa *Artemis* (RUTHERFORD et al., 2000; <http://www.sanger.ac.uk/Software/Artemis/>) (Figura 42).

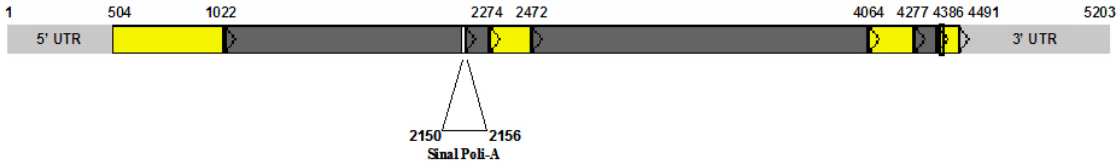


Figura 42. Estrutura do gene *4cl* obtido a partir da montagem do *shotgun* do clone BAC pelo programa *Artemis*.

Entretanto, diante das análises realizadas anteriormente neste trabalho, foi verificado, por inferências com os contigs do banco de dados do Projeto Genolyptus, que a região acima designada 5' UTR pode não estar corretamente determinada, visto que foram observadas identidades desta região com as ESTs e inferida como região de exon. Para a caracterização correta da região, se esta corresponde a uma região 5'UTR ou exônica, faz-se necessário o seqüenciamento de um maior número de clones *shotgun* para se obter uma cobertura maior do gene em estudo.

6. DISCUSSÃO

6.1 Mineração e análise das seqüências do gene *ccoaomt* no banco de dados

Um dos objetivos do presente trabalho foi a avaliação preliminar do grau de variabilidade genética em nível de seqüência nucleotídica existente em dois genes chaves na via de biossíntese de lignina. A mineração de ESTs de *Eucalyptus* foi realizada com um sistema integrado que interliga o sistema de armazenamento das seqüências do Projeto Genolyptus, o Sistema Genoma, a um programa de busca, similar ao encontrado em bancos de dados de domínio público.

A busca por ESTs similares aos genes de interesse foi realizada de duas maneiras. Para o gene *ccoaomt* foi utilizada a seqüência de referência, a qual estava presente nas análises de diversidade para justamente obter do banco seqüências que pudessem confirmar os polimorfismos encontrados naquela determinada região amplificada. A outra busca foi utilizando seqüências de isoformas disponíveis no banco de dados de domínio público para *Eucalyptus*.

A comparação dos clusters construídos pelo Sistema Genoma e retornados pela comparação com a seqüência de referência foi satisfatória no sentido em que foi possível certificar que a seqüência obtida pela amplificação com os iniciadores estava incompleta. Com a comparação entre as ESTs foi possível obter uma visão mais completa da região do gene obtida pelos fragmentos amplificados nas espécies de *Eucalyptus*.

Para o gene *ccoaomt*, os contigs 1 e 5 do cluster 6 foram idênticos e similares em 94,7% da seqüência. Nas análises das suas diferenças, foram encontradas duas regiões interessantes, uma inserção/deleção (indel) de 11 pares de base e um microssatélite dinucleotídeo, para o qual foi possível identificar, já *in silico*, polimorfismo de comprimento.

Na região do indel, por se tratar de uma inserção que aparentemente está presente em algumas espécies ou tecidos específicos, este indel poderia ser facilmente detectado utilizando os iniciadores utilizados neste trabalho. Entretanto, uma eletroforese de maior resolução seria

necessária para detectar o indel com robustez. Alternativamente, poderia ser considerada a construção de um iniciador de PCR que se anelasse na região imediatamente anterior à inserção de tal forma que somente os indivíduos que apresentassem a inserção, a região pudesse ser amplificada. Esse iniciador deverá ter o seu 3'OH livre, ou seja, a última base do iniciador deve coincidir com a posição de início da inserção, para possibilitar a amplificação da região em somente aqueles indivíduos que realmente possuírem a inserção. Essa inserção está presente no contig 5 o qual é o resultado do agrupamento de seqüências que, em sua maioria, são da espécie *E. globulus* e oriundas de xilema. Uma investigação mais detalhada da ocorrência deste indel no gene *ccoaomt* em nível intra e interespecífico, bem como inter tecidos, seria interessante para determinar a possível existência de algum padrão de distribuição e uma potencial correlação com função.

O segundo marcador detectado no gene *ccoaomt* foi um microssatélite de dinucleotídeo o qual pode fácil e rapidamente ser transformado em um marcador molecular mapeável. O conhecimento prévio da informação que este microssatélite está dentro do gene que codifica uma importante enzima da via de biossíntese de lignina pode ser vantajoso no momento da seleção de marcadores para mapeamento de QTL para características relacionadas com as propriedades da madeira, mais especificamente, teor e composição de lignina.

Um fato importante a ser considerado quanto ao microssatélite diz respeito à sua localização. Os microssatélites estão dispersos por todo o genoma mas a possibilidade de existência de microssatélite é maior em regiões de introns, visto que a alta repetição em regiões de exons e a presença de polimorfismos de comprimento não são esperadas, por influenciarem na formação da proteína, inserindo variações, alterações na sua composição ou até modificações na fase de leitura. A presença de repetições na região de 5'UTR é pouco descrita mas há estudos que demonstram a presença de elementos de repetição na extremidade

final da região 5 ‘UTR, os quais são importantes na regulação da acumulação de mRNA (DICKEY *et al*., 1998). Bhat *et al*. (2004) demonstraram que a repetição na região 5’UTR, quando presente, é suficiente para conferir uma modificação de 2,5 vezes a mais na expressão do mRNA. Portanto, este microssatélite presente na região 5’UTR necessita de estudos mais profundos quanto à influência dos polimorfismos de comprimento na regulação gênica.

A enzima CCoAOMT é uma metiltransferase, a qual retira um grupamento metil de uma adenosil metionina e adiciona ao feruloil-CoA. Em termos industriais, a presença substancial de compostos feruloil em extratos de celulose, facilita na retirada da lignina. Essa estratégia é amplamente desejada por produtores de papel que visam sua produção na extração de celulose de boa qualidade e em grandes quantidades. Possivelmente, a presença em maior quantidade da enzima CCoAOMT em espécies que possuam altos teores de lignina pode levar ao acúmulo de uma quantidade relativamente maior de unidades siringil, as quais são de extremo interesse para a indústria papeleira, além da geração de compostos feruloil que auxiliam na retirada da lignina. Tendo em vista este papel metabólico, o *ccoaomt* é um interessante gene candidato para estudos de associação. Um estudo interessante envolveria a utilização deste microssatélite possivelmente em desequilíbrio de ligação com outros polimorfismos no gene, causadores de diferenças fenotípicas, sejam eles codificadoras ou reguladores, para experimentos de genética de associação.

Uma característica que chama atenção entre os contigs obtidos do banco de dados é a alta similaridade encontrada na região 3’ UTR das seqüências. Normalmente, a região 3’ UTR apresenta-se em maiores extensões e mais polimórficas, o que não é observado nos alinhamentos entre as seqüências do banco. O que de fato é observado são seqüências com relativa similaridade entre a região de exons, introns e 3’ UTR e com singulares diferenças na região 5’ UTR, a qual era de se esperar que fosse mais semelhante entre as seqüências. Esta diferença sugere que as alterações nos níveis e composição de lignina entre espécies de

Eucalyptus podem estar relacionadas a níveis diferentes de expressão devido a alterações na região promotora e 5' UTR. Entretanto, vale lembrar que como se trata de análises baseadas em seqüências ESTs, as quais possuem qualidades baixas em suas extremidades, essas diferenças podem não ser diferenças próprias de cada contig, mas sim devido à qualidade da seqüência.

Dentre os singletons resultantes da busca no banco de dados do Projeto Genolyptus para o gene *ccoaomt*, foi obtido um derivado do seqüenciamento de pontas de BAC. Tendo em vista que ainda não foi disponibilizada uma biblioteca *shotgun* para o gene *ccoaomt*, a descoberta deste clone foi de grande valia. Desta forma, foi possível realizar uma análise não apenas com base em ESTs, mas sim parte da região gênica do gene em estudo. Em estudos por comparação, a seqüência encontrada foi comparada com duas isoformas do gene, 1 e 2, e conclui-se, por nível de similaridade, que a seqüência é pertencente a isoforma 1, descrita em *E. globulus*. Da mesma forma, buscas foram realizadas no próprio banco de dados para a isoforma 2 do mesmo gene e o resultado foi somente ESTs. Portanto, não foi encontrado no banco de dados do projeto, outra ponta de BAC que tivesse identidade com a isoforma 2.

Excetuando-se os contigs 1 e 5, os demais consensos encontrados para o gene *ccoaomt* no banco de dados do Genolyptus são similares à isoforma 2. A partir deste resultado é possível inferir que no banco de dados estão representadas as duas isoformas, e que portanto, ambas são expressas pelas espécies que constituem as bibliotecas presentes no banco de dados. Como descrito anteriormente, o contig 5 é exclusivamente constituído por ESTs advindas de xilema e portanto é provável que seja um local de exclusiva produção de isoformas 1 e nos demais locais prevaleça a expressão da isoforma 2.

6.2 Mineração e análise das seqüências do gene *4cl* no banco de dados

A busca por seqüências do gene *4cl* no banco de dados do Projeto Genolyptus foi realizada por comparações de seqüências de isoformas disponíveis no banco de dados de domínio público para o gênero *Arabidopsis*. Seqüências de aminoácidos das isoformas foram selecionadas no lugar de seqüências de nucleotídeos, principalmente para reduzir o índice de diferenças entre as espécies, visto que as seqüências de aminoácidos são mais conservadas que as de nucleotídeos. As comparações no banco de dados para cada uma das isoformas descritas para *Arabidopsis* revelaram a existência, no banco do Genolyptus, de um representante para cada uma das isoformas.

Obtidos os representantes no banco para cada uma das 4 isoformas da enzima 4CL, uma comparação foi realizada com a seqüência gênica obtida pela montagem do clone genômico derivado do seqüenciamento da biblioteca *shotgun* do BAC. O resultado revelou que a seqüência do BAC era um exemplar da isoforma 1 de 4CL.

O ATG inicial foi sugerido e a seqüência protéica resultante foi comparada com a seqüência existente para o gênero *Populus*, disponível em banco de dados de domínio público. Comparando com *Populus*, o códon inicial coincide com a seqüência gênica obtida pelo seqüenciamento do BAC, ou seja, foi obtida a seqüência codificadora completa para o gene *4cl* em *Eucalyptus grandis*, não descrita em literatura, nem disponível em banco de dados.

A partir das análises das seqüências de ESTs, foi possível inferir que no banco de dados do Genolyptus estão representadas as 4 isoformas descritas na literatura. Este resultado permite concluir que as espécies de *Eucalyptus* que estão presentes no banco, expressam as 4 isoformas mas não é possível diferenciar se a expressão é realizada em todos os estágios de desenvolvimento e em todas as partes da planta. Para tanto, fazem-se necessários estudos de expressão gênica em diferentes tecidos e estágios de desenvolvimento, via *Northern Blot* ou RT-PCR.

6.3 Amplificação de segmentos do gene 4CL

O trabalho inicial de amplificação de parte da região gênica de *4cl*, utilizando iniciadores descritos por Gion *et al.* (2000), foi parcialmente insatisfatório, por amplificar regiões em duplicidade. Esta situação provavelmente foi também observada nos experimentos realizados pelo autor, visto que a amplificação do fragmento era repetida, passando por uma etapa de excisão e eluição da banda referente ao tamanho esperado e re-amplificação do produto de PCR. Esta etapa não foi realizada neste trabalho e por este motivo não foi eliminada a amplificação múltipla para a região gênica e no momento do seqüenciamento foram observadas duplicidades de seqüências.

O desenho de novos iniciadores foi na tentativa de evitar a amplificação de múltiplos fragmentos. Nas amplificações com os novos pares de primers não foi visível a amplificação de vários fragmentos, mas no momento do seqüenciamento, a desordem de seqüências continuou a impossibilitar a leitura correta. Esse fato sugere a existência de mais de uma cópia do gene *4cl* no genoma do eucalipto. Esta, na verdade, seria a hipótese nula, visto que a participação deste em famílias gênicas, a existência de duplicidade do gene nos genomas e a existência de isoformas dentro de uma mesma espécie tem sido descrita para o gene *4cl* (CUKOVIC *et al.*, 2001; EHLTING *et al.*, 1999). Em *Populus*, por exemplo, no mínimo três diferentes isoformas foram detectadas (ALLINA *et al*., 1998; GRAND *et al*., 1983).

Assim como para *4cl*, outros genes que compõem a via de biossíntese de lignina também se apresentam em várias cópias. Entretanto, o número de cópias pode ter variação significativa entre espécies e por isso o fato de ter uma ou mais cópias em uma determinada espécie não é garantia de que na espécie alvo o comportamento seja o mesmo. Por exemplo, para o gene CCR, são descritos mais de uma cópia do gene em várias espécies de plantas. Em *Arabidopsis thaliana* 10 cópias putativas foram identificadas e em *Zea mays*, duas cópias foram reportadas (JONES *et al.*, 2001; PICHON *et al.*, 1998). Entretanto, para *Eucalyptus*

gunni uma cópia apenas foi encontrada (LACOMBE *et al*., 1997) assim como em *Eucalyptus globulus* e *E. tenuiramis*, apenas uma cópia do gene também foi descrita (POKE *et al.*, 2003).

Diante da possibilidade de existência de mais de uma cópia do gene em *Eucalyptus*, fazem-se necessários estudos que comprovem essa situação. A técnica de *Southern Blot* é uma das maneiras de se testar esta hipótese, utilizando como sonda a região amplificada pelos iniciadores, tanto os descritos por Gion *et al.* (2000) quanto os desenhados a partir de seqüência do banco de dados, específico para o gênero *Eucalyptus.*

Descrevendo o número de cópias existente no genoma das espécies de eucalipto estudadas, o próximo passo seria a descrição de quais isoformas estão representadas nesta espécie. De igual maneira, estudos que envolvam expressão gênica, identificação e comparação de seqüências seriam de grande valor, por desvendar quais isoformas estariam presentes no organismo e se haveria distinção de expressão por tecidos, órgãos ou espécie específicos.

Admitindo-se que, de fato, existem várias cópias para o gene *4cl* no genoma do eucalipto e que as regiões de anelamento dos iniciadores sejam conservadas para ambas às cópias, é possível sugerir que se trata de uma duplicação gênica. Por isso a confusão e a impossibilidade de certeza na definição das seqüências, onde os iniciadores para a reação de PCR não são suficientes para diferenciar as cópias, principalmente por anelarem em regiões idênticas, mas com a amplificação de seqüências distintas. Para solucionar esta questão, a única forma seria efetivamente a clonagem prévia dos diferentes produtos de PCR amplificados a partir do DNA genômico e o posterior seqüenciamento de uma amostra de clones. Uma outra forma de potencialmente avaliar isso seria via mapeamento genético de polimorfismos detectados dentro de uma seqüência amplificada e digeridas com enzimas de restrição (PCR-RFLP) ou via detecção e genotipagem de SNP.

6.4 Análise de polimorfismos no gene *ccoaomt*

A amplificação de segmentos do gene *ccoaomt* em três espécies de *Eucalyptus* teve como objetivo inicial o estudo da diversidade nucleotídica e haplotípica em um segmento deste gene. Foram observados SNPs tanto em regiões codificadoras como em não codificadoras deste segmento gênico.

A comparação entre os segmentos do gene *ccoaomt* se baseou na análise de uma mesma região, com igual número de bases e em confronto a uma seqüência de referência. De acordo com a seqüência de referência, para as três espécies em estudo, um total de 12 diferentes SNPs e uma inserção com 5 nucleotídeos foram os polimorfismos encontrados no fragmento do gene *ccoaomt* amplificado neste estudo. A deleção de apenas uma base em dois indivíduos de *E. urophylla* também foi encontrada.

Dentre os SNPs encontrados para os indivíduos de *E. grandis*, aquele na posição 238 implicaria na alteração do aminoácido, de uma valina para uma fenilalanina, ambos não polares, resultando em uma mutação sinônima. Esta alteração, por se encontrar na região de exon, pode ter uma grande importância na formação da estrutura final da proteína onde apenas um aminoácido modificado pode acarretar alterações drásticas na sua função. Esta hipótese deve ainda ser testada em estudos de expressão gênica e análises cristalográficas em mutantes naturais ou transgênicos, além da verificação da possibilidade de alteração de especificidade do substrato, caso o polimorfismo esteja relacionado ao sítio catalítico da enzima.

Em *E. globulus*, diferentemente de *E. grandis*, polimorfismos foram encontrados apenas nas regiões de introns. É bem descrito na literatura que a maior proporção de polimorfismo de seqüência é encontrada na região de introns, pela fraca ou ausente pressão de seleção (CHING *et al.*, 2002). Assim como em *E. globulus*, *E. grandis* e *E. urophylla* também apresentaram polimorfismos em introns. De igual maneira à *E. grandis*, *E. urophylla* também apresentou polimorfismos na região de exons.

Na posição 431, em *E. urophylla*, ocorreu uma mutação sinônima alterando o aminoácido valina por uma leucina. Esta alteração pode, de alguma forma, modificar a estrutura final da proteína mesmo conservando a característica apolar do aminoácido original. Já na posição 238, um polimorfismo em relação à seqüência utilizada como referência foi observada em *E. urophylla*. O polimorfismo estava fixado para todos os indivíduos, não caracterizando assim um polimorfismo intra-específico. Entretanto, a alteração com relação à seqüência de referência, alterou o aminoácido valina por uma fenilalanina, modificando o aminoácido mas conservando a sua característica apolar, da mesma forma como encontrado em *E. grandis*. Contudo, vale lembrar que nesta situação a alteração do aminoácido, mesmo mantendo as características apolares, pode trazer modificações significativas para a estrutura da proteína. O espaço físico necessário para o aminoácido inicial deverá ser alterado para comportar a presença de um anel aromático, presente na estrutura da fenilalanina, o que pode fisicamente trazer alterações significativas, assim como modificar a especificidade da enzima resultante, já que a alteração pode influenciar diretamente no sítio catalítico da enzima. Este resultado merece investigações mais detalhadas do efeito deste polimorfismo na atividade da enzima CCoAOMT em *E. urophylla* com relação às outras duas espécies.

Outro polimorfismo na região de exon foi detectado em *E. urophylla* na posição 229, o qual alteraria o aminoácido cisteína por uma arginina, resultando em uma mutação não sinônima. Este tipo de alteração é grave, por alterar o aminoácido e a sua característica original, inicialmente polar para um aminoácido básico. Este polimorfismo e sua conseqüente alteração na estrutura primária da enzima CCoAOMT em *E. urophylla* abrem o caminho para interessantes experimentos de bioquímica que podem gerar informações importantes sobre uma potencial relação entre fenótipos de propriedades químicas da madeira e estes polimorfismos. Vale ressaltar que *E. urophylla* é a única espécie que não ocorre no continente Australiano mas sim nas ilhas ao norte (Figura 4). Estes polimorfismos exclusivos desta

espécie e monomórficos dentro de espécie são possivelmente posteriores ao processo de especiação de *E. urophylla* e podem ser úteis para a identificação de germoplasmas, por exemplo. Além disso, a tipagem deste SNP pode ser extremamente útil para a inferência da participação de *E. urophylla* na geração de híbridos espontâneos no Brasil.

A mutação é a força geradora de variabilidade. Uma mutação causada por substituição nucleotídica, inserção/deleção, recombinação ou conversão gênica, pode se propagar na população por efeito de deriva, fluxo gênico e/ou seleção natural e, eventualmente, ser fixada na espécie (NEI & KUMAR, 2000). Entre as espécies estudadas, vários polimorfismos foram identificados como comuns e espécie-específicos. O polimorfismo na posição 167 se encontra fixado para a espécie *E. globulus*, com o alelo G em 100% dos indivíduos. Já para a espécie *E. grandis*, 73,1% dos indivíduos apresentaram o alelo G em homozigose e 10,4% em heterozigose, indicando que o alelo G ainda é o mais freqüente. Para *E. urophylla*, a diferença é significativa, com apenas 11,6% dos indivíduos homozigotos para o alelo G e a grande maioria dos indivíduos homozigotos para o alelo A. Estas diferenças em freqüências alélicas refletem o processo de especiação, que normalmente traz vantagens para a adaptação e permanência da espécie no meio ambiente, embora possa também ser apenas o resultado de efeito de deriva. Por estar localizado em região de intron, as diferenças para este polimorfismo são possivelmente resultantes de efeito de deriva por não ter conseqüências diretas na formação da proteína, mas pode haver influências indiretas, como sinais de ativação de expressão gênica, indicativo de "splicing alternativo" para a formação das isoformas.

O polimorfismo da posição 238 está fixado para a espécie *E. globulus* e *E. urophylla* e apresenta-se em 73,1% dos indivíduos em *E. grandis*. Este polimorfismo encontra-se na região de exon e a conseqüência de uma variação nesta região pode ser mais significativa que as alterações localizadas nas regiões de introns, por alterar o aminoácido, produto da leitura do códon de nucleotídeos. Por se tratar de seqüência gênica de uma enzima presente na via de

formação da madeira, este pode ser um polimorfismo importante que tem efeito direto sobre a atividade enzimática da CCoAOMT e conseqüentemente na qualidade da madeira da espécie. Assim como para o polimorfismo anteriormente descrito, estudos de associação devem ser ainda realizados para a confirmação desta hipótese.

Há outros polimorfismos que estão presentes em uma espécie e ausente em outras. No nosso estudo, foram observados os polimorfismos nas posições 10, 76, 154 e 368, presentes somente na espécie *E. grandis*, assim como nas posições 178, 179, 189, 229 e 431, presentes somente em *E. urophylla*. Os sítios polimórficos presentes em *E. globulus* são todos também observados em *E. grandis*, não apresentando polimorfismos específicos para esta espécie. Observação semelhante foi feita no gênero *Arabidopsis*, comparando os níveis de polimorfismo para o locus *Adh* (Álcool desidrogenase) para duas espécies, *A. thaliana,* com predominância de autopolinização, e *A. lyrata*, com cruzamento aberto. A distribuição dos sítios segregantes no gene foi muito diferente entre as duas espécies (INGVARSSON, 2005; SAVOLAINEN *et al.*, 2000). Esses polimorfismos espécie específicos podem ser importantes no sentido de formarem um conjunto de marcadores altamente discriminatórios, principalmente quando analisados na forma de haplótipos para a identificação de uma determinada espécie ou de híbridos delas derivados.

Em *E. urophylla* observou-se ainda um polimorfismo de grande interesse para estudos filogenéticos. No intron 2, observou-se em 53,5% dos indivíduos analisados, a presença de uma inserção de 5 nucleotídeos, TCTGT, que se apresenta como uma duplicação dos 6 nucleotídeos anteriores, ATCTGT. Todos os indivíduos que possuem essa inserção possuem o alelo A anterior à inserção, na posição 189, tornando o sítio polimórfico (Figura 43). Este alelo é indicado como sendo um sítio polimórfico específico para a espécie *E. urophylla*. Essa inserção, presente somente em indivíduos da espécie *E. urophylla*, pode também ser utilizada em programas de melhoramento e na caracterização de bancos de germoplasma.

Figura 43. Ilustração esquemática do alinhamento das seqüências obtidas por amplificação com iniciadores específicos em *E. urophylla*. Observações quanto à inserção encontrada em alguns indivíduos e a presença do alelo A antecedendo à inserção. Seqüências alinhadas pelo programa *SeqScape* com a indicação de qualidade de bases (barras em azul).

A construção de um oligonucleotídeo iniciador ancorado de forma que a última base no terminal 3' coincida exatamente com a posição 189, permitiria a amplificação exclusiva de indivíduos que apresentassem a inserção. Alternativamente, este indel de cinco bases poderia ser facilmente detectado por eletroforese em gel de poliacrilamida. Este marcador, juntamente com a análise dos vários SNPs com significativas diferenças de freqüências entre espécies serão altamente úteis para a investigação da formação de híbridos envolvendo *E. urophylla*.

6.5 Diversidade nucleotídica para segmentos do gene *ccoaomt*

Com o objetivo de quantificar a diversidade nucleotídica das espécies no segmento gênico de *ccoaomt*, foram estimados os principais parâmetros. Variações nucleotídicas foram encontradas tanto em regiões de exons quanto introns, porém, os polimorfismos encontrados na região de exons não foram admitidos nas análises de diversidade nucleotídica. Para a

inclusão dos dados obtidos nas análises estatísticas, é preciso que esses dados sejam validados no critério de número de observações. No caso dos polimorfismos de baixa freqüência, a observação de um número mínimo, correspondente a uma freqüência de 5% foi estipulada para declarar um SNP como sendo de fato polimórfico, evitando assim produtos de artefatos derivados de seqüenciamento. Com isso, os polimorfismos encontrados na região de exons para o amplicon gerado do gene *ccoaomt* não foram considerados como polimórficos e portanto eliminados das análises estatísticas.

Em plantas, a maioria dos dados de polimorfismo e diversidade em nível de nucleotídeo advém de poucas, mas bem estudadas, espécies de organismos modelos tais como *Arabidopsis*, arroz, milho e cevada. Somente alguns estudos são descritos sobre estimativas de níveis de polimorfismo de nucleotídeos e desequilíbrio de ligação em plantas de cruzamento aberto, lenhosas e perenes (INGVARSSON, 2005; DVORNYK *et al.*, 2002; GARCIA-GIL *et al.*, 2003; KADO *et al.*, 2003; POKE *et al*., 2003; KUMAR *et al*., 2004; BROWN *et al*., 2004). Agregando conhecimentos sobre o gênero *Eucalyptus* e contribuindo com informações sobre a diversidade vegetal, este estudo apresenta algumas das primeiras estimativas de diversidade nucleotídica para espécies do gênero, mais especificamente para regiões do gene *ccoaomt*.

O principal parâmetro que quantifica a diversidade nucleotídica é o parâmetro π. De uma forma geral para as três espécies, *E. grandis* (π = 0,00356) apresentou o dobro de diversidade nucleotídica do que *E. globulus* (π = 0,00168) e cerca de 1,4 vezes mais que *E. urophylla* (π = 0,00254). Esta última por sua vez, apresenta uma diversidade 1,5 vezes maior do que *E. globulus*, sendo esta a menos polimórfica das três espécies (Tabela 9). Uma estimativa de maior valor, π = 0,005, foi obtida ao se consolidar os dados das três espécies, o que se justifica pelo fato que desta forma aumenta-se o número de sítios polimórficos. Entretanto, quando analisadas as procedências de *E. grandis* em separado, a diversidade

nucleotídica para a procedência de Atherton (π = 0,00164) ficou entre o observado para as demais espécies, mas a de Pine Creek (π = 0,00381) manteve-se semelhante à estimada para a espécie *E. grandis* como um todo.

Comparações destas estimativas de diversidade nucleotídica de outras espécies arbóreas indica que de maneira geral as estimativas são congruentes. Por exemplo, em *Pinus sylvestris* a diversidade na região do *Fitocromo O* variou entre 0,0020 e 0,0037 em diferentes populações de *Pinus* para a região de introns e 0,0005 a 0,0020 para a região de exons (GARCÍA-GIL *et al*., 2003). Mesmo em espécies autógamas como *Arabidopsis lyrata*, a diversidade nucleotídica para o gene *Adh* 1 foi estimada em 0,0038 (SAVOLAINEN *et al.*, 2000), próximo ao observado em *E. grandis*.

É importante observar, entretanto, que estimativas de π podem variar em ordens de magnitude, dependendo de diversos fatores sendo os principais o sistema reprodutivo, história de vida, efeitos de amostragem, seja em termos de indivíduos e populações, bem como de quais e quantos genes são analisados. Brown *et al*. (2004) estimaram diversidade nucleotídica utilizando 32 genomas haplóides de *Pinus taeda*, uma espécie alógama florestal com ampla distribuição geográfica. Analisando um total de 18 Kb de seqüência distribuída por 19 locos, eles estimaram uma diversidade total, incluindo sítios silenciosos e não sinônimos de 0,00175 e observaram uma elevada e estatisticamente significativa heterogeneidade nesta estimativa composta (sítios silenciosos e não sinônimos consolidados) obtidas nos diferentes locos com valores variando de 0,00027 para o gene *c3h* até valores bem mais altos de 0,01728 para o gene *agp-4*. Para o gene *ccoaomt,* a diversidade nucleotídica foi estimada em 0,01975 para SNPs silenciosos, e zero para SNP não sinônimos o que gerou uma estimativa consolidada de máxima verossimilhança de 0,01200. A estimativa de π obtida neste trabalho para espécies de *Eucalyptus,* consideradas isoladamente, foi cerca de 4 a 7 vezes menor enquanto que ao se consolidar os dados das três espécies a estimativa foi de 0,005, ou seja, pouco menos da

metade do valor observado para *Pinus*. A diferença encontrada na comparação dos resultados pode ser devido ao fato de que foi analisada uma extensão de 440 pb para *Eucalyptus* enquanto que no trabalho de Brown *et al*. (2004) o segmento analisado de *ccoaomt* tinha 501 pb. Esta diferença de extensão é pequena, o que sugere que realmente a diversidade em *Eucalyptus* é menor do que em *Pinus*.

Pinus taeda possui uma distribuição geográfica muito ampla e a amostragem feita no estudo de Brown *et al*. (2004) abrangeu indivíduos de toda esta extensão. Por outro lado, o estudo em *Eucalyptus* envolveu procedências específicas de cada uma das três espécies, uma procedência por espécie, ou seja, uma população. Além disso, com exceção de *E. grandis* que tem uma ampla distribuição geográfica ao longo da costa oeste da Austrália, *E. urophylla* tem uma distribuição restrita às ilhas do norte e *E. globulus* também possui uma distribuição em manchas ao sul do continente e na ilha da Tasmânia. Estes aspectos da própria distribuição geográfica natural, juntamente com o fato de terem sido analisados indivíduos, embora geneticamente não relacionados, de uma mesma população, também pode explicar a menor diversidade encontrada. Observa-se ainda que *E. grandis*, a espécie com a maior distribuição geográfica e portanto maior oportunidade de fluxo gênico apresentou, de fato, valores mais altos de diversidade. Vale ressaltar, entretanto, que embora *E. grandis* tenha uma diversidade mais alta, não foram realizados testes estatísticos para verificar se há diferença significativa entre as estimativas obtidas para as três espécies.

Embora as estimativas de diversidade nucleotídica obtidas com base no gene *ccoaomt* se enquadrem dentro das estimativas correntes para espécies florestais alógamas, é importante ressaltar que esta estimativa deve ser considerada preliminar. Análises de um maior número de genes e maior extensão de seqüência poderá revelar estimativas mais próximas do parâmetro real e possivelmente confirmar a estimativa obtida neste trabalho a partir do gene *ccoaomt*.

Os valores de diversidade estimados pelo parâmetro θ foram semelhantes em *E. grandis* (θ = 0,00166) e *E. urophylla* (θ = 0,00179) e variaram significativamente em relação à *E. globulus*. Em comparação com humanos, o valor de θ estimado em 0,0012 (TARAZONA-SANTOS & TISHKOFF, 2005) está dentro da faixa observada para *Eucalyptus*. *Pinus taeda* apresentou novamente valor maior do que *Eucalyptus*, θ em média de 0,00407, com base nos 19 genes, muitos deles relacionados à formação da madeira (BROWN *et al*., 2004; NEALE & SAVOLAINEN, 2004).

A diversidade observada na região do gene *ccoaomt* resultou na observação de 1 SNP a cada 55 pb para *E. grandis,* 1 a cada 62,8 pb para *E. urophylla* e de 1 a cada 220 pb para *E. globulus.* Tenaillon *et al.* (2001) demonstraram em milho que a diversidade nucleotídica é de um SNP a cada 28 pb, e sugere-se que seja devido ao ativo sistema de transposon a criação de tanto polimorfismo (CHING *et al.*, 2002). Em um estudo realizado com soja foi encontrado em média, um SNP a cada 610 pb em regiões de códons enquanto em regiões não codificadoras, a média de polimorfismo é de um SNP a cada 206 pb (RAFALSKI, 2002). Para o gene *ccr* em *Eucalyptus globulus*, foi demonstrada a ocorrência de um SNP a cada 48 pb em exons e 33 pb em introns, com 50% a mais de variação calculada para a região de introns (POKE *et al.*, 2003). Neste mesmo estudo, a diversidade observada no gene *cad2* foi três vezes menor, com 1 SNP a cada 147 pb. Estes resultados, juntamente com as estimativas obtidas neste estudo para três espécies e um segmento de gene, demonstram claramente a grande variação na estimativa de diversidade nucleotídica que existe entre diferentes genes. Esta análise reforça a importância de se analisar diversos genes caso o objetivo seja a obtenção de uma estimativa mais próxima do parâmetro para o genoma como um todo. Por outro lado, também indica que estimativas de diversidade para o genoma todo possivelmente sempre terão uma elevadíssima variância associada e por isso não tenham grande relevância e utilidade. Cada gene, pela própria história evolutiva, frente aos vários eventos de mutação,

seleção, deriva, hibridização e recombinação, terá, naturalmente, uma diversidade diferente.

É esperado que em introns a taxa de mutação seja maior que em exons porque eles são não codificadoras e portanto a variação nesta região não afetará a produção da enzima, com exceção de sítios de splicing e seqüências promotoras. Contudo, é verificado que as mutações em ambas as regiões ocorrem em mesma taxa mas o quê ocorre é que nas regiões de exon, elas sofrem altas pressões de seleção e portanto não se propagam na população e são eliminadas (LEWIN, 1997). Um estudo realizado em *Zea mays* examinando o gene *Adh* 1 (GAUT & CLEGG, 1993) confirmou a presença de maior número de sítios polimórficos na região de introns. Em humanos, McCarthy *et al.* (2001) demonstraram que há variações quanto ao aparecimento de SNPs ao longo do genoma, com regiões de polimorfismos a cada 1000 pb e outras de ocorrência de um SNP a cada 50000 pb. Em *Eucalyptus*, para as três espécies em estudo, a ocorrência de SNPs na região de introns também superou à da região de exons de acordo com o esperado.

A análise comparativa realizada com os grupos de seqüências de origem do banco de dados (*in silico*) e as de resseqüenciamento por iniciadores específicos do gene de interesse demonstrou ser muito satisfatória. Os valores encontrados para os dois grupos estão condizentes com o esperado, o que sugere inferir que, para este gene e sob as condições que foram realizadas as comparações, as estimativas são compartilhadas.

A estimativa da diversidade nucleotídica foi reduzida à metade para o grupo das seqüências obtidas do banco de dados quando comparado ao grupo do resseqüenciamento, o que era esperado. A mineração de SNP *in silico,* em princípio, é uma valiosa ferramenta que rapidamente identifica potenciais marcadores a serem validados e os resultados neste estudo, sugerem que estimativas de freqüência de SNP derivados de bancos de dados de EST podem ser perfeitamente utilizados. Esta comparação é de grande valia em estudos de diversidade, já que a maioria dos projetos atualmente realizados englobam a formação de bancos de dados de

uma determinada espécie ou gênero, e as seqüências disponíveis desta forma, são seqüências EST, oriundas de cDNA.

A fim de aproveitar a massa de dados disponível, a idéia de reduzir gastos e tempo com o resseqüenciamento de fragmentos do genoma ou de genes em especifico, a mineração *in silico* poderá ser uma valiosa ferramenta para a busca das estimativas, importantes para uma diversidade de estudos, englobando QTLs, construção de mapas genéticos e físicos.

Segundo as análises preliminares realizadas com um reduzido número de seqüências e pequena extensão das seqüências analisadas, este tipo de substituição pode ser considerada condizente com a realidade encontrada nos resseqüenciamento. Porém, é importante observar que as condições utilizadas para a inferência desta comparação estão aquém do que possivelmente poderia ser encontrado. O número de seqüências utilizadas para a análise foi reduzido (apenas 22) e a extensão do fragmento analisado também foi pequena, apenas 440 pares de bases para as seqüências do resseqüenciamento, contra 400 pares de base das seqüências obtidas *in silico*. Além disso, vale ressaltar que a região analisada dos dois grupos não foi idêntica, e portanto, é esperado que haja diferenças quanto às estatísticas, visto que uma ou outra região pode apresentar mais sítios polimórficos ou maior freqüência de um alelo. A região analisada no grupo das seqüências *in silico*, pela própria propriedade das seqüências geradas, englobam apenas regiões de exons, ao contrário das seqüências de resseqüenciamento, que englobam tanto exons quanto introns. Como em regiões de exons a incidência de SNPs é menor, é esperado que haja diferenças quanto ao índice de diversidade nucleotídica entre os dois grupos, o que de fato ocorreu, reduzida à metade. Portanto, observando as ressalvas quanto à extensão do fragmento analisado e região de exon, é possivelmente viável a substituição das técnicas de resseqüenciamento pela mineração *in silico*.

6.6 O desequilíbrio de ligação no segmento amplificado do gene *ccoaomt*

O desequilíbrio de ligação, em geral, é uma conseqüência da proximidade física dos genes, ou seja, como resultado do fato dos genes estarem fisicamente presentes na mesma molécula de DNA. Entretanto, desequilíbrio de ligação pode ser criado por hibridização recente ou mantido por seleção natural se algumas combinações de alelos conferirem uma vantagem adaptativa quando juntas do que independentemente.

A extensão do desequilíbrio de ligação pode variar ao longo de todo o genoma. Em humanos, por exemplo, com hábito reprodutivo necessariamente alógamo, foram identificadas ilhas de desequilíbrio de ligação (GOLDSTEIN, 2001) onde todo o segmento é herdado conjuntamente e em outras regiões, a presença de altas taxas de recombinação. Em humanos europeus, populações estas com conhecido e recente histórico de forte efeito fundador, o comprimento do segmento em desequilíbrio de ligação pode se estender por até 60.000 pb. Isto significa que todos os genes neste segmento que envolve diferentes locos tendem a ser herdados em conjunto. Na mesma região genômica em populações de nigerianos, populações estas mais próximas de uma situação de panmixia e sem restrições recentes do tamanho populacional, o desequilíbrio de ligação nesta região se estende por menores distâncias.

Em plantas é esperado da mesma forma que fenômenos como gargalos genéticos e hábito reprodutivo autógamo favoreçam o incremento do desequilíbrio de ligação. Em espécies predominantemente auto-polinizadoras como *Arabidopsis* e arroz, o desequilíbrio de ligação se estende por longas distâncias físicas, acima de 200 Kb em *Arabidopsis* (NORDBORG, 2000) e em torno de 100 kb em arroz (GARRIS *et al.*, 2003). Por outro lado, em espécies de fecundação cruzada, como milho por exemplo, o desequilíbrio de ligação tende a declinar a níveis mínimos, a distâncias menores do que 1kb, como resultado em particular das altas taxas de recombinação (REMMINGTON *et al.*, 2001).

Neste trabalho com *Eucalyptus*, embora a extensão da seqüência estudada tenha sido

reduzida, apenas 440 pb, e muito poucos sítios polimórficos disponíveis para análise, o que resultou em muito poucos pontos no gráfico de dispersão para ajuste de curva, foi tentativamente estimada a extensão do desequilíbrio de ligação para a região do gene *ccoaomt*. Desequilíbrio de ligação significativo foi estimado ao longo de todo o segmento genômico analisado em duas das três espécies estudadas *E. grandis* e *E. urophylla,* sem redução de DL com o aumento de distância (Figuras 36 e 37). Para *E. globulus*, a baixa freqüência de polimorfismos no segmento analisado impossibilitou uma análise mais potente, embora os dados disponíveis tenham indicado também um significativo desequilíbrio de ligação ao longo de todo o segmento.

Esta análise, embora limitada a apenas um gene e com base em poucos sítios polimórficos, sugere que dentro de um gene e a distâncias menores do que aproximadamente 250 pb para a espécie *E. grandis* e aproximadamente 70 pb para a espécie *E. urophylla*, SNPs tendem a se encontrar em desequilíbrio de ligação. Os dados levantados não permitem inferências sobre o comportamento do DL além da extensão do segmento estudado e muito menos para o genoma com um todo.

Uma análise mais detalhada, envolvendo um maior número de segmentos genômicos e de maior comprimento, automaticamente geraria um maior número de estimativas de DL entre SNPs dois a dois o que resultaria em um grande número de pontos ao longo dos quais poderia ser então ajustada uma curva de tendência do decaimento do DL com a distância física com menor variância.

O teste D de Tajima foi aplicado para testar a hipótese nula de neutralidade, isto é verificar se as mutações observadas são resultantes apenas do processo natural de acúmulo de mutações neutras ou se alternativamente existe evidência a seleção natural está atuando no padrão de variação nucleotídica do segmento amplificado do gene *ccoaomt*. Valores de D não diferentes de zero indicam aderência total à teoria *neutra*; $D < 0$ sugere diversidade reduzida,

ou seja, seleção, e D > 0 sugere seleção balanceada. Analisando os valores obtidos de D, somente para *E. grandis,* o valor D de Tajima foi significativamente diferente e maior do que zero. O valor positivo reflete um excesso de alelos de freqüência intermediária, como o esperado sob um modelo de subdivisão populacional ou um antigo polimorfismo balanceado, ou seja, condizente com o modelo de seleção balanceada, embora não possa ser descartada a hipótese de um gargalo genético historicamente recente (TARAZONA-SANTOS & TISHKOFF, 2005). O valor alto D de 2,16 indica a presença de inúmeros alelos comuns e a existência de poucos variantes raros. Uma situação possível é que a população de *E. grandis* analisada, por ser originária de duas procedências, na região de Pine Creek, sudeste da Austrália, e Atherton, nordeste, abrangeu uma variedade de genótipos disponíveis nas duas regiões, em proporções médias. Além disso, a existência de algum evento relativamente recente de expansão populacional pode ter contribuído com o reduzido número de alelos raros. Alternativamente, o próprio processo de seleção branda realizado na coleta de material genético das procedências, buscando árvores de boa forma, crescimento e com sementes disponíveis, pode ter contribuído com um efeito de amostragem genética que se refletiu na diversidade ora observada no gene *ccoaomt*.

6.7 Diversidade Haplotípica

Maior diversidade haplotípica foi observada para *E. grandis* e *E. urophylla* em comparação com *E. globulus*. Houve o compartilhamento de três haplótipos entre as espécies, sendo que somente um era compartilhado pelas três espécies. Os demais haplótipos foram específicos para as espécies, ou seja, era possível identificar a espécie por haplótipo-específico, exceto para *E. globulus*. Entre as espécies *E. grandis* e *E. globulus* ocorreu compartilhamento de sítios variantes. Estes resultados estão de acordo com o esperado e distingue-se entre si pela distinta distribuição geográfica das espécies. O reduzido nível de

diversidade nucleotídica observado em *E. globulus*, está de acordo com o esperado pela distribuição geográfica e o que se sabe da história natural das espécies. *E. grandis* com ampla distribuição geográfica e ampla possibilidade de fluxo gênico; *E. urophylla* da mesma forma, embora isolada nas ilhas do norte e *E. globulus* sabidamente distribuída em populações bem estruturadas espacialmente com reduzido fluxo gênico pelas evidências elegantes de estudos de distribuição de variabilidade de haplótipos de DNA de cloroplasto (McKINNON *et al*., 2004).

6.8 Biblioteca de BAC

Neste trabalho, a biblioteca de BAC construída para a espécie *Eucalyptus grandis* permitiu a obtenção de seqüências completas do gene *4cl*, incluindo região 5'UTR e 3'UTR. A partir de um clone BAC selecionado para o gene *4cl* foi realizada uma sub-biblioteca por *Shotgun*. Um total de 1052 seqüências válidas foram utilizadas na montagem do BAC com o intuito de formar a seqüência completa do gene. Com o auxílio do programa *CAP3*, a montagem da seqüência foi realizada e gerado um fragmento com 5203 nucleotídeos, incluindo regiões de exons, intron e parte da região 3'UTR. As comparações com banco de dados de domínio público, bem como com o banco de dados do Projeto Genolyptus, foi importante para inferência quanto às posições de início e término de cada uma das regiões componentes da seqüência gênica de *4cl*.

A cobertura obtida do clone BAC alvo, com tamanho estimado de 130.000 pb, e assumindo um tamanho médio de seqüência válida de 350 pb, foi da ordem de 1052 seqüências x 400 pb = 420.800 pb, ou seja, cerca de 3 vezes. Um aspecto interessante na montagem do BAC foi o fato de que mesmo com uma cobertura relativamente pequena de apenas 3X, foi possível montar uma extensão considerável do gene, 5.203 pb apesar de não ter sido possível montar contíguos maiores do BAC pela alta freqüência de seqüências de

DNA repetitivo. Este resultado é interessante na medida que demonstra que mesmo com baixas coberturas de *shotgun* deverá ser possível obter rapidamente a montagem do gene alvo constituído obviamente por seqüências de baixa ou de cópia única as quais formam contíguos apesar da abundância de DNA repetitivo em sua volta. Este fenômeno tem sido observado no seqüenciamento *shotgun* de clones BAC de milho por exemplo durante o qual regiões ricas em genes rapidamente se consolidam em contíguos mais longos sendo que os genes aparecem em trechos únicos e contínuos de alta qualidade de seqüência (MESSING *et al*. 2004; ROUNSLEY S., com. pess.; VINSON *et al*., 2005).

A comparação da seqüência genômica do BAC com as seqüências de ESTs do banco determinou com significativa certeza que a seqüência gênica obtida não possuía ainda a região promotora mas incluía o primeiro exon com o códon inicial. Essa inferência foi obtida por comparações com o banco de dados público de proteína contra a nossa seqüência traduzida, no qual o resultado demonstrou similaridade com seqüência de 4CL na posição 1, ou seja, no "Start Codon" da seqüência. Desta maneira, por comparações com demais seqüências do banco de dados de domínio público, bem como com o banco do Genolyptus, obtivemos a seqüência codificadora completa do gene *4cl* para a espécie *Eucalyptus grandis*, não descrita ainda em literatura nem disponível em banco de dados.

A obtenção desta seqüência completa abre algumas possibilidades interessantes na continuidade deste trabalho. Por exemplo, um estudo detalhado da diversidade nucleotídica e padrões de DL ao longo deste gene podem agora ser realizado com base na utilização de vários pares de iniciadores específicos, cobrindo todo o gene. Estudos comparativos dentro da espécie e entre espécies, bem como em populações de clones fenotipados, podem ser realizados no sentido de buscar associações entre haplótipos específicos e variação quantitativa em propriedades químicas da madeira. Além disso é possível com maior probabilidade de sucesso realizar caminhamento ao longo do clone BAC via "primer

walking", para se chegar na região promotora, também de fundamental importância para estudos de associação, já que polimorfismos nessa região, podem ser determinantes no controle de expressão gênica.

7. Conclusões

As principais conclusões deste trabalho foram:

1) O banco de dados de EST do projeto Genolyptus apresenta uma riqueza de seqüências suficiente para que projetos de mineração de SNPs em genes específicos sejam desenvolvidos com sucesso. A representatividade de diferentes espécies e tecidos permite também encontrar diferentes isoformas conhecidas de um determinado gene alvo. Isto foi observado para ambos os genes estudados *ccoaomt* e *4cl*. Para *4cl*, por exemplo as 4 isoformas conhecidas estão presentes no banco.

2) O banco de dados do Projeto Genolyptus, por conter seqüências de pontas de BAC, assim como o acesso a um recurso genômico na forma de biblioteca de BAC, a qual pode facilmente ser submetida à triagem via PCR, possibilita a identificação e obtenção de fragmentos gênicos completos ou parciais. No caso da biblioteca de BAC, a seleção de um clone candidato para um determinado gene pode ser completamente seqüenciado via construção de uma sub-biblioteca por *shotgun*. Esta possibilidade foi comprovada neste trabalho seja para o gene *ccoaomt* para o qual uma ponta de BAC forneceu uma seqüência genômica parcial do gene, assim como para o gene *4cl,* para o qual foi gerada uma biblioteca *shotgun* e montada a seqüência quase completa de um gene.

3) Na análise *in silico* do gene *ccoaomt* foram encontradas duas regiões interessantes, uma inserção/deleção (indel) de 11 pares de bases e um microssatélite dinucleotídeo para o qual foi possível identificar, já *in silico*, polimorfismo de comprimento. Ambos polimorfismos poderiam se tornar marcadores interessantes para estudos de filogenia, hibridização e estudos de genética de associação.

4) Os dados de tentativa de resseqüenciamento de trechos do gene *4cl* indicaram que, de fato, existem várias cópias para o gene *4cl* no genoma do eucalipto e que as regiões de anelamento dos iniciadores utilizados são conservadas. Esta observação sugere que trata-se de uma duplicação gênica. Para se proceder com o resseqüenciamento deste gene, a única forma eficiente seria a clonagem prévia dos diferentes produtos de PCR amplificados a partir do DNA genômico e o posterior seqüenciamento de uma amostra de clones.

5) A análise de diversidade nucleotídica de fragmentos do gene *ccoaomt* amplificados de DNA genômico revelou a ocorrência de polimorfismos, seja em regiões de intron como em exons, sendo que a diversidade foi maior em regiões intrônicas. SNPs fixados ou quase fixados para determinados alelos bem como indel exclusivos foram observados em *E. urophylla*, a espécie disjunta que ocorre nas ilhas ao norte da Austrália. Estes polimorfismos podem ser altamente úteis para estudos de filogeografia da espécie e a detecção de eventos espontâneos de hibridação no Brasil. Alguns destes polimorfismos classificados como não sinônimos podem ainda ser interessantes alvos para estudos de atividade da enzima.

6) A diversidade observada na região do gene *ccoaomt* resultou na observação de 1 SNP a cada 55 pb para *E. grandis,* 1 a cada 62,8 pb para *E. urophylla* e de 1 a cada 220 pb para *E. globulus*. *E. grandis* (π = 0,00356) apresentou o dobro de diversidade nucleotídica do que *E. globulus* (π = 0,00168) e cerca de 1,4 vezes mais que *E. urophylla* (π = 0,00254). Observa-se ainda que *E. grandis*, a espécie com a maior

distribuição geográfica, e portanto maior oportunidade de fluxo gênico, apresentou de fato valores mais altos de diversidade.

7) Embora as estimativas de diversidade nucleotídica obtidas com base no gene *ccoaomt* se enquadrem dentro das estimativas correntes para espécies florestais alógamas, é importante ressaltar que esta estimativa deve ser considerada preliminar. Análises de um maior número de genes e maior extensão de seqüência poderá revelar estimativas mais próximas do parâmetro real e possivelmente confirmar a estimativa obtida neste trabalho a partir do gene *ccoaomt*.

8) Desequilíbrio de ligação significativo foi estimado ao longo de todo o segmento genômico analisado em duas das três espécies estudadas, *E. grandis* e *E. urophylla*, sem redução de DL com o aumento de distância. Esta análise, embora limitada a apenas um gene e com base em poucos sítios polimórficos, sugere que dentro de um gene e a distâncias menores do que aproximadamente 250 pb para a espécie *E. grandis* e aproximadamente 70 pb para a espécie *E. urophylla*, SNPs tendem a se encontrar em desequilíbrio de ligação. Os dados obtidos não permitem inferências sobre o comportamento do DL além da extensão do segmento estudado e muito menos para o genoma com um todo.

9) Somente para *E. grandis* o valor D de Tajima foi significativamente diferente e maior do que zero rejeitando a hipótese nula de neutralidade. O valor D de 2,16 indica a presença de inúmeros alelos comuns e a existência de poucos variantes raros. Uma possível explicação sugere que a população de *E. grandis* analisada, por ser originária de duas procedências, na região de Pine Creek, sudeste da Austrália, e Atherton,

nordeste, abrangeu uma variedade de genótipos disponíveis nas duas regiões, em proporções médias. Além disso, a existência de algum evento relativamente recente de expansão populacional pode ter contribuído com o reduzido número de alelos raros. Alternativamente, o próprio processo de seleção branda realizada na coleta de material genético desta procedência, buscando árvores de boa forma, crescimento e com sementes disponíveis, pode ter contribuído com um efeito de amostragem genética que se refletiu na diversidade ora observada no gene *ccoaomt*.

10) A diversidade haplotípica observada para *E. grandis* e *E. urophylla* foi maior em comparação com *E. globulus*. Houve o compartilhamento de três haplótipos entre as espécies, sendo que somente um era compartilhado pelas três espécies. Os demais haplótipos foram específicos para as espécies, ou seja, era possível identificar a espécie por haplótipo-específico, exceto para *E. globulus*. Entre as espécies *E. grandis* e *E. globulus* ocorreu compartilhamento de sítios variantes. Estes resultados estão de acordo com o esperado pela distribuição geográfica e o que se sabe da história natural das espécies. *E. grandis* com ampla distribuição geográfica e ampla possibilidade de fluxo gênico; *E. urophylla* da mesma forma, embora isolada nas ilhas do norte; e *E. globulus,* sabidamente distribuída em populações bem estruturadas espacialmente com reduzido fluxo gênico.

11) O resultado do seqüenciamento e montagem do clone BAC com o gene *4cl* demonstram que mesmo com baixas coberturas de *shotgun* deverá ser possível obter rapidamente a montagem de um gene alvo, apesar da abundância de DNA repetitivo, uma vez que regiões de baixa cópia rapidamente se consolidam em contíguos mais

longos sendo que os genes aparecem em trechos únicos e contínuos de alta qualidade de seqüência.

12) A obtenção da seqüência com o início de transcrição do gene *4cl* abre possibilidades interessantes de estudos detalhados da diversidade nucleotídica e padrões de DL ao longo deste gene em populações de clones fenotipados no sentido de buscar associação entre haplótipos específicos e variação quantitativa em propriedades químicas da madeira.

8. Referências Bibliográficas

ALLONA, I.; QUINN, M.; SHOOP, E.; *et al.* Analysis of xylem formation in pine by cDNA sequencing. **Proc. Natl. Acad. Sci. USA**, v. 95, p. 9693-9698, 1998.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; et al. Basic local alignment search tool. **J Mol Biol**, v. 215, p. 403-410, 1990.

ANDERSON, M.; ROBERTS, J. A. (Org.) **Arabidopsis**: Annual Plant Reviews, v.1. Boca Raton, FL : CRC Press LLC, 1998.

ANTEROLA, A. M.; LEWIS, N. G. Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. **Phytochemistry**, v. 61, p. 221-294, 2002.

ANTEROLA, A. M.; JEON, J-H.; DAVIN, L. B. *et al.* Transcriptional control of monolignol biosynthesis in *Pinus taeda:* Factors affecting monolignol ratios and carbon allocation in phenylpropanoid metabolism. **J Biol Chem**, v. 277, p. 18272-18280, 2002.

BAUCHER, M.; MONTIES, B.; VAN MONTAGU, M. *et al*. Biosynthesis and genetic engineering of lignin. **CRC Crit Rev Plant Sci**, v. 17, p. 125-197, 1998.

BAUCHER, M.; HALPIN, C.; PETIT-CONIL, M. *et al.* Lignin: Genetic Engineering and impact on Pulping. **Crit Rev Biochem Mol Biol**, v. 38, p. 305-350, 2003.

BERTOLUCCI, F.L.; RESENDE, G. D.; PENCHEL, R. Produção e utilização de híbridos de eucalipto. **Silvicultura**, v. 51, p. 12-16, 1995.

BHAT, S.; TANG, L.; KRUEGER, A. D. et al. The Fed-1 (CAUU)4 element is a 50 UTR dark-responsive mRNA instability element that functions independently of dark-induced polyribosome dissociation. **Plant Mol Biol**, v. 56, p. 761–773, 2004.

BIERMANN, C.J. **Handbook of Pulping and Papermaking**. 2nd ed. San Diego, CA: Academic Press, 1996. 754 p.

BLOUNT, J. W.; KORTH, K. L.; MASOUD, S. A.; *et al.* Altering expression of cinnamic acid 4-hydroxylase in transgenic plants provides evidence for a feedback loop the entry point into the phenylpropanoid pathway. **Plant Physiol**, v. 122, p. 107-116, 2000.

BOERJAN, W.; RALPH, J.; BAUCHER, M. Lignin biosynthesis. **Annu Rev Plant Biol**, v. 54, p. 519-546, 2003.

BOUDET, A. M.; CHABANNES, M.; GOFFNER, D. *et al*. Controlled down-regulation of genes involved in the last steps of lignin synthesis may significantly change the lignin profiles of plants. **Abstr Third Int Symp on Molecular Breeding of Wood Plants**, Tokyo, p. 1–21, 1998.

BROOKES, A. J. The essence of SNPs. **Gene**, v. 234, p. 177-186, 1999.

BROWN, G. R.; GILL, G. P.; KUNTZ, R. J.; *et al.* Nucleotide diversity and linkage disequilibrium in loblolly pine. **Proc Natl Acad Sci USA**, v. 101, n. 42, p. 15255-15260, 2004.

BRUNE, A.; ZOBEL, B. Genetic base populations, gene pools and breeding populations for *Eucalyptus* in Brazil. **Silvae genetica**, v. 30, n. 4/5, p. 146-149, 1981.

CEIMA. Sociedade espiritosantense de industrialização de madeiras, 2002. Disponível em: <http://www.ceima.com.br/euca.htm>. Acesso em 05 maio 2005.

CHABANNES, M.; RUEL, K.; YOSHINAGA, A. *et al*. In situ analysis of lignins in transgenic tobacco reveals a differential impact of individual transformations on the spatial patterns of lignin deposition at the cellular and subcellular levels. **Plant J**, v. 28, p. 271-282, 2001.

CHAPPLE, C. C. S.; VOGT, T.; ELLIS, B. E. *et al.* An Arabidopsis mutant defective in the general phenylpropanoid pathway. **Plant Cell**, v. 4, p. 1413-1424, 1992.

CHARLESWORTH, D.; BARTOLOME, C.; SCHIERUP, M. H. *et al*. Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. **Mol Biol Evol**, v. 20, n. 11, p. 1741-1753, 2003.

CHEN, C.; MEYERMANS, H.; BURGGRAEVE, B. *et al.* Cell-specific and conditional expression of caffeoyl-CoA O-methyltransferase in poplar. **Plant Physiol**, v. 123, n. 3, p. 853-868, 2000.

CHERNEY, J. H.; CHERNEY, D. J. R.; AKIN, D. E. *et al.* Potential of brown-midrib, low-lignin mutants for improving forage quality. **Advances in Agronomy**, v. 46, p. 157-198,

1991.

CHING, A.; CALDWELL, K. S.; JUNG, M. *et al*. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genet**, v. 3, n. 19, 2002.

CHRISTENSEN, J. H.; BAUCHER, M.; BOERJAN, W. *et al.*. Control of lignin biosynthesis. In: JAIN, S. M.; MINOCHA, S. C. (Org.). **Molecular Biology of Woody Plants***,* Volume 1. Norwell, USA: Kluwer Academic Publishers, 2000. 520 p.

CLAMP, M. ; CUFF, J.; SEARLE, S. M.; et al. The Jalview Java Alignment Editor. **Bioinformatics**, v. 12, p. 426-427, 2004.

CLARK, R. M.; LINTON, E.; MESSING, J. *et al*. Patterns of diversity in the genomic region near the maize domestication gene *tb1*. **Proc Natl Acad Sci USA**, v. 101, p. 700-707, 2004.

CREGAN, P.; RANDALL, N.; YOULIN, Z. Sequence variation, haplotype diversity and linkage disequilibrium in cultivated and wild soybean. In: **First International Conference on Legume Genomics and Genetics**. Plenary Session IV: Structural Genomics. Minneapolis-St. Paul, 2002.

CUKOVIC, D.; EHLTING, J.; VANZIFFLE, J. A. *et al*. Structure and evolution of 4-coumarate:coenzyme A ligase (4CL) gene families. **Biol Chem**, v. 382, n. 4, p. 645-54, 2001.

DEAN, J. F. D. Synthesis of lignin in transgenic and mutant plants. In: STEINBÜCHEL, A.; DOI, Y. **Biotechnology of biopolymers:** from synthesis to patents. Weinheim: Wiley-VCH

Verlang GmbH & Co., 2005. p. 3-26.

DELMER, D. P.; AMOR, Y. Cellulose biosynthesis. **Plant Cell**, v.7, p. 987-1000, 1995.

DICKEY, L.; PETRACEK, M., NGUYEN, T. *et al*. Light regulation of Fed-1 mRNA requires an element in the 50 untranslated region and correlates with differential polyribosome association. **Plant Cell**, v. 10, p. 475-484, 1998.

DOYLE, J. J.; DOYLE, J.L. Isolation of plant DNA from fresh tissue. **Focus**, v. 12, n. 1, p. 13-15, 1991.

DRAZEN, J. M.; YANDAVA, C. N.; DUBE, L. *et al.* Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. **Nat Genet**, v. 22, p. 168-170, 1999.

DVORNYK, V.; SIRVIÖ, A.; MIKKONEN, M.; *et al.* Low Nucleotide Diversity at the pal1 locus in the widely distributed Pinus sylvestris. **Mol Biol Evol**, v. 19, n. 2, p. 179-188, 2002.

EHLTING, J.; BÜTTNER, D.; WANG, Q.; *et al.* Three 4-coumarate:coenzyme A ligases in *Arabidopsis thaliana* represent two evolutionarily divergent classes in angiosperms. **Plant J**, v. 19, p. 9-20, 1999.

ELDRIDGE, K.; DAVIDSON, J.; HARWOOD, C. *et al*. **Eucalypt domestication and breeding**. New York: Oxford University Press, 1994. 288 p.

EWING, B.; GREEN, P. Basecalling of automated sequencer traces using phred. II. Error probabilities. **Genome Res**, v. 8, p. 186-194, 1998.

EWING, B.; HILLIER, L.; WENDL, M. *et al*. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Res**, v. 8, p. 175-185, 1998.

EXCOFFIER, L.; SLATKIN, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. **Mol Biol Evol**, v. 12, n. 5, p. 921-927, 1995.

FAO - FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **Pulp and Paper Capacities** - Survey 2000–2005. Rome: FAO, 2001.

FERREIRA, M. Escolha de Espécies de Eucalipto. **Circular Técnica IPEF**, v. 47, p. 1-30, 1979.

FERREIRA, M. E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética.** Brasília: Empresa Brasileira de Pesquisa Agropecuária-Embrapa, 1998. 220p.

FERRER, J-L.; ZUBIETA, C.; DIXON, R. A. *et al.* Crystal structures os alfalfa caffeoyl coenzyme A 3-O-methyltransferase. **Plant Physiol**, v. 137, p. 1009-1017, 2005.

FLINT-GARCIA, S. A.; THORNSBERRY, J. M.; BUCKLER VI, E. S. Structure of linkage disequilibrium in plants. **Annu Rev Plant Biol**, v. 54, p. 357-374, 2003.

FREUDENBERG, K.; NEISH, A. C. **Constitution and Biosynthesis of Lignin**. Berlin: Springer-Verlag. 1968. 129 p.

GARCÍA-GIL, M. R.; MIKKONEN, M.; SAVOLAINEN, O. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. **Mol Ecol**, v. 12, p. 1195-1206, 2003.

GARNIER-GÉRÉ, P.; BEDON, F.; POT, D. *et al*. DNA sequence polymorphism, linkage disequilibrium and haplotype structure in candidate genes of wood quality traits in maritime pine *Pinus pinaster* Ait. In: **11th international conference of the Plant and Animal Genome**. San Diego. 2003. P241.

GARRIS, A. J.; MCCOUCH, S. R.; KRESOVICH, S. Population structure and its e.ects on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice Oryza sativa L. **Genetics**, v. 165, p. 759-769, 2003.

GASTEIGER, E.; JUNG E.; BAIROCH, A. SWISS-PROT: Connecting biological knowledge via a protein database. **Curr Issues Mol Biol**, v. 3, p. 47-55, 2001.

GASTEIGER, E.; GATTIKER, A.; HOOGLAND, C. *et al.* ExPASy: the proteomics server for in-depth protein knowledge and analysis. **Nucleic Acids Res**, v. 31, p. 3784-3788, 2003.

GAUT, B. S.; CLEGG, M. T. Molecular evolution of the *Adh* 1 locus in the genus *Zea*. **Proc Natl Acad Sci USA**, v. 90, p. 5095-5099, 1993.

GAUT, B. S.; LONG, A. D. The Lowdown on Linkage Disequilibrium. **Plant Cell**, v. 15, p. 1502-1506, 2003.

GION, J-M.; RECH, P.; GRIMA-PETTENATI, J. *et al.* Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. **Mol Breed**, v. 6, p. 441-449, 2000

GISH, W.; STATES, D. J. Identification of protein coding regions by database similarity search. **Nat Genet**, v. 3, p. 266-272, 1993.

GOLDSTEIN, D. B. Islands of linkage disequilibrium. **Nat Genet**, v. 29, p. 109-111, 2001.

GONZÁLEZ-MARTÍNEZ, S. C.; BROWN, G. R.; ERSOZ, E. *et al.* Nucleotide diversity, linkage disequilibrium and adaptive variation in natural populations of loblolly pine. In: **12nd Conference International Plant and Animal Genomes**, San Diego, 2004. W3.

GRAND, C.; BOUDET, A.; BOUDET, A. M. Isoenzymes of hydroxycinnamate:CoA ligase from poplar stems: properties in tissue distribution. **Planta**, v. 158, p. 225-229, 1983.

GRATTAPAGLIA, D. Genômica florestal. In: MIR, L. (org.). **Genômica**. São Paulo: Atheneu, 2003. p. 917-934.

GRIFFITHS, A. J. F.; GELBART, W. M.; MILLER, J. H. *et al.* **Genética Moderna**. Rio de Janeiro: Guanabara Koogan, 2001. 589 p.

GUPTA, P. K.; RUSTGI, S.; KULWAL, P. L. Linkage disequilibrium and association studies

in higher plants: present status and future prospects. **Plant Mol Biol**, v. 57, p. 461-485, 2005.

HAGENBLAD, J.; NORDBORG, M. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. **Genetics**, v. 161, p. 289-298, 2002.

HARDING, S. A.; LESHKEVICH, J.; CHIANG, V. L.; *et al.* Differential substrate inhibition couples kinetically distinct 4- coumarate: coenzyme A ligases with spatially distinct metabolic roles in quaking aspen. **Plant Physiol**, v. 128, n. 2, p. 428-438, 2002.

HAWKINS, S.; GOFFNER, D.; BOUDET, A. M. Cinnamyl alcohol dehydrogenase polymorphism and its potential role in the control of lignin heterogeneity. **Acta Horticulturae** (ISHS), v. 381, p. 280-286, 1994. Disponível em: <http://www.actahort.org /books/381/381_35.htm>. Acesso em 09 Ago. 2005.

HAMRICK, J. L.; GODT, M. J. W. Effects of life history traits on genetic diversity in plant species. **Philos Trans R Soc Lond B Biol Sci**, v. 351, n. 1345, p. 1291-1298, 1996.

HEDRICK, P. W. Gametic disequilibrium measures: proceed with caution. **Genetics**, v. 117, p. 331-341, 1987.

HIGGINS, D.; THOMPSON, J.; GIBSON, T. *et al*. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res**, v. 22, p. 4673-4680, 1994.

HU, W-J.; KAWAOKA, A.; TSAI, C-J. *et al.* Compartmentalized expression of two

structurally and functionally distinct 4-coumarate: CoA ligase genes in aspen *Populus tremuloides*. **Proc Natl Acad Sci USA**, v. 95, n. 9, p. 5407–5412, 1998.

HU, W-J.; HARDING, S. A.; LUNG, J. *et al.* Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. **Nat Biotechnol**, v. 17, n. 8, p. 808-812, 1999.

HUANG, X.; MADAN, A. CAP3: A DNA sequence assembly program. **Genome Res**, v. 9, p. 868-877, 1999.

HUMPHREYS, J. M.; HEMM, M. R.; CHAPPLE, C. New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. **Proc Natl Acad Sci USA**, v. 96, p. 10045-10050, 1999.

INGVARSSON, P. K. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula L.*, Salicaceae). **Genetics**, v. 169, p. 945-953, 2005.

INOUE, K.; SEWALT, V. J. H.; BALLANCE, G. M. *et al*. Developmental expression and substrate specificities of alfafa caffeic acid 3-*O*-methyltransferase in caffeoyl coenzyme A 3-*O*-methyltransferase in relation to lignification. **Plant Physiol**, v. 117, p. 761-770, 1998.

IPEF - Instituto de Pesquisas e Estudos Florestais. **Chave de Identificação de Espécies Florestais**. Piracicaba-SP, 2004. Disponível em:

<http://www.ipef.br/identificacao/cief/especies/grandis.asp>. Acesso em 05 mai. 2005.

JONES, L.; ENNOS, A. R.; TURNER, S. R. Cloning and characterization of *irregular xylem4* (*irx4*): a severely lignin-deficient mutant of *Arabidopsis*. **Plant J**, v. 26, n. 2, p. 205-216, 2001.

JOUANIN, L.; GOUJON, T.; NADAÏ, V. *et al.* Lignification in transgenic poplars with extremely reduced caffeic acid o-methyltransferase activity. **Plant Physiol**, v. 123, n. 4, p. 1363-1373, 2000.

KADO, T.; YOSHIMARU, H.; TSUMURA, Y. *et al*. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae sensu lato). **Genetics**, v. 164, n. 4, p. 1547-1559, 2003.

KAJITA, S.; KATAYAMA, Y.; OMORI, S. Alterations in the biosynthesis of lignin in transgenic plants with chimeric genes for 4-coumarate:coenzyme A ligase. **Plant Cell Physiol**, v. 37, n. 7, p. 957-965, 1996.

KAJITA, S.; HISHIYAMA, S.; TOMIMURA, Y. *et al.* Structural characterization of modified lignin in transgenic tobbaco plants in which the activity of 4-coumarate:coenzyme A ligase is depressed. **Plant Physiol**, v. 114, p. 871-879, 1997a.

KAJITA, S.; MASHIRO, Y.; NISHIKUBO, N. *et al.* Immunological characterization of transgenic tobacco plants with a chimeric gene for 4-coumarate:CoA ligase that have altered lignin in their xylem tissue. **Plant Sci**, v. 128, p. 109-118, 1997b.

KIMURA, M.; CROW, J. F. The number of alleles that can be maintained in a finite population. **Genetics**, v. 49, p. 725-738, 1964.

KIMURA, M. Evolutionary rate at the molecular level. **Nature**, v. 217, p. 624-626, 1968.

KIMURA, M. **The neutral theory of molecular evolution**. Cambridge: Cambridge University Press, 1985. 384 p.

KOSIKOVA, B.; DURIS, M.; DEMIANOVA, V. Conversion of lignin biopolymer into surface-active derivates. **European Polymer Journal**, v. 36, p. 1209-1212, 2000.

KRAWEZAK, M.; REISS, J.; COOPER, D. N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. **Hum Genet**, v. 90, p. 41-54, 1992.

KRUGLYAK, L. The use of a genetic map of biallelic markers in linkage studies. **Nat Genet,** v. 17, p. 21-24, 1997.

KRUGLYAK, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. **Nat Genet**, v. 22, p. 139-144, 1999.

KUC, J.; NELSON, O. The abnormal lignins produced by the brown-midrib mutants of maize. I. The brown-midrib-1 mutant. **Arch Biochem Biophys**, v. 105, p. 103-113, 1964.

KÜHNL, T.; KOCH, U.; HELLER, W. *et al*. Elicitor induced s-adenosyl-l-methionine -

caffeoyl-coa 3-o-methyltransferase from carrot cell-suspension cultures. **Plant Sci**, v. 60, n. 1, p. 21-25, 1989.

KUMAR, S.; ECHT, C.; WILCOX, P. L. *et al*. Testing for linkage disequilibrium in the New Zealand radiata pine breeding population. **Theor Appl Genet**, v. 108, p. 292-298, 2004.

LACOMBE, E.; HAWKINS, S.; DOORSSELAERE, J. V. *et al*. Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: cloning, expression and phylogenetic relationships. **Plant J**, v. 11, n. 3, p. 429-441, 1997.

LAMB, C. J. *trans*-cinnamic acid as a mediator of the light-stimulated increase in hydroxycinnamoyl-CoA: quinate hydroxycinnamoyl transferase. **FEBS Lett**, v. 75, n. 1, p. 37-40, 1977.

LAUVERGEAT, V.; LACOMME, C.; LACOMBE, E. *et al.* Two cinnamoyl-CoA reductase (CCR) genes from Arabidopsis thaliana are differentially expressed during development and in response to infection with pathogenic bacteria. **Phytochemistry**, v. 57, p. 1187-1195, 2001.

LEE, D.; MEYER, K.; CHAPPLE, C. *et al. Anti-sense* suppression of 4- coumarate: coenzyme A ligase activity in Arabidopsis leads to altered lignin subunit composition. **Plant Cell**, v. 9, p. 1985-1998, 1997.

LEWIN, B. **Genes VI**. New York: Oxford University Press, 1997. 1260 p.

LEWONTIN, R. C. The interaction of selection and linkage. I. General considerations;

heterotic models. **Genetics**, v. 49, p. 49-67, 1964.

LI, L.; OSAKABE, Y.; JOSHI, C. P. *et al*. Secondary xylem-specific expression of caffeoyl-coenzyme A 3-O-methyltransferase plays an important role in the methylation pathway associated with lignin biosynthesis in loblolly pine. **Plant Mol Biol**, v. 40, p. 555-565, 1999.

LI, L.; POPKO, J. L.; UMEZAWA, T. *et al.* 5-hydroxyconiferyl aldehyde modulates enzymatic methylation for syringyl monolignol formation, a new view of monolignol biosyntesis in angiosperms. **J Biol Chem**, v. 275, n. 9, p. 6537-6545, 2000.

LI, Y.; SARKANEN, S. Thermoplastic with very high lignin contents. **ACS Symp. Ser**, v. 742, p. 351-366, 2000.

LI, L.; ZHOU, Y.; CHENG, X. *et al.* Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. **Proc Natl Acad Sci USA**, v. 100, n. 8, p. 4939-4944, 2003.

LIMA, W. P. **Impacto ambiental do eucalipto**. São Paulo: EDUSP, 1993. 301p.

LINDERMAYR, C.; MÖLLERS, B.; FLIEGMANN, J. *et al.* Divergent members of a soybean (Glycine max L.) 4-coumarate:coenzyme A ligase gene family. **Eur J Biochem**, v. 269, p. 1304-1315, 2002.

MARTZ, F.; MAURY, S.; PINÇON, G. *et al*. cDNA cloning, substrate specificity and expression study of tobacco caffeoyl-CoA 3-O-methyltransferase, a lignin biosynthetic

enzyme. **Plant Mol Biol**, v. 36, p. 427-437, 1998.

McCARTHY, L. C.; HOSFORD, D. A.; RILEY, J. H. *et al*. Single-nucleotide polymorphism alleles in the insulin receptor gene are associated with typical migraine. **Genomics**, v. 78, p. 135-149, 2001.

McKINNON, G. E.; JORDAN, G. J.; VAILLANCOURT, R. E. *et al*. Glacial refugia and reticulate evolution: the case of the Tasmanian eucalypts. **Philos Trans R Soc Lond B Biol Sci**, v. 29, n. 359, p. 275-284, 2004.

MESSING, J.; BHARTI, A. K.; KARLOWSKI, W. M. *et al*. Sequence composition and genome organization of maize. **Proc Natl Acad Sci USA**, v. 101, n. 40, p. 14349-14354, 2004.

MESSNER, K.; SREBOTNIK, E. Biopulping: an overview of developments in an environmentally safe paper-making technology. **FEMS Microbiol Rev**, v. 13, p. 351-364, 1994.

MEYERMANS, H.; MORREEL, K.; LAPIERRE, C. *et al.* Modifications in lignin and accumulation of phenolic glucosides in poplar xylem upon down-regulation of caffeoyl-coenzyme A O-methyltransferase, an enzyme involved in lignin biosynthesis. **J Biol Chem**, v. 275, p. 36899-36909, 2000.

MORGANTE, M.; SALAMINI, F. From plant genomics to breeding practice. **Curr Opin Biotechnol**, v. 14, p. 214-219, 2003.

MOTT, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. **Comput Appl Biosci**, v. 13, p. 477-478, 1997.

NEALE, D. B.; SAVOLAINEN, O. Association genetics of complex traits in conifers. **Trends Plant Sci**, v. 9, n. 7, p. 325-330, 2004.

NEI, M.; LI, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. **Proc Natl Acad Sci USA**, v. 76, n. 10, p. 5269-5273, 1979.

NEI, M. **Molecular evolutionary genetics**. New York: Columbia University Press, 1987. 512 p.

NEI, M.; KUMAR, S. **Molecular evolution and phylogenetics**. New York: Oxford University press, 2000. 333 p.

NICHOLSON, R. L.; HAMMERSCHMIDT, R. Phenolic compounds and their role in disease resistance. **Annu Rev Phytopathol**, v. 30, p. 369-389, 1992.

NICKERSON, D. A.; TAYLOR, S. L.; WEISS, K. M. *et al.* DNA sequence diversity in 9.7 kb region of the human lipoprotein lipase gene. **Nat Genet**, v. 19, p. 233-240, 1998.

NORDBORG, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. **Genetics**, v. 154, p. 923-929, 2000.

OLSEN, K. M.; HALLDORSDOTTIR, S. S.; STINCHCOMBE, J. R. *et al*. Linkage disequilibrium mapping of *Arabidopsis* CRY2 flowering time alleles. **Genetics**, v. 167, p. 1361-1369, 2004.

OSAKABE, K.; TSAO, C. C.; LI, L. *et al.* Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms. **Proc Natl Acad Sci USA**, v. 96, n. 16, p. 8955-8960, 1999.

PAKUSCH, A. E.; KNEUSEL, R. E.; MATERN, U. S-adenosyl-L-methionine:trans-caffeoyl-coenzyme A 3-O-methyltransferase from elicitor-treated parsley cell suspension cultures. **Arch Biochem Biophys**, v. 271, p. 488-494, 1989.

PARVATHI, K.; CHEN, F.; GUO, D. *et al.* Substrate preferences of *O*-methyltransferases in alfalfa suggests new pathways for 3-*O*-methylation of monolignols. **Plant J**, v. 25, n. 2, p. 193-202, 2001.

PEREIRA, J. C. D.; STURION, J. A.; HIGA, A. R. *et al*. **Características da madeira de algumas espécies de eucalipto plantadas no Brasil**. Colombo: Embrapa Florestas, 2000. 113p.

PICHON, M.; COURBOU, I.; BECKERT, M. *et al*. Cloning and characterization of two maize cDNAs encoding cinnamoyl-coA reductase (CCR) and differential expression of the corresponding genes. **Plant Mol Biol**, v. 38, p. 671-676, 1998.

PINÇON, G.; CHABANNES, M.; LAPIERRE, C. *et al.* Simultaneous down-regulation of

caffeic/5- hydroxy ferulic acid-O-methyltransferase I and cinnamoyl-coenzyme A reductase in the progeny from a cross between tobacco lines homozygous for each transgene. Consequences for plant development and lignin synthesis. **Plant Physiol**, v. 126, p. 145-155, 2001a.

PINÇON, G.; MAURY, S.; HOFFMANN, L. *et al.* Repression of *O*-methyltransferase genes in transgenic tobacco affects lignin synthesis and plant growth. **Phytochemistry**, v. 57, p. 1167-1176, 2001b.

PIQUEMAL, J.; LAPIERRE, C.; MYTON, K. *et al*. Down-regulation of cinnamoylcoA reductase induces significant changes of lignin profiles in transgenic tobacco plants. **Plant J**, v. 13, p. 71–83, 1998.

POKE, F. S.; VAILLANCOURT, R. E.; ELLIOTT, R. C. *et al*. Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase 2 (CAD2). **Mol Breed**, v. 12, p. 107-118, 2003

PRITCHARD, J. K.; PRZEWORSKI, M. Linkage disequilibrium in humans: Models and data. **Am J Hum Genet**, v. 69, n. 1, p. 1-14, 2001.

PRZEWORSKI, M. The signature of positive selection at randomly chosen loci. **Genetics**, v. 160, n. 3, p. 1179-1189, 2002.

RAFALSKI, A. Applications of single nucleotide polymorphisms in crop genetics. **Curr Opin Plant Biol**, v. 5, n. 2, p. 94-100, 2002.

RALPH, J. Recent advances in characterizing 'non-traditional' lignins and lignin-polysaccharide cross-linking. **9th International symposium on wood and pulping chemistry**, Montreal, Quebec, 1997. PL2.

RALPH, J.; MACKAY, J. J.; HATFIELD, R. D. *et al*. Abnormal lignin in a loblolly pine mutant. **Science**, v. 277, p. 235-239, 1997.

RALPH, J.; LAPIERRE, C.; MARITA, J. M. *et al.* Elucidation of new structures in lignins of CAD- and COMT-deficient plants by NMR. **Phytochemistry**, v. 57, p. 993-1003, 2001.

REMINGTON, D. L.; THORNSBERRY, J. M.; MATSUOKA, Y. *et al.* Structure of linkage disequilibrium and phenotypic associations in the maize genome. **Proc Natl Acad Sci USA**, v. 98, p. 11479-11484, 2001.

REVISTA DA MADEIRA. **No eucalipto a opção de futuro**. n. 31, 1997, p. 36-39.

REVISTA MADEIRA. **A madeira de eucalipto na indústria moveleira**. n. 70, ano 12, 2003. Mensal. Disponível em:< http://www.remade.com.br/revista/materia.php?edicao=70 &id=297>. Acesso em: 05 mai. 2005.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: the European Molecular Biology Open Software Suite. **Trends Genet**, v. 16 (6), p. 276-277, 2000.

ROZAS, J.; SÁNCHEZ-DELBARIO, J. C.; MESSENGER, X. *et al.* DnaSP, DNA polymorphism analyses by the coalescent and other methods. **Bioinformatics**, v. 19, n. 18, p.

2496-2497, 2003.

ROZEN, S.; SKALETSKY, H. J. Primer3 on the WWW for general users and for biologist programmers. **Methods Mol Biol**, v. 132, p. 365-386, 2000.

RUSSELL, J.; BOOTH, A.; FULLER, J. *et al.* A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. **Genome**, v. 47, p. 389-398, 2004.

RUTHERFORD, K.; PARKHILL, J.; CROOK, J. *et al*. Artemis: sequence visualization and annotation. **Bioinformatics**, v. 16, n. 10, p. 944-945, 2000.

SALAMOV, A. A.; NISHIKAWA, T.; SWINDELLS, M. B. Assessing protein coding region integrity in cDNA sequencing projects. **Bioinformatics**, v. 14, p. 384-390, 1998.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proc Natl Acad Sci USA**, v. 74, n. 12, p. 5463-5467, 1977.

SAVOLAINEN, O.; LANGLEY, C. H.; LAZZARO, B. P. *et al*. Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. **Mol Biol Evol**, v. 17, p. 645-655, 2000.

SCANAVACA JUNIOR, L. **Caracterização silvicultural, botânica e tecnológica do *Eucalyptus urophylla* S.T. Blake e de seu potencial para utilização em serraria**. Piracicaba, 2001. 108 p. Dissertação (mestrado) - Escola superior de Agricultura Luiz de

Queirroz, 2002. Disponível em: <http://www.teses.usp.br/teses/disponiveis/11/11142/tde-22052002-154220/publico/laerte.pdf> Acesso em 10 maio 2005.

SCHNEIDER, S.; ROESSLI, D.; EXCOFFIER, L. **Arlequin, Version 2000**: a software for population genetics data analysis. Switzerland: Genetics and Biometry Laboratory, University of Geneva, 2000.

SEDEROFF, R. R.; MACKAY, J. J.; RALPH, J. *et al.* Unexpected variation in lignin. **Curr Opin Plant Biol**, v. 2, p. 145–152, 1999.

SIMKO, I.; HAYNES, K. G.; EWING, E. E. *et al*. Mapping genes for resistance to Verticillium albo-atrum in tetraploid and diploid potato populations using haplotype association tests and genetic analysis. **Mol Genet Genomics**, v. 271, p. 522-531, 2004.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **J Mol Biol**, v. 147, p. 195-197, 1981.

STEPHAN, W.; LANGLEY, C. H. DNA polymorphism in *Lycopersicon* and crossing-over per physical length. **Genetics**, v. 150, p. 1585-1593, 1998.

STEPHENS, M.; SMITH, N. J.; DONNELLY, P. A new statistical method for haplotype reconstruction from population data. **Am J Hum Genet**, v. 68, p. 978-989, 2001.

STEPHENS, M.; DONNELLY, P. A comparison of Bayesian method for haplotype reconstruction from population genotype data. **Am J Hum Genet**, v. 73, p. 1162-1169, 2003.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. **Genetics**, v. 123, n. 3, p. 585-595, 1989.

TENAILLON, M. I.; SAWKINS, M. C.; LONG, A. D. *et al*. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays L.*). **Proc Natl Acad Sci USA**, v. 98, n. 16, p. 9161-9166, 2001.

TARAZONA-SANTOS, E.; TISHKOFF, S. A. Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus. **Genes Immun**, v. 6, p. 53-65, 2005.

THORNSBERRY, J. M.; GOODMAN, M. M.; DOEBLEY, J. *et al. Dwarf8* polymorphisms associate with variation in flowering time. **Nat Genet**, v. 28, p. 286-289, 2001.

THRING, R. W.; KATIKANENI, S. P. R.; BAKHSHI, N. N. The production of gasoline range hydrocarbons from Alcell (R) lignin using HZSM-5 catalyst. **Fuel Processing Technology**, v. 62, p. 17-30, 2000.

VIGNOLS, F.; RIGAU, J.; TORRES, M. A. *et al.* The brown midrib 3 (bm3) mutation in maize occurs in the gene encoding caffeic acid O-methyltransferase. **Plant Cell**, v. 7, n. 4, p. 407-416, 1995.

VIIKARI, L.; KANTELINEN, A.; SUNDQUIST, J. *et al.* Xylanases in bleaching: from an idea to the industry. **FEMS Microbiol Rev**, v. 13, p. 335-350, 1994.

VINSON, J. P.; JAFFE, D. B.; O'NEILL, K. *et al*. Assembly of polymorphic genomes: Algorithms and application to Ciona savignyi. **Genome Res**, v. 15, n. 8, p. 1127-1135, 2005.

VOGEL, K. P.; JUNG, H. G. Genetic modification of herbaceous plants for feed and fuel. **CRC Crit Rev Plant Sci**, v. 20, p. 15-49, 2001.

WALTERS, B. **Australian plants online**. 1998. Disponível em: <http://farrer.riv.csu.edu.au/ASGAP/eucalypt.html>. Acesso em 10 maio 2005.

WANG, D. G.; FAN, J. B.; SIAO, C. J. *et al.* Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. **Science**, v. 280, p. 1077-1082, 1998.

WATTERSON, G. A. On the number of segregating sites in genetical models without recombination. **Theor Popul Biol**, v. 7, p. 256-276, 1975.

ZHONG, R.; MORRISON III, H.; NEGREL, J. *et al.* Dual methylation pathways in lignin biosynthesis. **Plant Cell**, v. 10, p. 2033-2045, 1998.

ZHU, Y. L.; SONG, Q. J.; HYTEN, D. L. *et al.* Single nucleotide polymorphisms in soybean. **Genetics,** v. 163, p. 1123-1134, 2003.

ZUBIETA, C.; KOTA, P.; FERRER, J-L. *et al.* Structural basis for the modulation of lignin monomer methylation by caffeic acid/5-hydroxyferulic acid 3/5-*O*-methyltransferase. **Plant Cell**, v. 14, n. 6, p. 1265–1277, 2002.

UNIPROTKB/SWISS-PROT. North Carolina, USA. Disponível em: <http://us.expasy.org/cgi-bin/niceprot.pl?Q9C9W3#ref>. Acesso em 9 dez. 2003.

9.Apêndices

APÊNDICE B

Alinhamento dos Contigs do Cluster 6, utilizando o software ClustalW.

```
CLUSTAL W (1.82) multiple sequence alignment


CL6Contig3      -CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGCA-----GGAAGAAGCCGAGCAAACG 54
CL6Contig6      -CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGAAA----GGAAGAAGCCGAGCAAGCG 55
CL6Contig2      ACGGTCCGGATTCCCGGGTCGACCCACGCGTCCGCA-----GGAAGAAGCCGAGCAAACG 55
CL6Contig4      -CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGAGACACAGGAAGAAGCCGAGCAAACG 59
CL6Contig1      -CGGTCCGGATTCCCGGGTCGAC----------ATTTCTTGATCACAACATATTA--CA 46
CL6Contig5      -CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGATTTCTTGATCACAACATATTA--CA 57
                 **********************                 *     * *  *  *  *


CL6Contig3      AAGTTGCAGACGCCATTGGAAAAAAGACACGAAAGAGATAAAGAAGGAGCTTAAGAAGCA 114
CL6Contig6      AAGTTGCAGACGCCATTG-AAAAAAGACACGAAAGAGATCAAGAAGGAGCTTAAGAAGCA 114
CL6Contig2      AAGTTGCAGACGCCATTGGAAAAAAGACACGAAAGAGATCAAGAAGGAGCTTAAGAAGCA 115
CL6Contig4      AAGTTGCAGACGCCATTGGAAAAAAGACACTAAAGAGATCAAGAAGGAGCTTAAGAAGCA 119
CL6Contig1      ATATTCCTAA--GCACAGAGAGAGAGA---GAGAGAGAGAGAGAGAGAGTTTGAA----- 96
CL6Contig5      ATATTCCTAA--GCACAGAGAGAGAGA---GAGAGAGAGAGAGAGAGAGAGAGA-GAG 111
                *  ** *  *   **  *  * * ***    * *****   ***  ***    *


CL6Contig3      TCA---ATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCATTCGGAAGTC 171
CL6Contig6      TCA---ATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCATTCGGAAGTC 171
CL6Contig2      TCATCAATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCATTCGGAAGTC 175
CL6Contig4      TCA---------TCAATGGCAGAGCCTCAGCAGATCCAACCAGCGAAGCATTCGGAAGTC 170
CL6Contig1      TCA---ATGGCCACCGCCGGAGAGGAGAGCCAGACCCAAGCCGGGAGGCACCAGGAGGTT 153
CL6Contig5      TTA---ATGGCCACCGCGGGAGAGGAGAGCCAGAACCAAGCCGGGAGGCACCAGGAGGTT 168
                * *          *    * ****      **** **** * * ** ***   *** **


CL6Contig3      GGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTC 231
CL6Contig6      GGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTC 231
CL6Contig2      GGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTC 235
CL6Contig4      GGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTC 230
CL6Contig1      GGCCACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAATATATTTTGGAGACCAGCGTG 213
CL6Contig5      GGCCACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAATATATTTTGGAGACCAGCGTG 228
                *********   *** * ***** ******** ***** ** ** **************


CL6Contig3      TACCCAAGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGG 291
CL6Contig6      TACCCAAGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGG 291
CL6Contig2      TACCCAAGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGG 295
CL6Contig4      TACCCAAGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGG 290
CL6Contig1      TACCCAAGAGAGCCTGAGCCCATGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGG 273
CL6Contig5      TACCCAAGAGAGCCTGAGCCCATGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGG 288
                ************** *** ***************************** ************


CL6Contig3      AACCTGATGACCACATCGGCGGATGAAGGGCAGTTCCTGAACATGCTCCTCAAGCTCATC 351
CL6Contig6      AACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGCTCCTCAAGCTCATC 351
CL6Contig2      AACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGCTCCTCAAGCTCATC 355
CL6Contig4      AACCTGATGACCACATCGGCGGATGAAGGGCAGTTCCTGAACATGCTCCTCAAGCTCATC 350
CL6Contig1      AACATAATGACAACATCAGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCATC 333
CL6Contig5      AACATAATGACAACATCAGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCATC 348
                *** * ***** ***** ** ** ************ ********** ************


CL6Contig3      AACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC 411
CL6Contig6      AACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC 411
CL6Contig2      AACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC 415
CL6Contig4      AACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC 410
CL6Contig1      AACGCCAAGAACACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCT 393
CL6Contig5      AACGCCAAGAACACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCT 408
                ********************** ** **** *** ************** ********


CL6Contig3      CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAG 471
CL6Contig6      CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAG 471
CL6Contig2      CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAG 475
CL6Contig4      CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAG 470
CL6Contig1      CTTGCTCTTCCTGATGACGGAAAGATTTTGGCTATGGACATTAACAGAGAGAACTATGAA 453
CL6Contig5      CTTGCTCTTCCTGATGACGGAAAGATTTTGGCTATGGACATTAACAGAGAGAACTATGAA 468
```

```
                ********** ************** ***** ******** ** ** *******  **

CL6Contig3      ATCGGGCTGCCCGTCATCCAGAAGGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGC 531
CL6Contig6      ATCGGGCTGCCCGTCATCCAGAAGGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGC 531
CL6Contig2      ATCGGGCTGCCCGTCATCCAGAAGGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGC 535
CL6Contig4      ATCGGGCTGCCCGTCATCCAGAAGGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGC 530
CL6Contig1      CTTGGCCTGCCGGTCATCCAAAAAGCCGGTGTTGCCGACAAGATTGACTTCAGAGAAGGC 513
CL6Contig5      CTTGGCCTGCCGGTCATCCAAAAAGCCGGTGTTGCCGACAAGATTGACTTCAGAGAAGGC 528
                 * ** ***** ******** ** *****  ***** ******* ** ************

CL6Contig3      CCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGAACGTACGAC 591
CL6Contig6      CCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGAACGTACGAC 591
CL6Contig2      CCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGAACGTACGAC 595
CL6Contig4      CCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGAACGTACGAC 590
CL6Contig1      CCTGCTTTGCCTATTCTTGATCAGTTGATCGAAGATGGGAAG---CAAGGGTCGTTCGAC 570
CL6Contig5      CCTGCTTTGCCTATTCTTGATCAGTTGATCGAAGATGGGAAG---CAAGGGTCGTTCGAC 585
                *****  ****  * ********* *  *  ****** ****   ** **  *** ****

CL6Contig3      TTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGAC 651
CL6Contig6      TTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGAC 651
CL6Contig2      TTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGAC 655
CL6Contig4      TTCATATTCGTGGATGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGAC 650
CL6Contig1      TTCATATTCGTGGACGCGGACAAGGACAATTACCTCAACTACCACAAGAGGCTGATCGAG 630
CL6Contig5      TTCATATTCGTGGACGCGGACAAGGACAATTACCTCAACTACCACAAGAGGCTGATCGAG 645
                ************** ** *********** *** *************************

CL6Contig3      CTGGTCAAGGTTGGCGGCCTGATCGGATACGACAACACCCTGTGGAACGGTTCCGTGGTC 711
CL6Contig6      CTGGTCAAGGTTGGCGGCCTGATCGGATACNACAACACCCTGTGGAACGGCTCCGTGGTC 711
CL6Contig2      CTGGTCAAGGTTGGCGGCCTGATCGGATACGACAACACCCTGTGGAACGGCTCCGTGGTC 715
CL6Contig4      CTGGTCAAGGTTGGCGGCCTGATCGGATACGACAACACCCTGTGGAACGGCTCCGTGGTC 710
CL6Contig1      CTTGTCAAGGTTGGAGGCCTCATTGGCTACGACAACACCCTATGGAACGGCTCCGTGGTT 690
CL6Contig5      CTTGTCAAGGTTGGAGGCCTCATTGGCTACGACAACACCCTATGGAACGGCTCCGTGGTT 705
                ** *********** ***** ** ** *** ********** ******** ********

CL6Contig3      GCGCCCGCCGACGCGCCCCTCCGCAAGTACGTCCGGTACTACCGGGACTTCGTGCTGGAG 771
CL6Contig6      GC------------------------------------------------------------ 713
CL6Contig2      GCGCCCGCCGACGCGCCCCTCCGCAAGTACGTGCGGTACTACCGGGACTTCGTGCTGGAG 775
CL6Contig4      GCGCCCGCCGACGCGCCCCTCCGCAAGTACGTCCGGTACTACCGGGACTTCGTGCTGGAG 770
CL6Contig1      GCACCGCCGGACGCCCCGCTCAGGAAGTATGTGAGGTACTACAGGGATTTTGTGCTGGAG 750
CL6Contig5      GCGCCGCCGGACGCCCCGCTCAGGAAGTATGTGAGGTACTACAGGGATTTTGTGCTGGAG 765
                **

CL6Contig3      CTCAACAAGGCCCTCGCCGTGGACCCGAGGATCGAGATCTGCATGCTTCCCGTCGGGGAT 831
CL6Contig6      ------------------------------------------------------------
CL6Contig2      CTCAACAAGGCCCTCGCCGTGGACCCGAGGATCGAGATCTGCATGCTTCCCGTCGGGGAT 835
CL6Contig4      CTCAACAAGGCCCTCGCCGTGGACCCGAGGATCGAGATCTGCATGCTTCCCGTCGGGGAT 830
CL6Contig1      CTCAACAAGGCTCTTGCCGCTGATCCTAGGATTGAGATCTGCATGCTCCCCGTGGGTGAT 810
CL6Contig5      CTCAACAAGGCTCTTGCCGCTGATCCTAGGATTGAGATCTGCATGCTCCCCGTGGGTGAT 825


CL6Contig3      GGCATCACTCTCTGCCGCCGGGTCAGCTGATCACACCGTAATGGAGATCTCAAAAACAGT 891
CL6Contig6      ------------------------------------------------------------
CL6Contig2      GGCATCACCCTGTGCCGCCGGGTCAGTTGATCGCACCAGAAAGGAGATCTCAAAATCAAG 895
CL6Contig4      GGCATCACTCTGTGCCGCCGGGTCAGCTGATCACACCGGAAAGGAGATCTCAAAAACAAG 890
CL6Contig1      GGCATCACTCTCTGCCGTCGGATCAGCTGAGCAT-CTAATCTCAAGTCCTTATGATCA-- 867
CL6Contig5      GGCATCACTCTCTGCCGTCGGATCAGCTGAGCAT-CTAATCTCAAGTCCTTATGATCAGG 884


CL6Contig3      GATCGATGAAATGAGAAACTACCTTTAATAC-TTTCCTTCCTTTCTATTTTTTCCATCTT 950
CL6Contig6      ------------------------------------------------------------
CL6Contig2      CATTGATGAAATGAGGAACTACCCTTGATAA-TTTCCTTCCTTTCTATTTTTTCCATCTT 954
CL6Contig4      GATCGATGAAATGGGAAACTACCCTTAATACCTTTCCTTCCTTTCTATTTTTTCCATCTT 950
CL6Contig1      ------------------------------------------------------------
CL6Contig5      GTTCATTCTTAATGTAGAACCCACGAAAAAGAGAGAGATTTATGTATATCTTGTTGCTGT 944


CL6Contig3      CTGTCTTATGTTGTCTTTGAACCATTGAGCATGTATTTGTATTCAAATGAACGATTAAGG1010
CL6Contig6      ------------------------------------------------------------
CL6Contig2      CTGTCTTATATTGTCTTTGAACCGTTGAGCATGTATTTGTATTCAAATGAACGATTAAGG1014
CL6Contig4      CTGTCTTATATTGTCTTTGAACCGTTGAGCATGTATTTGTATTCAAATGAATGATTAAGG1010
CL6Contig1      ------------------------------------------------------------
```

```
CL6Contig5      TTCTTTTCCATGAACCTAGAAACGGGATTCGCAATTAAATGCCAAATTATGTTGCTGTTT1004


CL6Contig3      ATTGAGAAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAAAA1070
CL6Contig6      ------------------------------------------------------------
CL6Contig2      AGTGAGAAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAAAA1074
CL6Contig4      AGTGAGAAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAA--1068
CL6Contig1      ------------------------------------------------------------
CL6Contig5      CTCTTTTA------------------------------------------------1012


CL6Contig3      TTTATGCTACGAA 1083
CL6Contig6      ------------- 
CL6Contig2      GTTCATTTATGCT 1087
CL6Contig4      -------------
CL6Contig1      -------------
CL6Contig5      -------------
```

APÊNDICE C

Predição da posição ATG inicial para o contig 1 e contig 5 do cluster 6.

Contig 1

No. do ATG a partir da extremidade 5'	Frame	Início (bp)	Fim (bp)	Tamanho ORF	Stop codon?	Seqüência
1	1	100	837	246	Yes	MATAGEESQTQAGRHQEVGHKSLLQSDALYQYILET SVYPREPEPMKELREITAKHPWNIMTTSADEGQFLN MLLKLINAKNTMEIGVFTGYSLLATALALPDDGKIL AMDINRENYELGLPVIQKAGVADKIDFREGPALPIL DQLIEDGKQGSFDFIFVDADKDNYLNYHKRLIELVK VGGLIGYDNTLWNGSVVAPPDAPLRKYVRYYRDFVL ELNKALAADPRIEICMLPVGDGITLCRRIS

Contig 5

No. do ATG a partir da extremidade 5'	Frame	Início (bp)	Fim (bp)	Tamanho ORF	Stop codon?	Seqüência
1	1	115	852	246	Yes	MATAGEESQNQAGRHQEVGHKSLLQSDALYQYILE TSVYPREPEPMKELREITAKHPWNIMTTSADEGQF LNMLLKLINAKNTMEIGVFTGYSLLATALALPDDG KILAMDINRENYELGLPVIQKAGVADKIDFREGPA LPILDQLIEDGKQGSFDFIFVDADKDNYLNYHKRL IELVKVGGLIGYDNTLWNGSVVAPPDAPLRKYVRY YRDFVLELNKALAADPRIEICMLPVGDGITLCRRI S

APÊNDICE D

Alinhamento das seqüências nucleotídicas dos Contig 1 e 5 do Cluster 6. Alinhamento realizado pelo software Align (EMBOSS) pelo método water.

```
#=======================================
#
# Aligned_sequences: 2
# 1: CL6Contig1
# 2: CL6Contig5
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 882
# Identity:     859/882 (97.4%)
# Similarity:   859/882 (97.4%)
# Gaps:          15/882 ( 1.7%)
# Score: 5112.5
#
#
#=======================================

CL6Contig1         1 CGGTCCGGATTCCCGGGTCGAC-----------ATTTCTTGATCACAACA     39
                     |||||||||||||||||||||           |||||||||||||||||
CL6Contig5         1 CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGATTTCTTGATCACAACA     50

CL6Contig1        40 TATTACAATATTCCTAAGCAC----AGAGAGAGAGAGAGAGAGAGAGAGA     85
                     ||||||||||||||||||||    ||||||||||||||||||||||||||
CL6Contig5        51 TATTACAATATTCCTAAGCACAGAGAGAGAGAGAGAGAGAGAGAGAGAGA    100

CL6Contig1        86 GAGAGTTTGAATCAATGGCCACCGCCGGAGAGGAGAGCCAGACCCAAGCC    135
                     |||||...||.|.|||||||||||.||||||||||||||||.|||||||
CL6Contig5       101 GAGAGAGAGAGTTAATGGCCACCGCGGGAGAGGAGAGCCAGAACCAAGCC    150

CL6Contig1       136 GGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTCT    185
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       151 GGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTCT    200

CL6Contig1       186 TTACCAATATATTTTGGAGACCAGCGTGTACCCAAGAGAGCCTGAGCCCA    235
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       201 TTACCAATATATTTTGGAGACCAGCGTGTACCCAAGAGAGCCTGAGCCCA    250

CL6Contig1       236 TGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGGAACATAATGACA    285
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       251 TGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGGAACATAATGACA    300

CL6Contig1       286 ACATCAGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCATCAA    335
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       301 ACATCAGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCATCAA    350

CL6Contig1       336 CGCCAAGAACACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTG    385
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       351 CGCCAAGAACACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTG    400

CL6Contig1       386 CCACCGCTCTTGCTCTTCCTGATGACGGAAAGATTTTGGCTATGGACATT    435
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       401 CCACCGCTCTTGCTCTTCCTGATGACGGAAAGATTTTGGCTATGGACATT    450

CL6Contig1       436 AACAGAGAGAACTATGAACTTGGCCTGCCGGTCATCCAAAAAGCCGGTGT    485
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       451 AACAGAGAGAACTATGAACTTGGCCTGCCGGTCATCCAAAAAGCCGGTGT    500

CL6Contig1       486 TGCCGACAAGATTGACTTCAGAGAAGGCCCTGCTTTGCCTATTCTTGATC    535
```

```
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       501 TGCCGACAAGATTGACTTCAGAGAAGGCCCTGCTTTGCCTATTCTTGATC    550

CL6Contig1       536 AGTTGATCGAAGATGGGAAGCAAGGGTCGTTCGACTTCATATTCGTGGAC    585
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       551 AGTTGATCGAAGATGGGAAGCAAGGGTCGTTCGACTTCATATTCGTGGAC    600

CL6Contig1       586 GCGGACAAGGACAATTACCTCAACTACCACAAGAGGCTGATCGAGCTTGT    635
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       601 GCGGACAAGGACAATTACCTCAACTACCACAAGAGGCTGATCGAGCTTGT    650

CL6Contig1       636 CAAGGTTGGAGGCCTCATTGGCTACGACAACACCCTATGGAACGGCTCCG    685
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       651 CAAGGTTGGAGGCCTCATTGGCTACGACAACACCCTATGGAACGGCTCCG    700

CL6Contig1       686 TGGTTGCACCGCCGGACGCCCCGCTCAGGAAGTATGTGAGGTACTACAGG    735
                    |||||||.||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       701 TGGTTGCGCCGCCGGACGCCCCGCTCAGGAAGTATGTGAGGTACTACAGG    750

CL6Contig1       736 GATTTTGTGCTGGAGCTCAACAAGGCTCTTGCCGCTGATCCTAGGATTGA    785
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       751 GATTTTGTGCTGGAGCTCAACAAGGCTCTTGCCGCTGATCCTAGGATTGA    800

CL6Contig1       786 GATCTGCATGCTCCCCGTGGGTGATGGCATCACTCTCTGCCGTCGGATCA    835
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig5       801 GATCTGCATGCTCCCCGTGGGTGATGGCATCACTCTCTGCCGTCGGATCA    850

CL6Contig1       836 GCTGAGCATCTAATCTCAAGTCCTTATGATCA    867
                    ||||||||||||||||||||||||||||||||
CL6Contig5       851 GCTGAGCATCTAATCTCAAGTCCTTATGATCA    882
```

APÊNDICE E

Alinhamento das seqüências nucleotídicas dos contigs 2, 3, 4 e 6 do Cluster 6. Alinhamento realizado pelo software ClustalW.

```
CLUSTAL W (1.82) multiple sequence alignment


CL6Contig2      ACGGTCCGGATTCCCGGGTCGACCCACGCGTCCGCA-----GGAAGAAGCCGAGCAAACG 55
CL6Contig6      -CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGAAA----GGAAGAAGCCGAGCAAGCG 55
CL6Contig3      -CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGCA-----GGAAGAAGCCGAGCAAACG 54
CL6Contig4      -CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGAGACACAGGAAGAAGCCGAGCAAACG 59
                 ********************************       **************** **

CL6Contig2      AAGTTGCAGACGCCATTGGAAAAAAGACACGAAAGAGATCAAGAAGGAGCTTAAGAAGCA 115
CL6Contig6      AAGTTGCAGACGCCATTG-AAAAAAGACACGAAAGAGATCAAGAAGGAGCTTAAGAAGCA 114
CL6Contig3      AAGTTGCAGACGCCATTGGAAAAAAGACACGAAAGAGATAAAGAAGGAGCTTAAGAAGCA 114
CL6Contig4      AAGTTGCAGACGCCATTGGAAAAAAGACACTAAAGAGATCAAGAAGGAGCTTAAGAAGCA 119
                ****************** ********** ******** ********************

CL6Contig2      TCATCAATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCATTCGGAAGTC 175
CL6Contig6      TCA---ATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCATTCGGAAGTC 171
CL6Contig3      TCA---ATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCATTCGGAAGTC 171
CL6Contig4      TCA---------TCAATGGCAGAGCCTCAGCAGATCCAACCAGCGAAGCATTCGGAAGTC 170
                ***          * *  ********************** ************************

CL6Contig2      GGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTC 235
CL6Contig6      GGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTC 231
CL6Contig3      GGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTC 231
CL6Contig4      GGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTC 230
                ************************************************************

CL6Contig2      TACCCAAGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGG 295
CL6Contig6      TACCCAAGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGG 291
CL6Contig3      TACCCAAGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGG 291
CL6Contig4      TACCCAAGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGG 290
                ********************************************** ************

CL6Contig2      AACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGCTCCTCAAGCTCATC 355
CL6Contig6      AACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGCTCCTCAAGCTCATC 351
CL6Contig3      AACCTGATGACCACATCGGCGGATGAAGGGCAGTTCCTGAACATGCTCCTCAAGCTCATC 351
CL6Contig4      AACCTGATGACCACATCGGCGGATGAAGGGCAGTTCCTGAACATGCTCCTCAAGCTCATC 350
                ******************** ***************************************

CL6Contig2      AACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC 415
CL6Contig6      AACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC 411
CL6Contig3      AACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC 411
CL6Contig4      AACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC 410
                ************************************************************

CL6Contig2      CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAG 475
CL6Contig6      CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAG 471
CL6Contig3      CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAG 471
CL6Contig4      CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAG 470
                ************************************************************

CL6Contig2      ATCGGGCTGCCCGTCATCCAGAAGGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGC 535
CL6Contig6      ATCGGGCTGCCCGTCATCCAGAAGGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGC 531
CL6Contig3      ATCGGGCTGCCCGTCATCCAGAAGGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGC 531
CL6Contig4      ATCGGGCTGCCCGTCATCCAGAAGGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGC 530
                ************************************************************

CL6Contig2      CCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGAACGTACGAC 595
CL6Contig6      CCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGAACGTACGAC 591
```

```
CL6Contig3      CCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGAACGTACGAC 591
CL6Contig4      CCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGAACGTACGAC 590
                ************************************************************

CL6Contig2      TTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGAC 655
CL6Contig6      TTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGAC 651
CL6Contig3      TTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGAC 651
CL6Contig4      TTCATATTCGTGGATGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGAC 650
                ************** ********************************************

CL6Contig2      CTGGTCAAGGTTGGCGGCCTGATCGGATACGACAACACCCTGTGGAACGGCTCCGTGGTC 715
CL6Contig6      CTGGTCAAGGTTGGCGGCCTGATCGGATACNACAACACCCTGTGGAACGGCTCCGTGGTC 711
CL6Contig3      CTGGTCAAGGTTGGCGGCCTGATCGGATACGACAACACCCTGTGGAACGGTTCCGTGGTC 711
CL6Contig4      CTGGTCAAGGTTGGCGGCCTGATCGGATACGACAACACCCTGTGGAACGGCTCCGTGGTC 710
                ****************************** ****************** *********

CL6Contig2      GCGCCCGCCGACGCGCCCCTCCGCAAGTACGTGCGGTACTACCGGGACTTCGTGCTGGAG 775
CL6Contig6      GC------------------------------------------------------------ 713
CL6Contig3      GCGCCCGCCGACGCGCCCCTCCGCAAGTACGTCCGGTACTACCGGGACTTCGTGCTGGAG 771
CL6Contig4      GCGCCCGCCGACGCGCCCCTCCGCAAGTACGTCCGGTACTACCGGGACTTCGTGCTGGAG 770
                **
```

APÊNDICE F

Alinhamento da seqüência do contig 1 do cluster 6 com o singleton EUGR-BC pelo programa Align por meio do método needle.

```
#=======================================
#
# Aligned_sequences: 2
# 1: CN-EUGR-BC-002-003_F02_.f_015.ab1_Reverso
# 2: CL6Contig1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1163
# Identity:     437/1163 (37.6%)
# Similarity:   437/1163 (37.6%)
# Gaps:         700/1163 (60.2%)
# Score: 2330.0
#
#
#=======================================

CN-EUGR-BC-00      1 CGGACCATCACTGTCCTTATATACGTTGCATCATGCTTGCTCATAGAACT     50
                       |.||||..||...||
CL6Contig1         1           CGGTCCGGATTCCCG-------------------------     15

CN-EUGR-BC-00     51 TAGGTCAACTGCAACATTTCTTGATCACAACATATTACAATATTCCTAAG    100
                       |||       |.||||||||||||||||||||||||||||||||||
CL6Contig1        16 --GGT------CGACATTTCTTGATCACAACATATTACAATATTCCTAAG     57

CN-EUGR-BC-00    101 CAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGTTTGAATCAATGGCCAC    150
                     ||.||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig1        58 CACAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGTTTGAATCAATGGCCAC    107

CN-EUGR-BC-00    151 CGCCGGAGAGGAGAGCCAGACCCAAGCCGGGAGGCACCAGGAGGTTGGCC    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig1       108 CGCCGGAGAGGAGAGCCAGACCCAAGCCGGGAGGCACCAGGAGGTTGGCC    157

CN-EUGR-BC-00    201 ACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAAGTGAGTGTGAATCTT    250
                     |||||||||||||||||||||||||||||||||||
CL6Contig1       158 ACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAA---------------    192

CN-EUGR-BC-00    251 TATAGCTTTTGTTGGAATCGATGGATGATTCTGTTCCTTCTGCTGTATAA    300

CL6Contig1       193 --------------------------------------------------    192

CN-EUGR-BC-00    301 CTTGATCGATAGTTTCCAACTTGAGGTGTTGTGTCTCTGATCTTCAATCG    350

CL6Contig1       193 --------------------------------------------------    192

CN-EUGR-BC-00    351 TTTCTCTTTTTGCAGTATATTTTGGAGACCAGCGTGTACCCAAGAGAGCC    400
                             |||||||||||||||||||||||||||||||||||||
CL6Contig1       193 ---------------TATATTTTGGAGACCAGCGTGTACCCAAGAGAGCC    227

CN-EUGR-BC-00    401 TGAGCCCATGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGGTGAG    450
                     |||||||||||||||||||||||||||||||||||||||||||||
CL6Contig1       228 TGAGCCCATGAAGGAGCTCAGGGAAATAACAGCAAAACATCCAT------    271

CN-EUGR-BC-00    451 TTCGCATATGATTCAGGAACAGAACAGATCAGACTTCAATAAATGCCCTC    500

CL6Contig1       272 --------------------------------------------------    271

CN-EUGR-BC-00    501 TCTCATTCACGAGTTTGTTTCTTGAGCTGAACCCTTTTTGTTCCTTTGAA    550

CL6Contig1       272 --------------------------------------------------    271
```

```
CN-EUGR-BC-00    551 TATAAATTGCAAAAATAGGAACATAATGACAACATCAGCAGACGAAGGGC    600
                     ||||||||||||||||||||||||||||||||||
CL6Contig1       272 -----------------GGAACATAATGACAACATCAGCAGACGAAGGGC    304

CN-EUGR-BC-00    601 AGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCAAGAACACCATGGAG    650
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
CL6Contig1       305 AGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCAAGAACACCATGGAG    354

CN-EUGR-BC-00    651 ATTGGTGTCTTCACTGGCTACTCTCTCCTCGCCACCGCTCTTGCTCTTCC    700
                     ||||||||||||||||||||||||||||.||||||||||||||||||||
CL6Contig1       355 ATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTTGCTCTTCC    404

CN-EUGR-BC-00    701 TGATGACGGAAAGGT------------------CGGTTTGA--TTTGTTC    730
                     |||||||||||||.|                  |.|...||  |.||..|
CL6Contig1       405 TGATGACGGAAAGATTTTGGCTATGGACATTAACAGAGAGAACTATGAAC    454

CN-EUGR-BC-00    731 ------------TCTTCCAA------------GTCAACGAAGTT-TCTTC    755
                             ||.|||||            |.|.||.|..|| .||||
CL6Contig1       455 TTGGCCTGCCGGTCATCCAAAAAGCCGGTGTTGCCGACAAGATTGACTTC    504

CN-EUGR-BC-00    756 AAT-----C                                          759
                     |..     |
CL6Contig1       505 AGAGAAGGCCCTGCTTTGCCTATTCTTGATCAGTTGATCGAAGATGGGAA    554

CN-EUGR-BC-00    760                                                  759

CL6Contig1       555 GCAAGGGTCGTTCGACTTCATATTCGTGGACGCGGACAAGGACAATTACC    604

CN-EUGR-BC-00    760                                                  759

CL6Contig1       605 TCAACTACCACAAGAGGCTGATCGAGCTTGTCAAGGTTGGAGGCCTCATT    654

CN-EUGR-BC-00    760                                                  759

CL6Contig1       655 GGCTACGACAACACCCTATGGAACGGCTCCGTGGTTGCACCGCCGGACGC    704

CN-EUGR-BC-00    760                                                  759

CL6Contig1       705 CCCGCTCAGGAAGTATGTGAGGTACTACAGGGATTTTGTGCTGGAGCTCA    754

CN-EUGR-BC-00    760                                                  759

CL6Contig1       755 ACAAGGCTCTTGCCGCTGATCCTAGGATTGAGATCTGCATGCTCCCCGTG    804

CN-EUGR-BC-00    760                                                  759

CL6Contig1       805 GGTGATGGCATCACTCTCTGCCGTCGGATCAGCTGAGCATCTAATCTCAA    854

CN-EUGR-BC-00    760                 759

CL6Contig1       855 GTCCTTATGATCA    867
```

APÊNDICE G

Alinhamento dos singletons EUSP-FX e EUGL-XY pelo programa ClustalW.

```
CLUSTAL W (1.82) multiple sequence alignment


CN-EUSP-FX-002-018_A08_.g_053.      CGGTCCGGATTCCCGGGTCGACCACGCGTCCGCCCACGCGTCCGCTTTTT 50
CN_EUGL_XY_002_018_C07_.g_050       --------------------------------------------------


CN-EUSP-FX-002-018_A08_.g_053.      CTCTGGAAATGAAAGGGTATTAAACAAGTCCAAGATCAAGAGGCCTGCAT 100
CN_EUGL_XY_002_018_C07_.g_050       --------------------------------------------------


CN-EUSP-FX-002-018_A08_.g_053.      GCTTCCAGGAGCTCACCAACCCTGGACCAGAGAGAGAGAGAGAGAGTTTG 150
CN_EUGL_XY_002_018_C07_.g_050       --------------------------------------------------


CN-EUSP-FX-002-018_A08_.g_053.      AATCAATGGCCACCGCCGGAGAGGAGAGCCAGAACCAAGCCGGGAGGCAC 200
CN_EUGL_XY_002_018_C07_.g_050       ---CCCGGGTCGACCCACGCGTCCGGNACCAGAACCAAGCCGGGAGGCAC 47
                                       *   ** *  * *  * *    *  ********************


CN-EUSP-FX-002-018_A08_.g_053.      CAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAATA 250
CN_EUGL_XY_002_018_C07_.g_050       CAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAATA 97
                                    **************************************************


CN-EUSP-FX-002-018_A08_.g_053.      TATTTTGGAGACTAGCGTGTACCCAAGAGAGCCCGAGCCCATGAAGGAGC 300
CN_EUGL_XY_002_018_C07_.g_050       TATTTTGGAGACCAGCGTGTACCCAAGAGAGCCTGAGCCCATGAAGGAGC 147
                                    ************ ******************* ****************


CN-EUSP-FX-002-018_A08_.g_053.      TCAGGGAAATTACAGCAAAACATCCATGGAACATAATGACAACATCAGCA 350
CN_EUGL_XY_002_018_C07_.g_050       TCAGGGAAATAACAGCAAAACATCCATGGAACATAATGACAACATCAGCA 197
                                    ********** ****************************************


CN-EUSP-FX-002-018_A08_.g_053.      GACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCAAGAA 400
CN_EUGL_XY_002_018_C07_.g_050       GACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTC---AACGCCAAGAA 244
                                    ************************************   ***********


CN-EUSP-FX-002-018_A08_.g_053.      CACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTC 450
CN_EUGL_XY_002_018_C07_.g_050       CACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTC 294
                                    **************************************************


CN-EUSP-FX-002-018_A08_.g_053.      TTGCTCTTCCTGATGACGGAAAGATTTTGGCTATGGACATTAAC---AGA 497
CN_EUGL_XY_002_018_C07_.g_050       TTGCTCTTCCTGATGACGGAAAGATTTTGGCTATAGACGATAACTCAAGA 344
                                    ********************************** ***  ****   ***


CN-EUSP-FX-002-018_A08_.g_053.      GAGAACTATGAACTTGGCCTGCCGGTCATCCAAAA--AGCCG-GTGTTGC 544
CN_EUGL_XY_002_018_C07_.g_050       ATTGAGTATGGCAATCGAATGGCGCTCATTTCAGACCAGCTTTACGAGGC 394
                                        * ****    * *  ** ** ****   * *  ***     *  **


CN-EUSP-FX-002-018_A08_.g_053.      CCACAAGATTGACTTCAGAGAAGGCCCTG----------CTTTGCCTATT 584
CN_EUGL_XY_002_018_C07_.g_050       GGCCAAAACAAACTGCGATGGCGACTATGTGAATGTTGGCGATGATAATC 444
                                       *** *   *** *   *  * *  **         *  **   **


CN-EUSP-FX-002-018_A08_.g_053.      CTTGAT---CAGTTGATC-GAAGATGGGAAGCAAG---------GGACGT 621
CN_EUGL_XY_002_018_C07_.g_050       TTCAATGCGTCGTTGATCTGAATGCCATCAACGAGCTCACCTCCAAAGTC 494
                                     *  **     ******* ***       * * **          *


CN-EUSP-FX-002-018_A08_.g_053.      TCGAC--TTCATATTCGTGGACGCGGACAAGGACAATTACCTCAACTA-C 668
CN_EUGL_XY_002_018_C07_.g_050       CTGACGGTCCACATTC-TGGAGCCGGATTGCAGCAACGCCCTGCCCGATC 543
                                      ***  * ** **** ****  ****      ***   ***   * * *


CN-EUSP-FX-002-018_A08_.g_053.      CACAAGAGGCTGATCGAGCTTGTCAAGGTT----GGAGGCCTCATTGGCT 714
CN_EUGL_XY_002_018_C07_.g_050       CACAAGATGACGCTTCCGTTCGAAGATATCTCCGGGAAAACGCCACGGCC 593
                                    ******* *  * *   * * *   *  *     ***   * *   ***


CN-EUSP-FX-002-018_A08_.g_053.      ACGACAAC-----ACCCTATGGAACGGCTC-------------------- 739
CN_EUGL_XY_002_018_C07_.g_050       GCATCCACGGACGACCCCCTCAAAGGTCCTGCGCTCTGGTGCCGTAACTA 643
                                     *  * **     ****  *  ** * *


CN-EUSP-FX-002-018_A08_.g_053.      --------------------------------------------------
CN_EUGL_XY_002_018_C07_.g_050       CAACCACATGGTCTGCTATATCTGGGCGAACCTTAAAAGCGTCCAAGAAG 693
```

```
CN-EUSP-FX-002-018_A08_.g_053.      ----------------------------------------
CN_EUGL_XY_002_018_C07_.g_050       CACTTCAAGTCCGACCGGGGACGAAGACCTACTGGCCATAC 734
```

APÊNDICE H

Determinação das regiões de exons e introns para o singleton EUGL-XY e EUSP-FX, por comparação com a seqüência de referência obtida no trabalho através do programa *EST2Genome*.

```
Note Best alignment is between forward est and forward genome, and splice  sites
imply forward gene
Exon        58 100.0     1    58 SeqRef_Reverso    38    95 CN_EUGL_XY
+Intron    -20   0.0    59   188 SeqRef_Reverso
Exon        78  98.8   189   268 SeqRef_Reverso    96   175 CN_EUGL_XY
+Intron    -20   0.0   269   391 SeqRef_Reverso
Exon       115  97.6   392   515 SeqRef_Reverso   176   296 CN_EUGL_XY

Span       211  98.5     1   515 SeqRef_Reverso    38   296 CN_EUGL_XY

Segment     58 100.0     1    58 SeqRef_Reverso    38    95 CN_EUGL_XY
Segment     78  98.8   189   268 SeqRef_Reverso    96   175 CN_EUGL_XY
Segment     55 100.0   392   446 SeqRef_Reverso   176   230 CN_EUGL_XY
Segment      1 100.0   449   449 SeqRef_Reverso   231   231 CN_EUGL_XY
Segment     65 100.0   451   515 SeqRef_Reverso   232   296 CN_EUGL_XY


SeqRef_Reverso vs CN_EUGL_XY:

SeqRef_Reverso      1 CGGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTC     50
                      |||||||||||||||||||||||||||||||||||||||||||||||||
    CN_EUGL_XY     38 CGGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTC     87

SeqRef_Reverso     51 TTTACCAAgtgag.....tgcagTATATTTTGGAGACCAGCGTGTACCCA    215
                      ||||||||>>>>> 130 >>>>>|||||||||||||||||||||||||||||
    CN_EUGL_XY     88 TTTACCAA..............TATATTTTGGAGACCAGCGTGTACCCA    122

SeqRef_Reverso    216 AGAGAGCCTGAGCCCATGAAGGAGCTCAGGGACATAACAGCAAAACATCC    265
                      |||||||||||||||||||||||||||||||| ||||||||||||||||
    CN_EUGL_XY    123 AGAGAGCCTGAGCCCATGAAGGAGCTCAGGGAAATAACAGCAAAACATCC    172

SeqRef_Reverso    266 ATGgtgag.....aatagGAACATAATGACAACATCAGCAGACGAAGGGC    423
                      |||>>>>> 123 >>>>>|||||||||||||||||||||||||||||||||

    CN_EUGL_XY    173 ATG..............GAACATAATGACAACATCAGCAGACGAAGGGC    207

SeqRef_Reverso    424 AGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCAAGAACACCATGGAG    473
                      |||||||||||||||||||||||  | |||||||||||||||||||||||
    CN_EUGL_XY    208 AGTTCTTGAACATGCTTCTCAAG--C-TCAACGCCAAGAACACCATGGAG    254

SeqRef_Reverso    474 ATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTT    515
                      |||||||||||||||||||||||||||||||||||||||||||
    CN_EUGL_XY    255 ATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTT    296


Alignment Score: 211
```

```
Note Best alignment is between forward est and forward genome, and splice  sites
imply forward gene
Exon        58 100.0     1    58 SeqRef_Reverso   191   248 CN-EUSP-FX
+Intron    -20   0.0    59   188 SeqRef_Reverso
Exon        72  95.0   189   268 SeqRef_Reverso   249   328 CN-EUSP-FX
+Intron    -20   0.0   269   391 SeqRef_Reverso
Exon       124 100.0   392   515 SeqRef_Reverso   329   452 CN-EUSP-FX

Span       214  98.5     1   515 SeqRef_Reverso   191   452 CN-EUSP-FX

Segment     58 100.0     1    58 SeqRef_Reverso   191   248 CN-EUSP-FX
Segment     72  95.0   189   268 SeqRef_Reverso   249   328 CN-EUSP-FX
Segment    124 100.0   392   515 SeqRef_Reverso   329   452 CN-EUSP-FX


SeqRef_Reverso vs CN-EUSP-FX:

SeqRef_Reverso      1 CGGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTC     50
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
    CN-EUSP-FX    191 CGGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTC    240

SeqRef_Reverso     51 TTTACCAAgtgag.....tgcagTATATTTTGGAGACCAGCGTGTACCCA    215
                      ||||||||>>>>> 130 >>>>>|||||||||||||| ||||||||||||
    CN-EUSP-FX    241 TTTACCAA..............TATATTTTGGAGACTAGCGTGTACCCA    275

SeqRef_Reverso    216 AGAGAGCCTGAGCCCATGAAGGAGCTCAGGGACATAACAGCAAAACATCC    265
                      |||||||| ||||||||||||||||||||||| || ||||||||||||||
    CN-EUSP-FX    276 AGAGAGCCCGAGCCCATGAAGGAGCTCAGGGAAATTACAGCAAAACATCC    325

SeqRef_Reverso    266 ATGgtgag.....aatagGAACATAATGACAACATCAGCAGACGAAGGGC    423
                      |||>>>>> 123 >>>>>|||||||||||||||||||||||||||||||||
    CN-EUSP-FX    326 ATG..............GAACATAATGACAACATCAGCAGACGAAGGGC    360

SeqRef_Reverso    424 AGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCAAGAACACCATGGAG    473
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
    CN-EUSP-FX    361 AGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCAAGAACACCATGGAG    410

SeqRef_Reverso    474 ATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTT    515
                      |||||||||||||||||||||||||||||||||||||||||||
    CN-EUSP-FX    411 ATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTT    452


Alignment Score: 214
```

APÊNDICE I

Seqüências consenso dos contigs e singletons obtidos do banco de dados para o gene CCoOAMT.

>CL6Contig1 consenso
CGGTCCGGATTCCCGGGTCGACATTTCTTGATCACAACATATTACAATATTCCTAAGCACAGAGAGAGAGAGAGAGAGAGAGA
GAGAGAGTTTGAATCAATGGCCACCGCCGGAGAGGAGAGCCAGACCCAAGCCGGGAGGCACCAGGAGGTTGGCCACAAGTCTC
TCCTTCAGAGTGATGCTCTTTACCAATATATTTTGGAGACCAGCGTGTACCCAAGAGAGCCTGAGCCCATGAAGGAGCTCAGG
GAAATAACAGCAAAACATCCATGGAACATAATGACAACATCAGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCAT
CAACGCCAAGAACACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTTGCTCTTCCTGATGACGGAA
AGATTTTGGCTATGGACATTAACAGAGAGAACTATGAACTTGGCCTGCCGGTCATCCAAAAAGCCGGTGTTGCCGACAAGATT
GACTTCAGAGAAGGCCCTGCTTTGCCTATTCTTGATCAGTTGATCGAAGATGGGAAGCAAGGGTCGTTCGACTTCATATTCGT
GGACGCGGACAAGGACAATTACCTCAACTACCACAAGAGGCTGATCGAGCTTGTCAAGGTTGGAGGCCTCATTGGCTACGACA
ACACCCTATGGAACGGCTCCGTGGTTGCACCGCCGGACGCCCCGCTCAGGAAGTATGTGAGGTACTACAGGGATTTTGTGCTG
GAGCTCAACAAGGCTCTTGCCGCTGATCCTAGGATTGAGATCTGCATGCTCCCCGTGGGTGATGGCATCACTCTCTGCCGTCG
GATCAGCTGAGCATCTAATCTCAAGTCCTTATGATCA

>CL6Contig2 consenso
ACGGTCCGGATTCCCGGGTCGACCCACGCGTCCGCAGGAAGAAGCCGAGCAAACGAAGTTGCAGACGCCATTGGAAAAAAGAC
ACGAAAGAGATCAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCAT
TCGGAAGTCGGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCAAGAGAGCC
AGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCC
TGAACATGCTCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCC
CTTGCTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCATCCAGAA
GGCCGGCCTTGCCCACAAGATCGATTTCAGAGAAGGCCCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACC
ATGGAACGTACGACTTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGACCTGGTCAAG
GTTGGCGGCCTGATCGGATACGACAACACCCTGTGGAACGGCTCCGTGGTCGCGCCCGCCGACGCGCCCCTCCGCAAGTACGT
GCGGTACTACCGGGACTTCGTGCTGGAGCTCAACAAGGCCCTCGCCGTGGACCCGAGGATCGAGATCTGCATGCTTCCCGTCG
GGGATGGCATCACCCTGTGCCGCCGGGTCAGTTGATCGCACCAGAAAGGAGATCTCAAAATCAAGCATTGATGAAATGAGGAA
CTACCCTTGATAATTTCCTTCCTTTCTATTTTTTCCATCTTCTGTCTTATATTGTCTTTGAACCGTTGAGCATGTATTTGTAT
TCAAATGAACGATTAAGGAGTGAGAAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAAAAGTTCA
TTTATGCT

>CL6Contig3 consenso
CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGCAGGAAGAAGCCGAGCAAACGAAGTTGCAGACGCCATTGGAAAAAAGACA
CGAAAGAGATAAAGAAGGAGCTTAAGAAGCATCAATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCATTCGG
AAGTCGGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCAAGAGAGCCAGAG
TCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGGAACCTGATGACCACATCGGCGGATGAAGGGCAGTTCCTGAA
CATGCTCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCCCTTG
CTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCATCCAGAAGGCC
GGCCTTGCCCACAAGATCGATTTCAGAGAAGGCCCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGG
AACGTACGACTTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGACCTGGTCAAGGTTG
GCGGCCTGATCGGATACGACAACACCCTGTGGAACGGTTCCGTGGTCGCGCCCGCCGACGCGCCCCTCCGCAAGTACGTCCGG
TACTACCGGGACTTCGTGCTGGAGCTCAACAAGGCCCTCGCCGTGGACCCGAGGATCGAGATCTGCATGCTTCCCGTCGGGGA
TGGCATCACTCTCTGCCGCCGGGTCAGCTGATCACACCGTAATGGAGATCTCAAAAACAGTGATCGATGAAATGAGAAACTAC
CTTTAATACTTTCCTTCCTTTCTATTTTTTCCATCTTCTGTCTTATGTTGTCTTTGAACCATTGAGCATGTATTTGTATTCAA
ATGAACGATTAAGGATTGAGAAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAAAATTTATGCTA
CGAA

>CL6Contig4 consenso
CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGAGACACAGGAAGAAGCCGAGCAAACGAAGTTGCAGACGCCATTGGAAAAA
AGACACTAAAGAGATCAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGAGCCTCAGCAGATCCAACCAGCGAAGCATTCGGA
AGTCGGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCAAGAGAGCCAGAGT
CCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCCATGGAACCTGATGACCACATCGGCGGATGAAGGGCAGTTCCTGAAC
ATGCTCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCCCTTGC
TCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCATCCAGAAGGCCG
GCCTTGCCCACAAGATCGATTTCAGAGAAGGCCCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGGA
ACGTACGACTTCATATTCGTGGATGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGACCTGGTCAAGGTTGG
CGGCCTGATCGGATACGACAACACCCTGTGGAACGGCTCCGTGGTCGCGCCCGCCGACGCGCCCCTCCGCAAGTACGTCCGGT
ACTACCGGGACTTCGTGCTGGAGCTCAACAAGGCCCTCGCCGTGGACCCGAGGATCGAGATCTGCATGCTTCCCGTCGGGGAT
GGCATCACTCTGTGCCGCCGGGTCAGCTGATCACACCGGAAAGGAGATCTCAAAAACAAGGATCGATGAAATGGGAAACTACC
CTTAATACCTTTCCTTCCTTTCTATTTTTTCCATCTTCTGTCTTATATTGTCTTTGAACCGTTGAGCATGTATTTGTATTCAA
ATGAATGATTAAGGAGTGAGAAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAA

>CL6Contig5 consenso
CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGATTTCTTGATCACAACATATTACAATATTCCTAAGCACAGAGAGAGAGAG
AGAGAGAGAGAGAGAGAGAGAGAGAGTTAATGGCCACCGCGGGAGAGGAGAGCCAGAACCAAGCCGGGAGGCACCAGGAGG
TTGGCCACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAATATATTTTGGAGACCAGCGTGTACCCAAGAGAGCCTGAGCCC
ATGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGGAACATAATGACAACATCAGCAGACGAAGGGCAGTTCTTGAACAT
GCTTCTCAAGCTCATCAACGCCAAGAACACCATGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTTGCTC
TTCCTGATGACGGAAAGATTTTGGCTATGGACATTAACAGAGAGAACTATGAACTTGGCCTGCCGGTCATCCAAAAAGCCGGT
GTTGCCGACAAGATTGACTTCAGAGAAGGCCCTGCTTTGCCTATTCTTGATCAGTTGATCGAAGATGGGAAGCAAGGGTCGTT
CGACTTCATATTCGTGGACGCGGACAAGGACAATTACCTCAACTACCACAAGAGGCTGATCGAGCTTGTCAAGGTTGGAGGCC
TCATTGGCTACGACAACACCCTATGGAACGGCTCCGTGGTTGCGCCGCCGGACGCCCCGCTCAGGAAGTATGTGAGGTACTAC
AGGGATTTTGTGCTGGAGCTCAACAAGGCTCTTGCCGCTGATCCTAGGATTGAGATCTGCATGCTCCCCGTGGGTGATGGCAT
CACTCTCTGCCGTCGGATCAGCTGAGCATCTAATCTCAAGTCCTTATGATCAGGGTTCATTCTTAATGTAGAACCCACGAAAA
AGAGAGAGATTTATGTATATCTTGTTGCTGTTTCTTTTCCATGAACCTAGAAACGGGATTCGCAATTAAATGCCAAATTATGT
TGCTGTTTCTCTTTTA

>CL6Contig6 consenso
CGGTCCGGATTCCCGGGTCGACCCACGCGTCCGAAAGGAAGAAGCCGAGCAAGCGAAGTTGCAGACGCCATTGAAAAAAGACA
CGAAAGAGATCAAGAAGGAGCTTAAGAAGCATCAATGGCAGCCAACGCAGAGCCTCAGCAGACCCAACCAGCGAAGCATTCGG
AAGTCGGCCACAAGAGCCTCTTGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCAAGAGAGCCAGAG
TCCATGAAGGAGCTCAGGGAAATAACAGCAAAACATCCATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAA
CATGCTCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACCGGCTACTCTCTCCTCGCCACCGCCCTTG
CTCTTCCCGATGACGGAAAGATCTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCATCCAGAAGGCC
GGCCTTGCCCACAAGATCGATTTCAGAGAAGGCCCTGCCCTGCCGCTCCTTGATCAGCTCGTGCAAGATGAGAAGAACCATGG
AACGTACGACTTCATATTCGTGGACGCCGACAAGGACAACTACATCAACTACCACAAGAGGCTGATCGACCTGGTCAAGGTTG
GCGGCCTGATCGGATACNACAACACCCTGTGGAACGGCTCCGTGGTCGC

>EUGR-BC consenso
GATTGAAGAAACTTCGTTGACTTGGAAGAGAACAAATCAAACCGACCTTTCCGTCATCAGGAAGAGCAAGAGCGGTGGCGAGG
AGAGAGTAGCCAGTGAAGACACCAATCTCCATGGTGTTCTTGGCGTTGATGAGCTTGAGAAGCATGTTCAAGAACTGCCCTTC
GTCTGCTGATGTTGTCATTATGTTCCTATTTTTGCAATTTATATTCAAAGGAACAAAAAGGGTTCAGCTCAAGAAACAAACTC
GTGAATGAGAGAGGGCATTTATTGAAGTCTGATCTGTTCTGTTCCTGAATCATATGCGAACTCACCATGGATGTTTTGCTGTT
ATTTCCCTGAGCTCCTTCATGGGCTCAGGCTCTCTTGGGTACACGCTGGTCTCCAAAATATACTGCAAAAAGAGAAACGATTG
AAGATCAGAGACACAACACCTCAAGTTGGAAACTATCGATCAAGTTATACAGCAGAAGGAACAGAATCATCCATCGATTCCAA
CAAAAGCTATAAAGATTCACACTCACTTGGTAAAGAGCATCACTCTGAAGGAGAGACTTGTGGCCAACCTCCTGGTGCCTCCC
GGCTTGGGTCTGGCTCTCCTCTCCGGCGGTGGCCATTGATTCAAACTCTCTCTCTCTCTCTCTCTCTCTCTCTGCTTAG
GAATATTGTAATATGTTGTGATCAAGAAATGTTGCAGTTGACCTAAGTTCTATGAGCAAGCATGATGCAACGTATATAAGGAC
AGTGATGGTCCG

>EUSP-FX consenso
CGGTCCGGATTCCCGGGTCGACCACGCGTCCGCCCACGCGTCCGCTTTTTCTCTGGAAATGAAAGGGTATTAAACAAGTCCAA
GATCAAGAGGCCTGCATGCTTCCAGGAGCTCACCAACCCTGGACCAGAGAGAGAGAGAGAGAGTTTGAATCAATGGCCACCGC
CGGAGAGGAGAGCCAGAACCAAGCCGGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGATGCTCTTTACCAAT
ATATTTTGGAGACTAGCGTGTACCCAAGAGAGCCCGAGCCCATGAAGGAGCTCAGGGAAATTACAGCAAAACATCCATGGAAC
ATAATGACAACATCAGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCATCAACGCCAAGAACACCATGGAGATTGG
TGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTTGCTCTTCCTGATGACGGAAAGATTTTGGCTATGGACATTAACAGAG
AGAACTATGAACTTGGCCTGCCGGTCATCCAAAAAGCCGGTGTTGCCCACAAGATTGACTTCAGAGAAGGCCCTGCTTTGCCT
ATTCTTGATCAGTTGATCGAAGATGGGAAGCAAGGGACGTTCGACTTCATATTCGTGGACGCGGACAAGGACAATTACCTCAA
CTACCACAAGAGGCTGATCGAGCTTGTCAAGGTTGGAGGCCTCATTGGCTACGACAACACCCTATGGAACGGCTC

>EUGL-XY-002 consenso
CCCGGGTCGACCCACGCGTCCGGNACCAGAACCAAGCCGGGAGGCACCAGGAGGTTGGCCACAAGTCTCTCCTTCAGAGTGAT
GCTCTTTACCAATATATTTTGGAGACCAGCGTGTACCCAAGAGAGCCTGAGCCCATGAAGGAGCTCAGGGAAATAACAGCAAA
ACATCCATGGAACATAATGACAACATCAGCAGACGAAGGGCAGTTCTTGAACATGCTTCTCAAGCTCAACGCCAAGAACACCA
TGGAGATTGGTGTCTTCACTGGCTACTCTCTCCTTGCCACCGCTCTTGCTCTTCCTGATGACGGAAAGATTTTGGCTATAGAC
GATAACTCAAGAATTGAGTATGGCAATCGAATGGCGCTCATTTCAGACCAGCTTTACGAGGCGGCCAAAACAAACTGCGATGG
CGACTATGTGAATGTTGGCGATGATAATCTTCAATGCGTCGTTGATCTGAATGCCATCAACGAGCTCACCTCCAAAGTCCTGA
CGGTCCACATTCTGGAGCCGGATTGCAGCAACGCCCTGCCCGATCCACAAGATGACGCTTCCGTTCGAAGATATCTCCGGGAA
AACGCCACGGCCGCATCCACGGACGACCCCCTCAAAGGTCCTGCGCTCTGGTGCCGTAACTACAACCACATGGTCTGCTATAT
CTGGGCGAACCTTAAAAGCGTCCAAGAAGCACTTCAAGTCCGACCGGGGACGAAGACCTACTGGCCATAC

>EUGR-ML-A04
AAACACTAAACAGCACCCCTTTTTCCATTTATTGGCATGCACTGTGAGACTTAACAGACTGTATCCATTCGAAATCGCGCGCA
AAAGAAGTGAATACAACAAAGGGTACAGACAAAGAAATGCCGTTTATATAATTGCTTATGATCCTGATTTGTACCTGCAAAGA
ACTTTGCATCTCAACACCCTATCGTGAACATAAAATCCGGGCATCACCATGATCTTCACCCGGCTCGAACTGTGCTGTCTCTG
CCGATCCACGAGTATGTCCTCCATGTAGTGCGGATCAAAACTTCTGTTCTCTTCGACCCTCAAAATCCCGATGGGTGGGCTGA
AAGAAAAGGCAAGCAGATGAAGCAGCCATATGCATTTGGCGGACACAAAGAACGCTTGCAGGAGTTGCTCTGGCCACGGTCTG
GTCCAATTCAAGGTCGTGATGATGCAACTCATCTTCTGATCGCAGAACTTGCTGAACTCT

>EUGR-ML_F05
AATTTTAGTACTTGATTGCATCATTTAGAAGTTGAGTCGGAGATGCCGTCCCATCATCCTACTCATGCTGATACTAAGAACGT
ATCTTTGAATGGATTTTCTTGGAAGAATATTTCCTTTCTACTGAGGGACTAATATGTAAATAATATATACTTTGCTTGCATAC
TTCTAGTGATCTGTCAGGTAGCTCTTCATCACTCACTTATGCGCGTTCACTTGATGGGAATGGTCCTCGCGGTAAGTCATTCA
GTCGCTATCCCATGCATTGGCAAAGAGTCCAGTTCTTTTGGCCTCTTGTAGTTCTTCAACAGGAGGGAAGAACAAACAACACT
GACAGAGGAGGCTGCCATTGCCACTCCAGCAATCCAAGGTGGTAGCCGGAATCCTGTAAATGGGAAGAGGATCCCCGCTGCTA
TTGGCATTCCTATGACGTTGTAACCGAAGGCCCAGATGTAGTT

>EUGL-XY-001
GCGTCCGCATCACTCTCTGCCGCCGGGCTCAGCTGATCACACCGGAAAGGAGATCTCAAAAACAAGGATCGATGAAATGAGAA
ACTACCTTTAATACTTTCCTTCCTTTCTATTTTTTCCATCTTCTGTCTTATGTTGTCTTTGAACCGTTGAGCATGTATTTGTA
TTCAAATGAACGATTAAGGATTGAGAAGAAGTTGCTAATTTGGCTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGGGGGGGGGCCCCCTTAAAAAAATCCCCTTTTTCCCCCGGGGGGGG
GGGAAAAAATTTTTTTTTTTTGGGGGCCCCCAAATTTTTTTTCGGGGGGGCGTTTTTAAAACCGGGGGGGGGGAAAAAAAA

>CL5089Contig1 consenso
CCACGCGTCCGGGAAGAAGCCGAGCAAGCGAAGTTGCAGACGCCATTGAAAAAAGACACGAAAGAGATCAAGAAGGAGCTTAA
GAAGCATCACCCTGTGCCGCCGGGTCAGTTGATCGCACCAGAAAGGAGATCTCAAAATCAAGCATTGATGAAATGAGGAACTA
CCCTTGATAATTTCCTTCCTTTCTATTTTTTCCATCTTCTGTCTTATATTGTCTTTGAACCGTTGAGCATGTATTTGTATTCA
AATGAACGATTAAGGAGTGAGAAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAAAAGTTCATTT
ATGCA

APÊNDICE J

Parte do alinhamento do contig 1 do cluster 5809 com o contig 2 do cluster 6, representante no banco de dados do Genolyptus da isoforma 2 do gene CCoAOMT. Alinhamento realizado pelo software ClustalW.

```
CL5089Contig1       ---------------CCACGCGTCCGGGAAGAA--GCCGAG----CAAGC-GAAGTTGCA 38
CL6Contig2          CGCCGACGCGCCCCTCCGCAAGTACGTGCGGTACTACCGGGACTTCGTGCTGGAGCTCAA 780
                                  ** *  ** ** *  * *   *** *    *  ** * ** *  *

CL5089Contig1       GACGCCATTGAAAAAAGACACGAAA---GAGATCAAGAAGGAGC---TTAAGAA--GCAT 90
CL6Contig2          CAAGGCCCTCGCCGTGGACCCGAGGATCGAGATCTGCATGCTTCCCGTCGGGGATGGCAT 840
                     * * *  *       *** ***     ******   * *   *   *   * *  ****

CL5089Contig1       CACCCTGTGCCGCCGGGTCAGTTGATCGCACCAGAAAGGAGATCTCAAAATCAAGCATTG 150
CL6Contig2          CACCCTGTGCCGCCGGGTCAGTTGATCGCACCAGAAAGGAGATCTCAAAATCAAGCATTG 900
                    ************************************************************

CL5089Contig1       ATGAAATGAGGAACTACCCTTGATAATTTCCTTCCTTTCTATTTTTTCCATCTTCTGTCT 210
CL6Contig2          ATGAAATGAGGAACTACCCTTGATAATTTCCTTCCTTTCTATTTTTTCCATCTTCTGTCT 960
                    ************************************************************

CL5089Contig1       TATATTGTCTTTGAACCGTTGAGCATGTATTTGTATTCAAATGAACGATTAAGGAGTGAG 270
CL6Contig2               TATATTGTCTTTGAACCGTTGAGCATGTATTTGTATTCAAATGAACGATTAAGGAGTGAG
1020
                    ************************************************************

CL5089Contig1       AAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAAAAGTTCAT 330
CL6Contig2               AAGAAGTTGCTAATTTGGCTTACAATGGAGCTACATTCAAATTGTATAATAAAAGTTCAT
1080
                    ************************************************************

CL5089Contig1       TTATGCA 337
CL6Contig2          TTATGCT 1087
                    ******
```

APÊNDICE K

Seqüências nucleotídicas e protéicas dos contigs utilizados no estudo, os quais foram o resultado da busca no banco de dados para o gene 4CL.

>CL273Contig1
CGTCCGCTCGCTCGCGCGGAACCAGTCCGGCGAGATCTGCATCCGGGGTCACCAGATCATGAAAGGTTATCTGAA
CGACGCCGAGCCGACCGCAAATACCATAGACAAAGAAGGGTGGCTGCACACCGGCGACATCGGCTACATAGACGA
TGACGACGAGCTCTTCATTGTCGATCGGTTGAAGGAACTCATCAAGTACAAGGGCTTCCAGGTTGCTCCGGCCGA
GCTAGAGGCAATGCTGATTGCACACCCAAGTATCTCGGATGCCGCTGTTGTGCCGATGAAGGATGAGGTTGCCGG
TGAGGTTCCTGTTGCATTCGTGGTGAAATCCAATGGTTCCGTAATCACCGAGGACGAAATCAAGCAATACATCTC
GAAGCAGGTCGTGTTTTACAAGAGGATCAAGCGGGTTTTCTTCACGGACGCAATTCCGAAAGCCCCCTCCGGAAA
AATCTTGAGGAAGGACCTAAGAGCAAAGTTGGCCTCTGGTGTTTACAATTAATTTCTCATACCCTTTTCTTTTTC
AACCCTGCCCCTGTACTTGCTTAAAGACCCATGTAGTTGAAATGAATGTAACCTCTTCGGAGGGGCCAAATATGG
AAGGGGGAAAGAAAGACATATGGCGATGATTTGATTTCACATGCTATTGTAATGTATTTATTGTTTCAATTCCGA
ATTAGACAAAGTGCTTAAAGCTCTCTTTTCGGGATTTTTTTTTCATTAATGTATATAATTGCGGA

>CL273ctg1
MKGYLNDAEPTANTIDKEGWLHTGDIGYIDDDDELFIVDRLKELIKYKGFQVAPAELEAMLIAHPSISDAAVVPM
KDEVAGEVPVAFVVKSNGSVITEDEIKQYISKQVVFYKRIKRVFFTDAIPKAPSGKILRKDLRAKLASGVYN

>CL273Contig2
CGTCCGCCAAACGCTCACCTTCTCATCATCAGCCCTCTCTTTCTCTCTCTCTCTCTCTCTCTCTCGATTCTCC
GCCCCGCCACGACAATGGAGGCGAAGCCGTCGGAGCAGCCCCGCGAGTTCATCTTCCGGTCGAAGCTCCCCGACA
TCTACATTCCCGACAACCTCTCCCTCCACGCCTACTGCTTCGAGAACATCTCCGAGTTCGCCGACCGCCCCTGCG
TCATCAACGGGGCCACCGGCCGGACCTACACCTATGCCGAGGTCGAGCTGATCTCCCGCCGGGTCTCAGCCGGCC
TCAACGGGCTCGGCGTCGGACAGGGCGACGTGATCATGCTGCTCCTCCAGAACTGCCCTGAGTTCGTGTTCGCAT
TCCTCGGCGCGTCCTACCGGGGCGCCATCAGCACGACCGCCAACCCGTTCTACACCCCGGGAGAGATCGCCAAGC
AGGCCTCAGCTGCCCAGGCCAAGATCGTAATCACGCAGGCCGCGTACGCCGACAAGGTGAGGCCGTTCGCGGAGG
AGAACGGGGTCAAGGTCGTGTGCATCGATACCGCGCCGGAGGGCTGCCTGCACTTCTCGGAATTGATGCAGGCGG
ATGAGAACGCCGCCCCCGCGGCGGACGTCAAGCCGGACGACGTCTTGGCACTCCCCTATTCGTCGGGCACGACGG
GGCTCCCCAAGGGGGTGATGCTCACGCACAGGGGTCAAGTGAGTAGCGTGGCGCAGCAGGTCGACGGAGACAACC
CCAACTTGTACTTCCACAAGGAGGACGTGATCCTGTGCACGCTCCCGTTGTTCCACATATACTCCCTCAACTCGG
TGATGTTCTGCGCGCTCCGTGTCGGCGCTGCCATCCTGATCATGCAGAAGTTCGAGATCGTGGCGCTGATGGAGC
TCGTGCAGCGGTACCGGGTGACGATCCTGCCCATCGTCCCGCCAATCGTGCTGGCGATCGCCAAGAGCGCCGAGG
TGGACCGGTACGACCTGTCGTCGATCCGGACCATCATGTCGGGTGCGGCCCCGATGGGGAAGGAGCTCGAGGACA
CCGTGCGAGCCAAGCTGCCGAATGCCAAGCTCGGACAGGGCTATGGGATGACGGAGGCGGGCCCGGTGCTGGCAA
TGTGCCTGGCATTTGCAAAGGAGCCGTTCGAGATCAAGTCAGGCGCGTGCGGGACCGTCGTGAGGAACGCGGAGA
TGAAGATCGTCGACCCGGAGACAGGGGCCTCGCTCCCGCGGAACCAGGCCGGCGAGATCTGCATCCGGGGTCACC
AGATCATGAAAGGTTATCTGAACGACCCCGAAGCGACCGCTAATACCATAGACAAAGAAGGGTGGCTGCACACCG
GCGACATCGGCTACATAGACGATGACGACGAGCTCTTCATTGTCGATCGGTTGAAGGAACTCATCAAGTACAAGG
GCTTCCAGGTCGCTCCGGCCGAATTAGAGGCAATGCTGATTGCACACCCAAGTATCTCGGATGCTGCCGTTGTGC
CGATGAAGGATGAGGTTGCCAGTGAGGTTCCTGTTGCATTCGTGGTGAAATCCAATGGTTCCGTAATCACTGAGG
ACGAAATCAAGCAATACATCTCGAAGCAGGTCGTGTTTTACAAGAGGATCAACCGGGTTTTCTTCACGGACGCAA
TTCCGAAAGCCCCCTCCGGCAAAATCTTGAGGAAGGACCTAAGAGCAAAGTTGGCCTCCGGTGTTTACAATTAAT
TTCTCATACCCTTTTCTTTTTCAACCCTGCCCCTGTACTTGCTTAAAGACCCATGTAGTTGAAATGAATGTAACC
TCTTCGGAGGGGCCAAATATGGAAGGGGGAAAGAAAGACATATGGCGATGATTTGATTTCACATGCTATTGTAAT
GTATTTATTGTTTCAATTCCGAATTAGACAAAGTGCTTAAAGCTCTCTTTTCGGATTTTTTTTTTCATTAATGTA
TAATAATTGCGGACATTACAATATAGTGTACAACGTGATTTGAGCTTGATGAATTACAAGATTGGAAGAACT

>CL273Ctg2
MLLLQNCPEFVFAFLGASYRGAISTTANPFYTPGEIAKQASAAQAKIVITQAAYADKVRPFAEENGVKVVCIDTA
PEGCLHFSELMQADENAAPAADVKPDDVLALPYSSGTTGLPKGVMLTHRGQVSSVAQQVDGDNPNLYFHKEDVIL
CTLPLFHIYSLNSVMFCALRVGAAILIMQKFEIVALMELVQRYRVTILPIVPPIVLAIAKSAEVDRYDLSSIRTI
MSGAAPMGKELEDTVRAKLPNAKLGQGYGMTEAGPVLAMCLAFAKEPFEIKSGACGTVVRNAEMKIVDPETGASL
PRNQAGEICIRGHQIMKGYLNDPEATANTIDKEGWLHTGDIGYIDDDDELFIVDRLKELIKYKGFQVAPAELEAM
LIAHPSISDAAVVPMKDEVASEVPVAFVVKSNGSVITEDEIKQYISKQVVFYKRINRVFFTDAIPKAPSGKILRK
DLRAKLASGVYN

>CL4405Contig1

GGTCCGGATTCCCGGGTCGACCCACGCGTCCGCGGACGCGTGGGTTTACACTTTTGAAACCGGACTTTGAGCTGA
TATTGAATCGCTCCTTTTTTAGAAAAAGGTGGTGGTTACTAAAGTTTAAATCATGAGTTAAGTTTTCCTTACAGCT
GCAAAAGAAAAGGACAGAGAGAAAGAAAGGACCCGAAATTGATGAGTGGACCTAAGAGGGAGGCTTTGCTTGAAT
GTGGAAGCAGTTGAACTCATTAAATGACTTCTGCTCGCAATCACATGATCTATTCACTCAATTACATTCTTTCGG
TGGTTAGAACTCCATCGCTTTCATTTCTATTGCGAAAGCTTTAACCATCCTGTTCAATTTCCTGCACTGTTTACC
TACAGAACATTTCTGTATTCTCGTCATCTTTCTCCTCCAATTTTAAATGCAAGACGGTAACGCGCATAACTTTAT
ATTTAACATCCCTATAAAGGGTTTGTGATTTCGATTGGCTTTTAGCATTTGTTGACGCGGTGCAGGGATATGGAA
TGACGGAAGCGGGACCGGTGCTTTCTATGTGCTTGGGGTTCGCCAAGCAACCCTTCCCAACCAAATCGGGTTCGT
GCGGGACGGTTGTTCGGAATGCAGAGCTCAAAGTCATCGACCCCGAGACCGGTTCCTCCCTTGGCTACAACCAGC
CCGGCGAGATATGCATTCGTGGCCAACAAATTATGAAAGGATACCTGAACGACCCCGAGGCGACTTCGATCACCA
TTGACGCGGATGGCTGGCTTCACACCGGTGACATAGGCTATGTCGATGATGATGATGAGATCTTCATTGTGGACA
GAGTGAAGGAAATAATCAAATTCAAGGGGTTCCAGGTGCCACCAGCCGAGCTTGAAGCGCTTCTAGTAAGCCACC
CATCCATTGCGGATGCAGCTGTCGTCCCGCAAAAGGATGAAGTCGCCGGTGAAGTCCCGGTTGCATTTGTGGTGA
GATCGAATGGCTTTGAACTGACGGAAGAAGCAGTAAAGGAGTTCATAGCCAAACAGGTGGTTTTCTATAAGAAGC
TTCACAAGGTCCACTTCGTGCATGCAATCCCAAAGTCTCCGTCCGGGAAGATACTGAGGAAAGATCTCAGAGCCA
AGTTAGCTACTGCAGCCCCCATTTCTTAAATTCGTTAGGTTGTGATTTCGGTGATTTGTAAGCATTTTTACTGGT
GGTGGGTTCTCCATGTTCTTCTTCCTTTTTCTTTACCGTCTTTCCTGTTATAATTGCATAAACAGAAGTGGGGAG
AACAAGAAGTGGTGAGCAATGTGATCCTCTTGAAGATTCAATAATATTGTCGATCCTTCCCA

>CL4405ctg1
MTEAGPVLSMCLGFAKQPFPTKSGSCGTVVRNAELKVIDPETGSSLGYNQPGEICIRGQQIMKGYLNDPEATSIT
IDADGWLHTGDIGYVDDDDEIFIVDRVKEIIKFKGFQVPPAELEALLVSHPSIADAAVVPQKDEVAGEVPVAFVV
RSNGFELTEEAVKEFIAKQVVFYKKLHKVHFVHAIPKSPSGKILRKDLRAKLATAAPIS

>CL2848Contig1
AAGGCGTGATCTTGACTCCCCGGAATTTCATCCCGGCATCCCTGATGATGACCATGGATCAAGAAATGGCGGGGG
ACATGCACCGTGTGTTCCTCTGCGTCCTGCCCATGTTCCACGTGTTCGGGCTCGCGGTGATCGCTTATTCGCAGC
TCCAGAAGGGGAACGCGCTCGTGTCGATGGGGAGGTTCGAATTCGACTCGCTCTTAAGGGCGGTCGAGAAGTACA
GGATCACGCATTTGTGGGTTGTCCCCCCCCATTGTACTTGCTTTGGCTAAGCAGAGTGCGGTGAGGAAGTATGACC
TCTCGTCCTTGAAGCACATTGGTTCCGGCGCTGCACCTTTAGGGAAGGATGTGATGGAGGATTGTGCTAAGAATT
TCCCACAGGCCGACGTGATGCAGGGTTATGGTATGACAGAAACTTGTGGGATCGTCTCTGTGGAGAATGCAAATT
TCGGCCCTCGGCATACTGGTTCCGCTGGACAACTAGTTGCAGGAGTTGAAGCTCAAGTTATCAGCGTGGATACAC
TAAAATCTCTTCCCCCTAATCAGTTAGGGGAAATATGGGTTCGTGGACCTAACATGATGAAAGGATATTATAACA
ATCCACAAGCAACTAAATTGACAATTGATAACAAGGGTTGGGTGCACACTGGAGACCTTGGATATTTTGATGAGG
AAGGGCAACTATATGTTGTTGATCGAATCAAAGAGCTCATCAAGTACAAAGGTTTTCAGATTGCTCCAGCTGAGC
TTGAAGGACTCCTTCTTTCACATCCTGAAATTTTAGATGCTGTTGTCATTCCATTTCCTGATGCTGAAGCTGGTG
AAGTTCCTATTGCATATGTCGTTCGCTCACCTACCAGCTCTCTAACTGA

>CL2848ctg1
MMTMDQEMAGDMHRVFLCVLPMFHVFGLAVIAYSQLQKGNALVSMGRFEFDSLLRAVEKYRITHLWVVPPIVLAL
AKQSAVRKYDLSSLKHIGSGAAPLGKDVMEDCAKNFPQADVMQGYGMTETCGIVSVENANFGPRHTGSAGQLVAG
VEAQVISVDTLKSLPPNQLGEIWVRGPNMMKGYYNNPQATKLTIDNKGWVHTGDLGYFDEEGQLYVVDRIKELIK
YKGFQIAPAELEGLLLSHPEILDAVVIPFPDAEAGEVPIAYVVRSPTSSLT

APÊNDICE L

Alinhamento dos contigs obtidos na busca realizada no banco de dados para o gene 4CL com as suas isoformas. Alinhamento realizado pelo programa Align pelo método water.

```
#=======================================
#
# Aligned_sequences: 2
# 1: CL273Ctg2
# 2: 4CL1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 468
# Identity:     347/468 (74.1%)
# Similarity:   400/468 (85.5%)
# Gaps:           8/468 ( 1.7%)
# Score: 1831.0
#
#
#=======================================

CL273Ctg2          1 MLLLQNCPEFVFAFLGASYRGAISTTANPFYTPGEIAKQASAAQAKIVIT     50
                     ||||.||||||.:||.||:|||.:|.||||:||.||||||.|:..|::||
4CL1              94 MLLLPNCPEFVLSFLAASFRGATATAANPFFTPAEIAKQAKASNTKLIIT    143

CL273Ctg2         51 QAAYADKVRPFAEENGVKVVCID-----TAPEGCLHFSELMQADENAAPA     95
                     :|.|.||::|...::||.:||||     ..||||||.|:||.|:...|:..
4CL1             144 EARYVDKIKPLQNDDGVVIVCIDDNESVPIPEGCLRFTELTQSTTEASEV    193

CL273Ctg2         96 AD---VKPDDVLALPYSSGTTGLPKGVMLTHRGQVSSVAQQVDGDNPNLY    142
                     .|   :.||||:||||||||||||||||:|.|:|||||||:|||||
4CL1             194 IDSVEISPDDVVALPYSSGTTGLPKGVMLTHKGLVTSVAQQVDGENPNLY    243

CL273Ctg2        143 FHKEDVILCTLPLFHIYSLNSVMFCALRVGAAILIMQKFEIVALMELVQR    192
                     ||.:|||||.||:||||:|||:|.|.||||||||||.||||..|:||:||
4CL1             244 FHSDDVILCVLPMFHIYALNSIMLCGLRVGAAILIMPKFEINLLLELIQR    293

CL273Ctg2        193 YRVTILPIVPPIVLAIAKSAEVDRYDLSSIRTIMSGAAPMGKELEDTVRA    242
                     .:||:.|:||||||||||:|.::|||||||.:.|||||:||||||.|.|
4CL1             294 CKVTVAPMVPPIVLAIAKSSETEKYDLSSIRVVKSGAAPLGKELEDAVNA    343

CL273Ctg2        243 KLPNAKLGQGYGMTEAGPVLAMCLAFAKEPFEIKSGACGTVVRNAEMKIV    292
                     |.|||||||||||||||||.|.|||||||.:||||||||||||||||||
4CL1             344 KFPNAKLGQGYGMTEAGPVLAMSLGFAKEPFPVKSGACGTVVRNAEMKIV    393

CL273Ctg2        293 DPETGASLPRNQAGEICIRGHQIMKGYLNDPEATANTIDKEGWLHTGDIG    342
                     ||:||.||.|||.||||||||||||||:|.|||.||||:|||||||||
4CL1             394 DPDTGDSLSRNQPGEICIRGHQIMKGYLNNPAATAETIDKDGWLHTGDIG    443

CL273Ctg2        343 YIDDDDELFIVDRLKELIKYKGFQVAPAELEAMLIAHPSISDAAVVPMKD    392
                     .|||||||||||||||||||||||||||||:||.||.|:|.|||.||:
4CL1             444 LIDDDDELFIVDRLKELIKYKGFQVAPAELEALLIGHPDITDVAVVAMKE    493

CL273Ctg2        393 EVASEVPVAFVVKSNGSVITEDEIKQYISKQVVFYKRINRVFFTDAIPKA    442
                     |.|.||||||||..|.::||::||::||||||||||:||||::||||
4CL1             494 EAAGEVPVAFVVKSKDSELSEDDVKQFVSKQVVFYKRINKVFFTESIPKA    543

CL273Ctg2        443 PSGKILRKDLRAKLASGV    460
                     ||||||||||||||:|:
4CL1             544 PSGKILRKDLRAKLANGL    561
```

```
#=======================================
# Aligned_sequences: 2
# 1: CL273ctg1
# 2: 4CL2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 147
# Identity:     117/147 (79.6%)
# Similarity:   134/147 (91.2%)
# Gaps:           0/147 ( 0.0%)
# Score: 613.0
#
#
#=======================================

CL273ctg1          1 MKGYLNDAEPTANTIDKEGWLHTGDIGYIDDDDELFIVDRLKELIKYKGF     50
                     |||||||...||:||||:|||||||:|:|||||||||||||||||||||
4CL2             410 MKGYLNDPLATASTIDKDGWLHTGDVGFIDDDDELFIVDRLKELIKYKGF    459

CL273ctg1         51 QVAPAELEAMLIAHPSISDAAVVPMKDEVAGEVPVAFVVKSNGSVITEDE    100
                     ||||||||::||.||.|:|.|||.||:|.||||||||:|..|.|:|||
4CL2             460 QVAPAELESLLIGHPEINDVAVVAMKEEDAGEVPVAFVVRSKDSNISEDE    509

CL273ctg1        101 IKQYISKQVVFYKRIKRVFFTDAIPKAPSGKILRKDLRAKLASGVYN    147
                     |||::||||||||||.:|||||:||||||||||||||:||:|:.|
4CL2             510 IKQFVSKQVVFYKRINKVFFTDSIPKAPSGKILRKDLRARLANGLMN    556


#=======================================
#
# Aligned_sequences: 2
# 1: CL4405ctg1
# 2: 4CL3
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 202
# Identity:     171/202 (84.7%)
# Similarity:   187/202 (92.6%)
# Gaps:           0/202 ( 0.0%)
# Score: 889.0
#
#
#=======================================

CL4405ctg1         1 MTEAGPVLSMCLGFAKQPFPTKSGSCGTVVRNAELKVIDPETGSSLGYNQ     50
                     ||||||||||.|||||:|.||||||||||||||||:..||..||||||
4CL3             359 MTEAGPVLSMSLGFAKEPIPTKSGSCGTVVRNAELKVVHLETRLSLGYNQ    408

CL4405ctg1        51 PGEICIRGQQIMKGYLNDPEATSITIDADGWLHTGDIGYVDDDDEIFIVD    100
                     |||||||||||||.|||||||||.|||.:||||||||||||:||||||||
4CL3             409 PGEICIRGQQIMKEYLNDPEATSATIDEEGWLHTGDIGYVDEDDEIFIVD    458

CL4405ctg1       101 RVKEIIKFKGFQVPPAELEALLVSHPSIADAAVVPQKDEVAGEVPVAFVV    150
                     |:||:||||||||||||||:||::|.||||||||||.||||||||||||
4CL3             459 RLKEVIKFKGFQVPPAELESLLINHHSIADAAVVPQNDEVAGEVPVAFVV    508

CL4405ctg1       151 RSNGFELTEEAVKEFIAKQVVFYKKLHKVHFVHAIPKSPSGKILRKDLRA    200
                     ||||.::|||.|||::||||||||:||||.||.:|||||||||||||:|
4CL3             509 RSNGNDITEEDVKEYVAKQVVFYKRLHKVFFVASIPKSPSGKILRKDLKA    558

CL4405ctg1       201 KL    202
                     ||
4CL3             559 KL    560
```

```
#=======================================
#
# Aligned_sequences: 2
# 1: CL273ctg1
# 2: 4CL4
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 141
# Identity:     111/141 (78.7%)
# Similarity:   126/141 (89.4%)
# Gaps:           0/141 ( 0.0%)
# Score: 585.0
#
#
#=======================================

CL273ctg1          1 MKGYLNDAEPTANTIDKEGWLHTGDIGYIDDDDELFIVDRLKELIKYKGF     50
                     |||||||.|.||.||||:|||||||||::|||||:||||||||||:||:
4CL4             424 MKGYLNDPEATARTIDKDGWLHTGDIGFVDDDDEIFIVDRLKELIKFKGY    473

CL273ctg1         51 QVAPAELEAMLIAHPSISDAAVVPMKDEVAGEVPVAFVVKSNGSVITEDE    100
                     |||||||||:||:||||.|||||.|||||.|||||||.:|.||.:|||:
4CL4             474 QVAPAELEALLISHPSIDDAAVVAMKDEVADEVPVAFVARSQGSQLTEDD    523

CL273ctg1        101 IKQYISKQVVFYKRIKRVFFTDAIPKAPSGKILRKDLRAKL    141
                     :|.|::||||.||||||.|||.:.||||.||||||||||||
4CL4             524 VKSYVNKQVVHYKRIKMVFFIEVIPKAVSGKILRKDLRAKL    564
```

APÊNDICE M

Alinhamento do contig 2 do cluster 273, representante da isoforma 1 do gene 4CL no banco de dados, com a seqüência gênica obtida pela montagem de seqüências geradas pelo seqüenciamento da biblioteca *shotgun* de BAC. Alinhamento e inferências realizadas por meio do programa *EST2Genome*.

```
Note Best alignment is between forward est and forward genome, and splice  sites
imply forward gene
Exon       979  97.9     1  1022 4CL_EUGR        68  1088 CL273Contig2
+Intron    -20   0.0  1023  2273 4CL_EUGR
Exon       193  98.5  2274  2472 4CL_EUGR      1089  1287 CL273Contig2
+Intron    -20   0.0  2473  4063 4CL_EUGR
Exon       198  96.3  4064  4277 4CL_EUGR      1288  1501 CL273Contig2
+Intron    -20   0.0  4278  4385 4CL_EUGR
Exon       101  98.1  4386  4490 4CL_EUGR      1502  1606 CL273Contig2

Span      1411  97.8     1  4490 4CL_EUGR        68  1606 CL273Contig2

Segment    484  98.6     1   498 4CL_EUGR        68   565 CL273Contig2
Segment    497  97.5   500  1022 4CL_EUGR       566  1088 CL273Contig2
Segment    193  98.5  2274  2472 4CL_EUGR      1089  1287 CL273Contig2
Segment    198  96.3  4064  4277 4CL_EUGR      1288  1501 CL273Contig2
Segment    101  98.1  4386  4490 4CL_EUGR      1502  1606 CL273Contig2


4CL_EUGR vs CL273Contig2:

    4CL_EUGR      1 GATTCTCCGCCCCGCCACGACAATGGAGGCGAAGCCGTCGGAGCAGCCCC     50
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2     68 GATTCTCCGCCCCGCCACGACAATGGAGGCGAAGCCGTCGGAGCAGCCCC    117

    4CL_EUGR     51 GCGAGTTCATCTTCCGGTCGAAGCTCCCCGACATCTACATTCCCGACAAC    100
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    118 GCGAGTTCATCTTCCGGTCGAAGCTCCCCGACATCTACATTCCCGACAAC    167

    4CL_EUGR    101 CTCTCCCTCCACGCCTACTGCTTCGAGAACATCTCCGAGTTCGCCGACCG    150
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    168 CTCTCCCTCCACGCCTACTGCTTCGAGAACATCTCCGAGTTCGCCGACCG    217

    4CL_EUGR    151 CCCCTGCGTCATCAACGGGGCCACCGGCCGGACCTACACCTATGCCGAGG    200
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    218 CCCCTGCGTCATCAACGGGGCCACCGGCCGGACCTACACCTATGCCGAGG    267

    4CL_EUGR    201 TCGAGCTGATCTCCCGCCGGGTCTCAGCCGGCCTCAACGGGCTCGGCGTC    250
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    268 TCGAGCTGATCTCCCGCCGGGTCTCAGCCGGCCTCAACGGGCTCGGCGTC    317

    4CL_EUGR    251 GGACAGGGCGACGTGATCATGCTGCTCCTCCAGAACTGCCCTGAGTTCGT    300
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    318 GGACAGGGCGACGTGATCATGCTGCTCCTCCAGAACTGCCCTGAGTTCGT    367

    4CL_EUGR    301 GTTCGCGTTCCTCGGCGCGTCCTACCGGGGCGCCATCAGCACGACCGCGA    350
                    |||||| |||||||||||||||||||||||||||||||||||||||||| |
CL273Contig2    368 GTTCGCATTCCTCGGCGCGTCCTACCGGGGCGCCATCAGCACGACCGCCA    417

    4CL_EUGR    351 ACCCGTTCTACACCCCGGGCGAGATCGCCAAGCAGGCCTCAGCTGCCCGG    400
                    |||||||||||||||||| |||||||||||||||||||||||||||||| |
CL273Contig2    418 ACCCGTTCTACACCCCGGGAGAGATCGCCAAGCAGGCCTCAGCTGCCCAG    467

    4CL_EUGR    401 GCCAAGATCGTGATCACGCAGGCCGCGTTCGCCGACAAGGTGAGGCCGTT    450
                    |||||||||| ||||||||||||||| ||||||||||||||||||||||||
CL273Contig2    468 GCCAAGATCGTAATCACGCAGGCCGCGTACGCCGACAAGGTGAGGCCGTT    517

    4CL_EUGR    451 CGCGGAGGAGAACGGGGTGAAGGTCGTGTGCATCGATACCGCGCCGGAGG    500
```

```
                    |||||||||||||||||| |||||||||||||||||||||||||||||| |
CL273Contig2    518 CGCGGAGGAGAACGGGGTCAAGGTCGTGTGCATCGATACCGCGCCGGA-G    566

    4CL_EUGR    501 GGCTGCCTGCACTTCTCGGAATTGATGCAGGCGGACGAGAACGCCGCCCC    550
                    |||||||||||||||||||||||||||||||||||| ||||||||||||||
CL273Contig2    567 GGCTGCCTGCACTTCTCGGAATTGATGCAGGCGGATGAGAACGCCGCCCC    616

    4CL_EUGR    551 CGCGGCGGACGTCAAGCCGGACGACGTCTTGGCGCTCCCCTATTCGTCGG    600
                    |||||||||||||||||||||||||||||||||| ||||||||||||||||
CL273Contig2    617 CGCGGCGGACGTCAAGCCGGACGACGTCTTGGCACTCCCCTATTCGTCGG    666

    4CL_EUGR    601 GCACGACGGGGCTTCCCAAGGGAGTGATGCTTACGCACAGGGGTCAAGTG    650
                    |||||||||||||| |||||||| |||||||| ||||||||||||||||||
CL273Contig2    667 GCACGACGGGGCTCCCCAAGGGGGTGATGCTCACGCACAGGGGTCAAGTG    716

    4CL_EUGR    651 ACCAGCGTGGCGCAGCAGGTCGACGGAGACAACCCCAACTTGTACTTCCA    700
                    |  ||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    717 AGTAGCGTGGCGCAGCAGGTCGACGGAGACAACCCCAACTTGTACTTCCA    766

    4CL_EUGR    701 CAAGGAGGACGTGATCCTGTGCACGCTCCCGTTGTTCCACATATACTCCC    750
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    767 CAAGGAGGACGTGATCCTGTGCACGCTCCCGTTGTTCCACATATACTCCC    816

    4CL_EUGR    751 TCAACTCGGTGATGTTCTGCGCGCTCCGTGTAGGCGCCGCCATCCTGATC    800
                    |||||||||||||||||||||||||||||||| ||||| ||||||||||||
CL273Contig2    817 TCAACTCGGTGATGTTCTGCGCGCTCCGTGTCGGCGCTGCCATCCTGATC    866

    4CL_EUGR    801 ATGCAGAAGTTCGAGATCGTGGCGCTGATGGAGCTCGTGCAGCGGTACCG    850
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    867 ATGCAGAAGTTCGAGATCGTGGCGCTGATGGAGCTCGTGCAGCGGTACCG    916

    4CL_EUGR    851 GGTGACGATCCTGCCCATTGTCCCGCCGATCGTGCTGGAGATCGCCAAGA    900
                    |||||||||||||||||| |||||||| |||||||||| ||||||||||||
CL273Contig2    917 GGTGACGATCCTGCCCATCGTCCCGCCAATCGTGCTGGCGATCGCCAAGA    966

    4CL_EUGR    901 GCGCCGAGGTGGACCGGTACGACCTGTCGTCGATCCGGACCATCATGTCG    950
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2    967 GCGCCGAGGTGGACCGGTACGACCTGTCGTCGATCCGGACCATCATGTCG   1016

    4CL_EUGR    951 GGTGCGGCCCCGATGGGGAAGGAGCTCGAGGACACCGTGCGAGCCAAGCT   1000
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2   1017 GGTGCGGCCCCGATGGGGAAGGAGCTCGAGGACACCGTGCGAGCCAAGCT   1066

    4CL_EUGR   1001 GCCCAATGCCAAGCTCGGACAGgtgaa......tgcagGGCTATGGGATG   2285
                    ||| ||||||||||||||||>>>>> 1251 >>>>>|||||||||||||
CL273Contig2   1067 GCCGAATGCCAAGCTCGGACAG.................GGCTATGGGATG   1100


    4CL_EUGR   2286 ACGGAGGCGGGCCCGGTGCTGGCAATGTGCCCGGCATTTGCAAAGGAGCC   2335
                    |||||||||||||||||||||||||||||| ||||||||||||||||||
CL273Contig2   1101 ACGGAGGCGGGCCCGGTGCTGGCAATGTGCCTGGCATTTGCAAAGGAGCC   1150

    4CL_EUGR   2336 GTTCGAGATCAAGTCAGGCGCATGCGGGACCGTCGTGAGGAACGCGGAGA   2385
                    |||||||||||||||||||| ||||||||||||||||||||||||||||||
CL273Contig2   1151 GTTCGAGATCAAGTCAGGCGCGTGCGGGACCGTCGTGAGGAACGCGGAGA   1200

    4CL_EUGR   2386 TGAAGATCGTCGACCCGGAGACAGGGGCCTCGCTCGCGCGGAACCAGGCC   2435
                    |||||||||||||||||||||||||||||||||| ||||||||||||||
CL273Contig2   1201 TGAAGATCGTCGACCCGGAGACAGGGGCCTCGCTCCCGCGGAACCAGGCC   1250

    4CL_EUGR   2436 GGCGAGATCTGCATCCGGGGTCACCAGATCATGAAAGgtacg......ta   2472
                    ||||||||||||||||||||||||||||||||||||>>>>> 1591 >>
CL273Contig2   1251 GGCGAGATCTGCATCCGGGGTCACCAGATCATGAAAG.............   1287

    4CL_EUGR   2472 aagGTTATCTGAACGACGCCGAGGCGACCGCAAATACCATAGACAAAGAA   4110
                    >>>|||||||||||||| |||| |||||||| ||||||||||||||||||
CL273Contig2   1287 ...GTTATCTGAACGACCCCGAAGCGACCGCTAATACCATAGACAAAGAA   1334
```

```
    4CL_EUGR   4111 GGGTGGCTGCACACCGGCGACATCGGCTACATAGACGATGACGACGAGCT   4160
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2   1335 GGGTGGCTGCACACCGGCGACATCGGCTACATAGACGATGACGACGAGCT   1384

    4CL_EUGR   4161 CTTCATTGTCGATCGGTTGAAGGAACTCATCAAGTACAAGGGCTTCCAGG   4210
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2   1385 CTTCATTGTCGATCGGTTGAAGGAACTCATCAAGTACAAGGGCTTCCAGG   1434

    4CL_EUGR   4211 TTGCTCCGGCCGAGCTAGAGGCAATGCTGATTGCACACCCAAGTATCTCG   4260
                    | ||||||||||||  ||||||||||||||||||||||||||||||||||
CL273Contig2   1435 TCGCTCCGGCCGAATTAGAGGCAATGCTGATTGCACACCCAAGTATCTCG   1484

    4CL_EUGR   4261 GATGCCGCTGTTGTGCCgtaag.....atcagGATGAAGGATGAGGTTGC   4403
                    ||||| || ||||||||>>>>> 108 >>>>>||||||||||||||||||
CL273Contig2   1485 GATGCTGCCGTTGTGCC..............GATGAAGGATGAGGTTGC   1519

    4CL_EUGR   4404 CGGTGAGGTTCCTGTTGCATTCGTGGTGAAATCCAATGGTTCCGTAATCA   4453
                    | ||||||||||||||||||||||||||||||||||||||||||||||||
CL273Contig2   1520 CAGTGAGGTTCCTGTTGCATTCGTGGTGAAATCCAATGGTTCCGTAATCA   1569

    4CL_EUGR   4454 CCGAGGACGAAATCAAGCAATACATCTCGAAGCAGGT   4490
                    | ||||||||||||||||||||||||||||||||||
CL273Contig2   1570 CTGAGGACGAAATCAAGCAATACATCTCGAAGCAGGT   1606

Alignment Score: 1411
```

APÊNDICE N

Predição do ATG inicial (“Start Codon”) para o contig 2 do cluster 273 realizado pelo programa *ATGpr*.

No. do ATG a partir da extremidade 5’	Frame	Início (bp)	Fim (bp)	Tamanho ORF	Stop codon?	Seqüência
3	3	336	1721	462	Yes	MLLLQNCPEFVFAFLGASYRGAISTTANPFYTPGEIAKQAS AAQAKIVITQAAYADKVRPFAEENGVKVVCIDTAPEGCLHF SELMQADENAAPAADVKPDDVLALPYSSGTTGLPKGVMLTH RGQVSSVAQQVDGDNPNLYFHKEDVILCTLPLFHIYSLNSV MFCALRVGAAILIMQKFEIVALMELVQRYRVTILPIVPPIV LAIAKSAEVDRYDLSSIRTIMSGAAPMGKELEDTVRAKLPN AKLGQGYGMTEAGPVLAMCLAFAKEPFEIKSGACGTVVRNA EMKIVDPETGASLPRNQAGEICIRGHQIMKGYLNDPEATAN TIDKEGWLHTGDIGYIDDDDELFIVDRLKELIKYKGFQVAP AELEAMLIAHPSISDAAVVPMKDEVASEVPVAFVVKSNGSV ITEDEIKQYISKQVVFYKRINRVFFTDAIPKAPSGKILRKD LRAKLASGVYN
7	3	828	1721	298	Yes	MFCALRVGAAILIMQKFEIVALMELVQRYRVTILPIVPPIV LAIAKSAEVDRYDLSSIRTIMSGAAPMGKELEDTVRAKLPN AKLGQGYGMTEAGPVLAMCLAFAKEPFEIKSGACGTVVRNA EMKIVDPETGASLPRNQAGEICIRGHQIMKGYLNDPEATAN TIDKEGWLHTGDIGYIDDDDELFIVDRLKELIKYKGFQVAP AELEAMLIAHPSISDAAVVPMKDEVASEVPVAFVVKSNGSV ITEDEIKQYISKQVVFYKRINRVFFTDAIPKAPSGKILRKD LRAKLASGVYN
6	3	693	1721	343	Yes	MLTHRGQVSSVAQQVDGDNPNLYFHKEDVILCTLPLFHIYS LNSVMFCALRVGAAILIMQKFEIVALMELVQRYRVTILPIV PPIVLAIAKSAEVDRYDLSSIRTIMSGAAPMGKELEDTVRA KLPNAKLGQGYGMTEAGPVLAMCLAFAKEPFEIKSGACGTV VRNAEMKIVDPETGASLPRNQAGEICIRGHQIMKGYLNDPE ATANTIDKEGWLHTGDIGYIDDDDELFIVDRLKELIKYKGF QVAPAELEAMLIAHPSISDAAVVPMKDEVASEVPVAFVVKS NGSVITEDEIKQYISKQVVFYKRINRVFFTDAIPKAPSGKI LRKDLRAKLASGVYN
8	3	867	1721	285	Yes	MQKFEIVALMELVQRYRVTILPIVPPIVLAIAKSAEVDRYD LSSIRTIMSGAAPMGKELEDTVRAKLPNAKLGQGYGMTEAG PVLAMCLAFAKEPFEIKSGACGTVVRNAEMKIVDPETGASL PRNQAGEICIRGHQIMKGYLNDPEATANTIDKEGWLHTGDI GYIDDDDELFIVDRLKELIKYKGFQVAPAELEAMLIAHPSI SDAAVVPMKDEVASEVPVAFVVKSNGSVITEDEIKQYISKQ VVFYKRINRVFFTDAIPKAPSGKILRKDLRAKLASGVYN
4	3	591	1721	377	Yes	MQADENAAPAADVKPDDVLALPYSSGTTGLPKGVMLTHRGQ VSSVAQQVDGDNPNLYFHKEDVILCTLPLFHIYSLNSVMFC ALRVGAAILIMQKFEIVALMELVQRYRVTILPIVPPIVLAI AKSAEVDRYDLSSIRTIMSGAAPMGKELEDTVRAKLPNAKL GQGYGMTEAGPVLAMCLAFAKEPFEIKSGACGTVVRNAEMK IVDPETGASLPRNQAGEICIRGHQIMKGYLNDPEATANTID KEGWLHTGDIGYIDDDDELFIVDRLKELIKYKGFQVAPAEL EAMLIAHPSISDAAVVPMKDEVASEVPVAFVVKSNGSVITE DEIKQYISKQVVFYKRINRVFFTDAIPKAPSGKILRKDLRA KLASGVYN

APÊNDICE O

Alinhamento das seqüências utilizadas para a análise de diversidade nucleotídica *in silico*. As seqüências utilizadas são constituintes do contig 2 do cluster 6 e todas abrangem uma idêntica região de 400 pb. Alinhamento realizado através do software *ClustalW*. Os círculos enfatizam os sítios polimórficos.

```
CLUSTAL W (1.82) multiple sequence alignment


CN-EUSP-FX-002-011_C10_.g_070.      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN-EUSP-FX-002-012_E09_.g_067.      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
RS-EUGR-XY-006-001_RS-EUGR-XY-      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_UESC_GO_AF_R_A07_.g_049.ab1      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_EUGL_XY_002_023_E09_.g_067.      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_GO_AF_R_B02_.g_013.ab1_CN-C      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_EUGL_XY_002_023_G06_.g_040.      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_GO_AF_R_B03_.g_025.ab1_CN-C      AT-CTAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
A10.esd_RS-010_PE-XY-001_genol      ---------GGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 41
CL6Contig2/93-491                   AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_UESC_GO_AF_R_G04_.g_024.ab1      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_EUGL_XY_002_014_F11_.g_091.      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
B06.ab1_MG-010_GR-PU-003_genol      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
RS-EUPE-XY-004-017_RS-EUPE-XY-      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
AF-EUGL-XY-001-059_C04_.g_086.      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_GO_AF_R_A10_.g_069.ab1_CN-C      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
RS-EUGR-PU-001-001_RS-EUGR-PU-      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_GO_AF_R_D09_.g_074.ab1_CN-C      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
CN_EUPE_XY_003_021_A12_.g_085.      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
D02.ab1_GO-094_GR-SE-001_genol      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
E06.esd_RS-049_GL-XY-001_genol      ATACAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 50
G12.ab1_MG-023_PE-XY-001_genol      AT-CAAGAAGGAGCTTAAGAAGCATCATCAATGGCAGCCAACGCAGAGCC 49
                                              ****************************************


CN-EUSP-FX-002-011_C10_.g_070.      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN-EUSP-FX-002-012_E09_.g_067.      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
RS-EUGR-XY-006-001_RS-EUGR-XY-      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_UESC_GO_AF_R_A07_.g_049.ab1      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_EUGL_XY_002_023_E09_.g_067.      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_GO_AF_R_B02_.g_013.ab1_CN-C      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_EUGL_XY_002_023_G06_.g_040.      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_GO_AF_R_B03_.g_025.ab1_CN-C      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
A10.esd_RS-010_PE-XY-001_genol      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 91
CL6Contig2/93-491                   TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_UESC_GO_AF_R_G04_.g_024.ab1      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_EUGL_XY_002_014_F11_.g_091.      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
B06.ab1_MG-010_GR-PU-003_genol      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
RS-EUPE-XY-004-017_RS-EUPE-XY-      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
AF-EUGL-XY-001-059_C04_.g_086.      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_GO_AF_R_A10_.g_069.ab1_CN-C      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
RS-EUGR-PU-001-001_RS-EUGR-PU-      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_GO_AF_R_D09_.g_074.ab1_CN-C      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
CN_EUPE_XY_003_021_A12_.g_085.      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
D02.ab1_GO-094_GR-SE-001_genol      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
E06.esd_RS-049_GL-XY-001_genol      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 100
G12.ab1_MG-023_PE-XY-001_genol      TCAGCAGACCCAACCAGCGAAGCATTCGGAAGTCGGCCACAAGAGCCTCT 99
                                    **************************************************


CN-EUSP-FX-002-011_C10_.g_070.      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN-EUSP-FX-002-012_E09_.g_067.      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
RS-EUGR-XY-006-001_RS-EUGR-XY-      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_UESC_GO_AF_R_A07_.g_049.ab1      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_EUGL_XY_002_023_E09_.g_067.      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_GO_AF_R_B02_.g_013.ab1_CN-C      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_EUGL_XY_002_023_G06_.g_040.      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_GO_AF_R_B03_.g_025.ab1_CN-C      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
A10.esd_RS-010_PE-XY-001_genol      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 141
CL6Contig2/93-491                   TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_UESC_GO_AF_R_G04_.g_024.ab1      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_EUGL_XY_002_014_F11_.g_091.      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
```

```
B06.ab1_MG-010_GR-PU-003_genol      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
RS-EUPE-XY-004-017_RS-EUPE-XY-      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
AF-EUGL-XY-001-059_C04_.g_086.      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_GO_AF_R_A10_.g_069.ab1_CN-C      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
RS-EUGR-PU-001-001_RS-EUGR-PU-      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_GO_AF_R_D09_.g_074.ab1_CN-C      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
CN_EUPE_XY_003_021_A12_.g_085.      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
D02.ab1_GO-094_GR-SE-001_genol      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
E06.esd_RS-049_GL-XY-001_genol      TGCAGAGCGATGCTCTCTACCAGTACATATAGGAGACCAGCGTCTACCCA 150
G12.ab1_MG-023_PE-XY-001_genol      TGCAGAGCGATGCTCTCTACCAGTACATATTGGAGACCAGCGTCTACCCA 149
                                    ****************************** ******************

CN-EUSP-FX-002-011_C10_.g_070.      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN-EUSP-FX-002-012_E09_.g_067.      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
RS-EUGR-XY-006-001_RS-EUGR-XY-      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_UESC_GO_AF_R_A07_.g_049.ab1      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_EUGL_XY_002_023_E09_.g_067.      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_GO_AF_R_B02_.g_013.ab1_CN-C      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_EUGL_XY_002_023_G06_.g_040.      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_GO_AF_R_B03_.g_025.ab1_CN-C      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
A10.esd_RS-010_PE-XY-001_genol      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 191
CL6Contig2/93-491                   AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_UESC_GO_AF_R_G04_.g_024.ab1      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_EUGL_XY_002_014_F11_.g_091.      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
B06.ab1_MG-010_GR-PU-003_genol      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
RS-EUPE-XY-004-017_RS-EUPE-XY-      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
AF-EUGL-XY-001-059_C04_.g_086.      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_GO_AF_R_A10_.g_069.ab1_CN-C      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
RS-EUGR-PU-001-001_RS-EUGR-PU-      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_GO_AF_R_D09_.g_074.ab1_CN-C      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
CN_EUPE_XY_003_021_A12_.g_085.      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
D02.ab1_GO-094_GR-SE-001_genol      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
E06.esd_RS-049_GL-XY-001_genol      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 200
G12.ab1_MG-023_PE-XY-001_genol      AGAGAGCCAGAGTCCATGAAGGAGCTCAGGGAAATAACAGCCAAACATCC 199
                                    ************ ************************************

CN-EUSP-FX-002-011_C10_.g_070.      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN-EUSP-FX-002-012_E09_.g_067.      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
RS-EUGR-XY-006-001_RS-EUGR-XY-      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_UESC_GO_AF_R_A07_.g_049.ab1      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_EUGL_XY_002_023_E09_.g_067.      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_GO_AF_R_B02_.g_013.ab1_CN-C      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_EUGL_XY_002_023_G06_.g_040.      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_GO_AF_R_B03_.g_025.ab1_CN-C      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
A10.esd_RS-010_PE-XY-001_genol      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 241
CL6Contig2/93-491                   ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_UESC_GO_AF_R_G04_.g_024.ab1      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_EUGL_XY_002_014_F11_.g_091.      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
B06.ab1_MG-010_GR-PU-003_genol      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
RS-EUPE-XY-004-017_RS-EUPE-XY-      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
AF-EUGL-XY-001-059_C04_.g_086.      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_GO_AF_R_A10_.g_069.ab1_CN-C      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
RS-EUGR-PU-001-001_RS-EUGR-PU-      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_GO_AF_R_D09_.g_074.ab1_CN-C      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
CN_EUPE_XY_003_021_A12_.g_085.      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
D02.ab1_GO-094_GR-SE-001_genol      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
E06.esd_RS-049_GL-XY-001_genol      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 250
G12.ab1_MG-023_PE-XY-001_genol      ATGGAACCTGATGACCACATCGGCTGATGAAGGGCAGTTCCTGAACATGC 249
                                    ************************************************

CN-EUSP-FX-002-011_C10_.g_070.      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN-EUSP-FX-002-012_E09_.g_067.      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
RS-EUGR-XY-006-001_RS-EUGR-XY-      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_UESC_GO_AF_R_A07_.g_049.ab1      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_EUGL_XY_002_023_E09_.g_067.      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_GO_AF_R_B02_.g_013.ab1_CN-C      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_EUGL_XY_002_023_G06_.g_040.      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_GO_AF_R_B03_.g_025.ab1_CN-C      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
A10.esd_RS-010_PE-XY-001_genol      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 291
CL6Contig2/93-491                   TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_UESC_GO_AF_R_G04_.g_024.ab1      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_EUGL_XY_002_014_F11_.g_091.      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
B06.ab1_MG-010_GR-PU-003_genol      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
RS-EUPE-XY-004-017_RS-EUPE-XY-      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
AF-EUGL-XY-001-059_C04_.g_086.      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_GO_AF_R_A10_.g_069.ab1_CN-C      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
```

```
RS-EUGR-PU-001-001_RS-EUGR-PU-      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_GO_AF_R_D09_.g_074.ab1_CN-C      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
CN_EUPE_XY_003_021_A12_.g_085.      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
D02.ab1_GO-094_GR-SE-001_genol      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
E06.esd_RS-049_GL-XY-001_genol      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 300
G12.ab1_MG-023_PE-XY-001_genol      TCCTCAAGCTCATCAACGCCAAGAACACCATGGAGATCGGCGTCTACACC 299
                                    ************************************************

CN-EUSP-FX-002-011_C10_.g_070.      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCTGATGACGGAAAGAT 349
CN-EUSP-FX-002-012_E09_.g_067.      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCTGATGACGGAAAGAT 349
RS-EUGR-XY-006-001_RS-EUGR-XY-      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCTGATGACGGAAAGAT 349
CN_UESC_GO_AF_R_A07_.g_049.ab1      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCTGATGACGGAAAGAT 349
CN_EUGL_XY_002_023_E09_.g_067.      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
CN_GO_AF_R_B02_.g_013.ab1_CN-C      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
CN_EUGL_XY_002_023_G06_.g_040.      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
CN_GO_AF_R_B03_.g_025.ab1_CN-C      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
A10.esd_RS-010_PE-XY-001_genol      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 341
CL6Contig2/93-491                   GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
CN_UESC_GO_AF_R_G04_.g_024.ab1      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
CN_EUGL_XY_002_014_F11_.g_091.      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
B06.ab1_MG-010_GR-PU-003_genol      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
RS-EUPE-XY-004-017_RS-EUPE-XY-      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
AF-EUGL-XY-001-059_C04_.g_086.      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
CN_GO_AF_R_A10_.g_069.ab1_CN-C      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
RS-EUGR-PU-001-001_RS-EUGR-PU-      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
CN_GO_AF_R_D09_.g_074.ab1_CN-C      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
CN_EUPE_XY_003_021_A12_.g_085.      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
D02.ab1_GO-094_GR-SE-001_genol      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGANAGAT 349
E06.esd_RS-049_GL-XY-001_genol      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 350
G12.ab1_MG-023_PE-XY-001_genol      GGCTACTCTCTCCTCGCCACCGCCCTTGCTCTTCCCGATGACGGAAAGAT 349
                                    ********************************** ********* ****

CN-EUSP-FX-002-011_C10_.g_070.      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN-EUSP-FX-002-012_E09_.g_067.      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
RS-EUGR-XY-006-001_RS-EUGR-XY-      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_UESC_GO_AF_R_A07_.g_049.ab1      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_EUGL_XY_002_023_E09_.g_067.      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_GO_AF_R_B02_.g_013.ab1_CN-C      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_EUGL_XY_002_023_G06_.g_040.      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_GO_AF_R_B03_.g_025.ab1_CN-C      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
A10.esd_RS-010_PE-XY-001_genol      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 391
CL6Contig2/93-491                   CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_UESC_GO_AF_R_G04_.g_024.ab1      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_EUGL_XY_002_014_F11_.g_091.      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
B06.ab1_MG-010_GR-PU-003_genol      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
RS-EUPE-XY-004-017_RS-EUPE-XY-      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
AF-EUGL-XY-001-059_C04_.g_086.      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_GO_AF_R_A10_.g_069.ab1_CN-C      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
RS-EUGR-PU-001-001_RS-EUGR-PU-      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_GO_AF_R_D09_.g_074.ab1_CN-C      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
CN_EUPE_XY_003_021_A12_.g_085.      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
D02.ab1_GO-094_GR-SE-001_genol      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
E06.esd_RS-049_GL-XY-001_genol      CTTGGCCATGGACATCAATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 400
G12.ab1_MG-023_PE-XY-001_genol      CTTGGCCATGGACATCNCATAGGGAGAACTTCGAGATCGGGCTGCCCGTCA 399
                                    ********** ***************************************
```

APÊNDICE P

“Draft” do artigo científico a ser submetido à revista TAG.

Nucleotide Sequence of a BAC DNA encoding 4-Coumarate Coenzyme A Ligase from *Eucalyptus grandis*

Neiva, S.[1]; Pappas Junior, G.[1]; Brommonschenkel, S.[2]; Grattapaglia, D.[1, 3]

[1] Graduate Program in Genomic Sciences and Biotechnology, Universidade Catolica de Brasília, Campus II - SGAN 916 modulo B, Brasília 70790-160 DF, Brazil.

[2] Universidade Federal de Lavras - UFV

[3]Plant Genetics Laboratory, Embrapa - Recursos Genéticos e Biotecnologia, Parque Estação Biológica, Brasília 70770-970 DF, Brazil.

* Corresponding author:

Dario Grattapaglia

Plant Genetics Laboratory, Embrapa - Recursos Genéticos e Biotecnologia

Parque Estação Biológica, Brasília 70770-970 DF, Brazil

Phones: office: +55-61-4484652 mobile: 55-61-99712142

Fax: +55-61-3403624

e-mail: dario@cenargen.embrapa.br

Resumo

A lignina é um importante composto fenólico constituinte da madeira. A capacidade de sintetizar lignina foi uma essencial adaptação evolucionária das plantas do meio aquático para o terrestre A via de biossíntese dos precursores da lignina é a mais bem conhecida no processo de formação da madeira e tem sido foco de estudos nas suas várias enzimas. A 4CL (4-Coumarate:CoA ligase; EC 6.2.1.12) tem um importante papel na via de biossíntese de lignina e compostos secundários, mediando o último passo do metabolismo geral. A construção de uma biblioteca de BAC pelo Projeto Genolyptus possibilitou, através de uma triagem via PCR por clones que possuíssem o gene alvo, a identificação de um clone, a partir do qual uma biblioteca shotgun foi construída. O seqüenciamento e montagem do clone BAC resultaram na obtenção da seqüência completa da região codante do gene 4cl com 5.203 pb. A obtenção desta seqüência com o início de transcrição, inédita para Eucalyptus, abre possibilidades interessantes de estudos detalhados da diversidade nucleotídica e padrões de DL ao longo deste gene em populações de clones fenotipados, no sentido de buscar associações entre haplótipos específicos e variação quantitativa em propriedades químicas da madeira.

Palavras chave

Eucalyptus, diversidade nucleotídica, CCoAOMT, 4CL, BAC.

Introdução

A lignina é, depois da celulose, o segundo mais abundante biopolímero terrestre, agregando aproximadamente 30% do carbono orgânico da biosfera. A capacidade de sintetizar lignina foi uma essencial adaptação evolucionária das plantas do meio aquático para o terrestre (Boerjan et al., 2003; Nicholson & Hammerschmidt, 1992). Crucial pra a integridade estrutural da parede celular, a lignina permite a rigidez e condição ereta da planta (Chabannes et al., 2001; Jones, 2001). Além disso, fornece à parede celular a capacidade de resistência à água, permite o transporte de solutos através do sistema vascular, impede a perda de água excessiva por transpiração e atua na proteção da planta contra o ataque de patógenos (Humphreys , 1999).

A via de biossíntese dos precursores da lignina é a mais bem conhecida no processo de formação da madeira e tem sido o foco da experimentação em biologia molecular florestal. A maioria dos genes que codificam para as enzimas desta via, bem como fatores de transcrição e proteínas de parede, foram clonados e caracterizados particularmente em Pinus taeda (Allona et al., 1998), para as espécies arbóreas.

A enzima 4-coumarate:CoA ligase faz parte da via de formação da lignina mediando o último passo do metabolismo geral de fenilpropanóides (Lee et al., 1997; Hu et al., 1998). A enzima 4CL é membro de uma superfamília de enzimas formadoras de adenilatos que compartilham mecanismos de reação comum com a formação de um intermediário do substrato adenilato na presença de ATP e Magnésio, seguido de uma esterificação com Coenzima A (CoA ligase), 4´-fosfopanteteíne ou oxidação por molécula de oxigênio. Os produtos da reação, ésteres de hidroxicinamoil CoA, servem como substratos para a via específica de formação de fenilpropanóides. Possui EC 6.2.1.12 que corresponde a uma ligase, formando ligações carbono-enxofre, uma ligase ácido-tiol, com reação principal: ATP + 4-coumarate + CoA = AMP + diphosphate + 4-coumaroyl-CoA.

A grande importância dos fenilpropanóides para as plantas diz respeito à proteção contra UV que lhe é conferida e a formação de lignina. Os compostos estruturais de ubiquitina no xilema de plantas

vasculares sugerem que a evolução da via de fenilpropanóides foi um passo crucial para o processo de colonização da terra pelas plantas terrestres.

O *Eucalyptus*, da família das *Myrtaceae*, árvore originária da Austrália de fácil adaptação em várias condições climáticas de fácil adaptação em várias outras condições climáticas. O eucalipto possui um genoma de aproximadamente 630 Mb com um conjunto haplóide de 11 cromossomos e o gênero é conhecido por sua grande variabilidade genética. Esta decorre do seu hábito alógamo e da resposta à pressão de seleção causada pelas alterações do meio ambiente.

Atualmente, a grande utilização do eucalipto visa a indústria de polpa e papel, mas não dispensa a sua grande utilidade na construção de casas, produção de óleos essenciais e em projetos de reflorestamento e preservação de florestas nativas.

Um projeto visando o estudo completo do gênero *Eucalytus*, o Projeto Genolyptus, utilizou como ferramenta necessária para o estudo, a construção de uma biblioteca de BAC do eucalipto. O *Eucalyptus grandis* foi a espécie escolhida e o sistema de clonagem em BAC foi selecionado por permitir a clonagem estável de grandes fragmentos de DNA e por serem facilmente manipulados. A partir de um clone positivo para o gene *4cl*, uma biblioteca *shotgun* do clone foi construída e seqüenciada.

No presente trabalho, nós clonamos e analisamos uma cópia nuclear do gene *4cl* de *Eucalyptus grandis* e determinamos sua seqüência nucleotídica, ainda não descrita para o gênero.

Material e Métodos

Biblioteca de BAC No âmbito do projeto Genolyptus foi construída uma biblioteca genômica de BAC de um indivíduo da espécie Eucalyptus grandis (S. Brommonschenkel com. pess.). Essa biblioteca composta por 20160 clones com tamanho médio de 120 a 150 kb fornece uma cobertura estimada de 4 vezes do genoma do eucalipto. Esta biblioteca foi submetida a uma triagem via PCR com o objetivo de

identificar um ou mais clones BAC contendo o gene alvo para posteriores estudos da estrutura gênica completa, incluindo promotor, região codante e introns. Os clones positivos apresentaram fragmento de tamanho aproximado de 600pb, amplificado pela utilização de um par de primers descritos por Gion *et al*. (2000) para o gene *4cl*.

O isolamento do DNA de BAC foi realizado pela técnica de lise alcalina (Sambrook et al., 1989).

Triagem da biblioteca de BAC Uma estratégia hierárquica de "pools" de clones BAC foi utilizada para rapidamente alcançar o gene alvo. Os 20.160 clones BAC foram organizados em 210 microplacas de 96 poços. Mini-preparações de DNA de BACs foram realizadas em "pools" de 96 clones de uma mesma placa. As extrações de DNA dos clones BAC foram realizadas para cada "pool" de 96 clones resultando em 210 pools, os quais foram organizados em grupos de 6 pools, formando um superpool. Ao total, foram obtidos 35 superpools (Figura1).

Reações de PCR para a triagem foram realizadas a partir do superpool. Aquele que fosse positivo, era desmembrado em pools, pool e placa, até chegar ao clone. A reação de PCR foi realizada em um volume total de 25 µl contendo 2µl do DNA do clone BAC, 1x de tampão pht 10X IIB (Phoneutria – pht), 0,4 µM de cada um dos primers, 0,4 mM de cada dNTPs, 0,1 mg.ml-1 de BSA, 2 mM de MgCl2 e 2,5U de Taq DNA polimerase (Pheneutria – pht). As condições para a reação foram as seguintes: 95° C por 4 min, 30 ciclos de 92° C por 45s, 65° C por 45s, 72° C por 1min30s, extensão final de 72° C por 10 min e incubação a 12° C por tempo indeterminado. Os produtos da reação foram analisados por eletroforese em gel de agarose 1%, corado com brometo de etídeo e comparados com marcador molecular 1 Kb plus DNA Ladder (Invitrogen, Carlsbad, CA – EUA) (Figura 2).

Biblioteca *Shotgun* de BAC Uma biblioteca genômica *shotgun* do clone BAC, selecionado por conter o gene alvo para a enzima 4CL, foi produzida na UFV pelo Prof. Sergio Brommonschenkel e sua equipe. Foram gerados 960 subclones do BAC os quais continham tamanho médio de 1 kb. Os clones foram submetidos à extração de DNA por lise alcalina (Sambrook et al., 1989) e seqüenciados por ambas as extremidades 3' e 5' com o kit Big Dye™ sob a seguinte reação de volume total de 10 µl: 1 a

3 µl de DNA, 1 µl de primer a 3,2 µM (T3 e T7), 2 µl de Tampão 2,5x (200 mM Tris-HCl pH 9,0; 5mM MgCl2), 2 µl de Big Dye™. As amplificações foram realizadas em termociclador sob o seguinte programa: 96°C por 2min, 25 ciclos de 96° C por 45s, 50° C por 30s e 60° C por 3min. Após os 25 ciclos, incubar a 60° C por 4min e a 4° C por 2min.

Após a reação de seqüenciamento, os produtos foram purificados, desnaturados e submetidos à eletroforese em seqüenciador 3700 da Applied Biosystems.

Para o desenvolvimento dos contigs, as seqüências dos clones foram analisadas quanto à qualidade de bases e alinhadas para a formação de um contig único, por meio do programa CAP3 (Huang & Madan, 1999). Comparações com seqüências de nucleotídeos e aminoácidos foram realizadas por meio do programa ClustalW (Higgins et al., 1994) e a inferência do codon inicial pelo programa ATGprediction (Salamov et al., 1998).

Resultados e Discussão

A biblioteca de BAC construída para a espécie *Eucalyptus grandis* teve como grande objetivo a obtenção de seqüências completas da região gênica, incluindo promotor, região 5'UTR e 3'UTR, a fim de comparações quanto às variações nucleotídicas e associações fenotípicas da espécie. As buscas por clones que apresentassem homologia com o gene *4cl* por meio de amplificações com iniciadores específicos geraram como resultado apenas um clone, P164H12 (Figura 2). Este clone foi submetido à construção de uma biblioteca *shotgun* para facilitar o seqüenciamento completo do clone, visto que os fragmentos obtidos variavam em torno de 1 kb, contra 150 kb, observado no clone BAC.

As seqüências resultantes do seqüenciamento dos clones da biblioteca *shotgun* foram avaliadas quanto à qualidade e alinhadas para montar um contig único, contendo a seqüência completa do gene *4cl*. A seqüência montada constitui um segmento contínuo de 5203 pb, com 4 regiões de exons, 3 introns, um sítio putativo de poliadenilação e uma pequena região 3'UTR. As inferências quanto às

regiões foram realizadas por comparação com banco de dados de *Arabidopsis* e com as ESTs do banco de dados do Projeto Genolyptus, para a inferência correta das posições de início e término de cada uma das regiões componentes da seqüência gênica de *4cl*. A seqüência foi graficamente visualizada com o auxilio da montagem pelo programa Artemis (Rutherford et al., 2000) (Figura 3).

Outra inferência, com relação ao códon inicial de inicio da ORF, foi gerada pelo programa ATGpr, indicando a posição número 90 para o ATG inicial da seqüência de 4CL obtida. Alinhamentos comparativos da seqüência de aminoácidos gerada a partir da tradução da ORF com inicio na posição 90 constataram que a metionina inicial foi corretamente inferida. Comparações com a seqüência protéica de Populus (gi|2911799|) confirmaram o início da região codante do segmento do gene *4cl* de *Eucalyptus*. Desta forma pode-se sugerir que a seqüência codante para a enzima 4CL encontra-se completa, visto que a espécie *Populus* é passível de comparação com *Eucalyptus*. Vale ressaltar ainda que a inferência sobre a posição do ATG inicial é aceitável visto que a seqüência analisada de *Populus* é uma seqüência validada, ou seja, confirmada por outros seqüenciamentos e/ou por experimentos. Desta maneira, por comparações com seqüências do banco de dados de domínio público, bem como com o banco de dados do Projeto Genolyptus, obtivemos a seqüência codante completa do gene *4cl* para a espécie *Eucalyptus grandis*, não descrita ainda em literatura nem disponível em banco de dados.

A obtenção da seqüência codante completa de *4cl* deixa portas abertas para novas estratégias de estudo do gene como desenho de iniciadores específicos para *Eucalyptus* para amplificação da região gênica de *4cl* em outras espécies de eucalipto, bem como a possibilidade de obtenção, por “primer walking”, da região promotora, também de fundamental importância para estudos de associação, já que essa região, ao se encontrar polimórfica em várias espécies, é o determinante para a variação de expressão gênica. Estudos detalhados da diversidade nucleotídica e padrões de DL ao longo deste gene em populações de clones fenotipados poderão ser realizados, no sentido de buscar associações entre haplótipos específicos e variação quantitativa em propriedades químicas da madeira.

Legenda das figuras

Figura 1. Formação dos pools e superpools para a triagem da biblioteca para os genes *ccoaomt* e *4cl*.

Figura 2. Triagem de uma biblioteca de BAC. Gel de agarose dos produtos amplificados com iniciadores para 4CL nos 35 superpools de BAC. A-H, superpools; M, marcador 1 Kb Plus DNA ladder (Invitrogen, Carlsbad, CA).

Figura 3. Ilustração esquemática do segmento do gene *4cl* montado a partir de seqüências oriundas da biblioteca *shotgun* do clone BAC, selecionado via triagem por PCR de uma biblioteca BAC. As regiões em amarelo, cinza escuro e cinza claro correspondem, respectivamente, às regiões de exon, intron e 3'UTR.

Agradecimentos

Ao Dr. Sergio Brommonschenkel e sua equipe pela construção das bibliotecas de BAC e *shotgun* do clone de BAC. Este trabalho foi financiado pelo Projeto Genolyptus e empresas colaboradoras, Dell computadores e CNPq.

Referências

Allona, I.; Quinn, M.; Shoop, E.; et al. Analysis of xylem formation in pine by cDNA sequencing. Proc. Natl. Acad. Sci. USA, v. 95, p. 9693-9698, 1998.

Boerjan, W.; Ralph, J.; Baucher, M. Lignin biosynthesis. Annual Review of Plant Biology, n. 54, p. 519–546, 2003.

Chabannes, M.; Ruel, K.; Yoshinaga, A. et al. In situ analysis of lignins in transgenic tobacco reveals a differential impact of individual transformations on the spatial patterns of lignin deposition at the cellular and subcellular levels. The Plant Journal, v. 28, p. 271–82, 2001.

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

Higgins, D.; Thompson, J.; Gibson, T.; Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., v. 22, p. 4673-4680, 1994.

Hu, W-J. ; Kawaoka, A.; Tsai, C-J.; et al. Compartmentalized expression of two structurally and functionally distinct 4-coumarate: CoA ligase genes in aspen Populus tremuloides. Proc. Natl. Acad. Sci. USA, v. 95 (9), p. 5407–5412, 1998.

Huang, X.; Madan, A. CAP3: A DNA sequence assembly program. Genome Research, v. 9, p. 868-877,

1999

Humphreys, J. M.; Hemm, M. R.; Chapple, C. New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. Proc. Natl. Acad. Sci. USA, v. 96, p. 10045–10050, 1999.

Jones, L.; Ennos, A. R.; Turner, S. R. Cloning and characterization of irregular xylem4 (irx4): a severely lignin-deficient mutant of Arabidopsis. Plant Journal, v. 26(2), p. 205-216, 2001.

Lee, D.; Meyer, K.; Chapple, C.; et al. Antisense suppression of 4- coumarate: coenzyme A ligase activity in Arabidopsis leads to altered lignin subunit composition. The Plant Cell, v. 9, p. 1985-1998, 1997.

Nicholson, R. L.; Hammerschmidt, R. Phenolic compounds and their role in disease resistance. Annual Review of Phytopathology, v. 30, p. 369–389, 1992.

Rutherford, K.; Parkhill, J.; Crook, J. Horsnell, T. et al. Artemis: sequence visualization and annotation. Bioinformatics, v. 16 (10), p. 944-945, 2000.

Salamov, A. A.; Nishikawa, T.; Swindells, M. B. Assessing protein coding region integrity in cDNA sequencing projects. Bioinformatics, v. 14, p. 384-390, 1998.

Sambrook, J.; Fritsch, E. F.; Maniatis, T. Molecular Cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, USA. 1989 2nd edição.

Figura 1

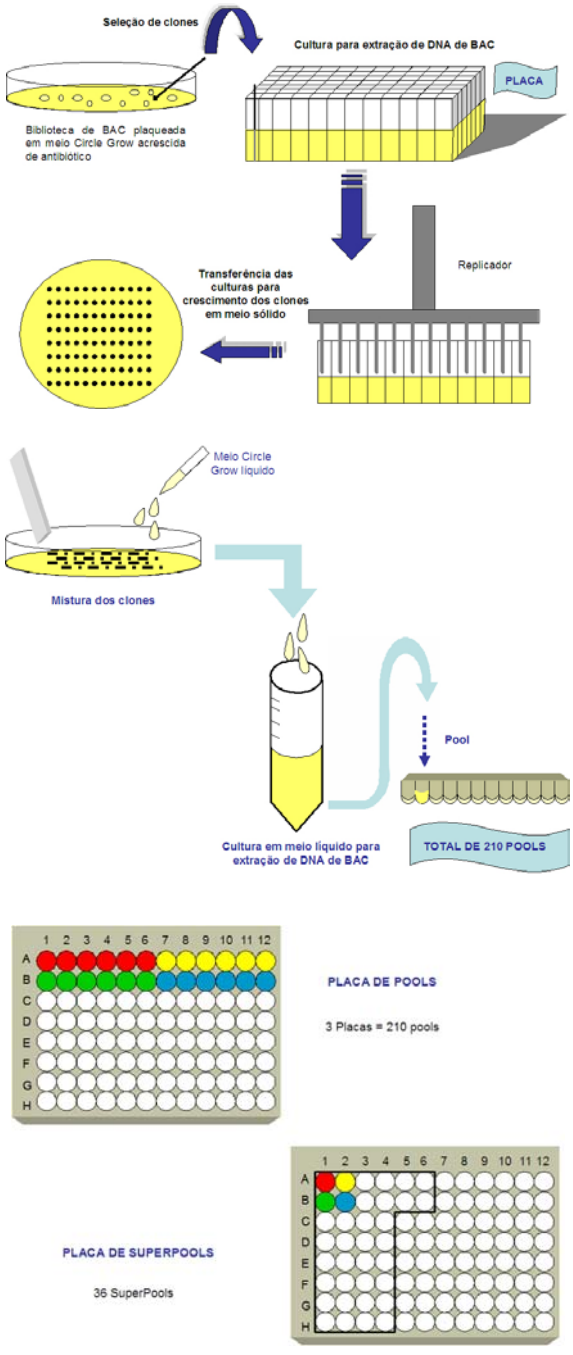

Seleção de clones
Cultura para extração de DNA de BAC
PLACA
Biblioteca de BAC plaqueada em meio Circle Grow acrescida de antibiótico
Replicador
Transferência das culturas para crescimento dos clones em meio sólido
Meio Circle Grow líquido
Mistura dos clones
Pool
Cultura em meio líquido para extração de DNA de BAC
TOTAL DE 210 POOLS
1 2 3 4 5 6 7 8 9 10 11 12
A B C D E F G H
PLACA DE POOLS
3 Placas = 210 pools
PLACA DE SUPERPOOLS
36 SuperPools

Figura 2

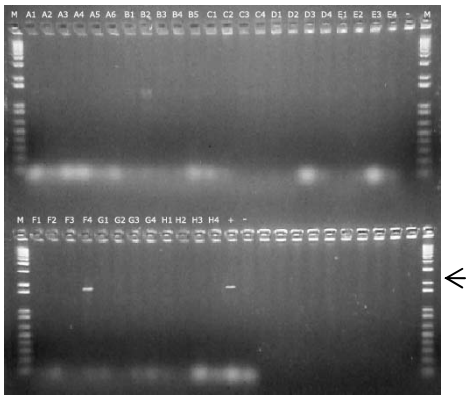
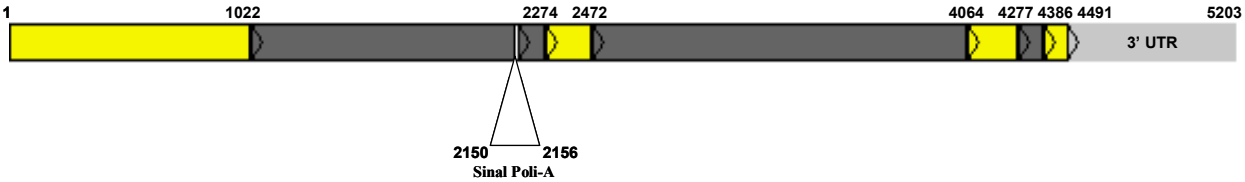

M A1 A2 A3 A4 A5 A6 B1 B2 B3 B4 B5 C1 C2 C3 C4 D1 D2 D3 D4 E1 E2 E3 E4 - M
M F1 F2 F3 F4 G1 G2 G3 G4 H1 H2 H3 H4 + -

Figura 3


1
1022
2274
2472
4064
4277
4386
4491
5203
3' UTR
2150
2156
Sinal Poli-A