UNIVERSIDADE CATÓLICA DE BRASÍLIA

Alexandre Marques Póvoa

FREQÜÊNCIA DE SNPS, ESTRUTURA DE HAPLÓTIPOS E DESEQUILÍBRIO DE LIGAÇÃO PARA OS GENES *CAD2* E *COMT2* DA VIA DE LIGNIFICAÇÃO EM *Eucalyptus*

Dissertação apresentada ao Programa de Pós-Graduação "*Stricto Sensu*" em Ciências Genômicas e Biotecnologia para obtenção do Título de Mestre em Ciências Genômicas e Biotecnologia.

Orientador: Prof. Dr. Georgios J. Pappas Jr
Co-orientador: Prof. Dr. Dario Grattapaglia

Brasília
2005

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. Georgios J. Pappas Jr

Dr. Dario Grattapaglia

Dr. Rinaldo Wellerson Pereira

Dr. Alexandre Coelho

Brasília
UCB

“Quanto mais aumenta nosso conhecimento, mais evidente fica nossa ignorância”

John Kennedy

Aos meus pais Ernani e Maria de Lourdes, e irmãos Fernando e Marcelo,

Pela educação e incentivo durante toda minha vida

Dedico

AGRADECIMENTOS

Ao Prof. Dr. Georgios J. Pappas Jr, pelo apoio muito importante, pela ajuda e oportunidade dada para realização deste trabalho.

Ao Prof. Dr. Dario Grattapaglia, pela co-orientação, dicas e estratégias formuladas.

Ao Prof. Dr. Rinaldo Wellerson Pereira, pelas sugestões durante o desenvolvimento de presente trabalho e fundamental ajuda na análise dos dados.

Ao Prof. Dr. Sérgio Brommonschenkel pela preparação dos clones de BAC.

À pesquisadora Marília Pappas por sua grande contribuição tanto de bancada quanto de solução de problemas.

À Eva, pelos inúmeros momentos de colaboração.

Às empresas Suzano Celulose e Papel, Klabin Floresta, Aracruz Celulose, pelo fornecimento do material biológico.

A todos os professores do Programa de Pós-Graduação "*Stricto Sensu*" em Ciências Genômicas e Biotecnologia / UCB, pela importante colaboração em meu aperfeiçoamento profissional.

Aos amigos da pós-graduação da UCB, em especial à Camila, Sandra, Andrea, Elisa, e todos os outros que estiveram comigo nesta luta.

A todos os funcionários do Programa de Pós-Graduação "*Stricto Sensu*" em Ciências Genômicas e Biotecnologia / UCB, pela amizade, atenção e presteza no auxílio, especialmente Francisco Fábio Gomes da Costa, Alessandra Maria Moreira Reis e André Luiz Cardoso Ramos.

A CAPES e DELL, pela concessão das bolsas de mestrado.

À minha família pelo apoio e incentivo.

À minha namorada Nathália Bueno pelo incentivo, compreensão e fundamental apoio.

A todos aqueles que colaboraram direta ou indiretamente e foram muito importantes para a realização deste trabalho.

RESUMO

No Brasil, a produção do eucalipto movimenta bilhões de reais por ano. A madeira de eucalipto tem seu principal uso na indústria de papel e celulose, onde o processo de polpagem constitui-se uma etapa bastante onerosa para separar a celulose dos outros componentes da madeira. Dentre estes componentes da madeira, encontra-se a lignina, um heteropolímero fenólico composto majoritariamente, em angiospermas, pelos monômeros guaiacil e siringil. Quanto maior a proporção de siringil, mais interessante é para a indústria papeleira, isto porque facilita no processo de polpagem, diminuindo a poluição, e os custos envolvidos. Desta forma, a via de formação desse polímero passou a ser alvo de estudos de forma a identificar genes correlacionados à qualidade da madeira. A partir destes genes candidatos, foram iniciados estudos de forma a obter sua seqüência completa e avaliar sua diversidade nucleotídica, de forma a procurar as bases moleculares da qualidade da madeira. Estes estudos foram realizados para os genes cinamil-álcool desidrogenase (*CAD2*) e ácido cafeico 3-O-metiltransferase (*COMT2*). Primeiramente foi realizada uma triagem de clones de cromossomos artificiais de bactéria (BAC) de forma a encontrar quais clones continham o gene de interesse. Com os clones identificados, somente para o gene *CAD2*, foi realizado um "shotgun" objetivando a obtenção de sua inédita seqüência completa. Foi obtida a seqüência completa para o gene *CAD2* em *E. grandis*, além de suas regiões regulatórias. Estudos de seqüência deste gene demonstraram bastante conservação dentre as espécies de *Eucalyptus*, além de uma expressão preferencial no xilema. Foi também realizada uma análise de diversidade nucleotídica para regiões dos genes *CAD2* e *COMT2*, sendo observada uma grande conservação nos indivíduos de *E. globulus*, tanto para o gene *CAD2* quanto para *COMT2*, assim como variados níveis de desequilíbrio de ligação observados em quase todas análises. Pouca diversidade foi encontrada dentro de cada espécie para os dois genes estudados, enquanto que entre as mesmas esta diversidade demonstrou-se bem maior. Ocorreu uma distinção clara entre os indivíduos de *E. grandis* estudados, fato que pode ser explicado pela variação de microclimas e pela dispersão. Novos estudos devem ser realizados de forma a analisar uma região mais extensa destes genes, além do controle de expressão dos genes da via de lignificação.

Palavras-chave: *Eucalyptus*, lignina, *CAD* e *COMT*.

ABSTRACT

The eucalypt production circulate billion of reais per year. The eucalypt wood has it's principal use in the paper and cellulose industry, where the pulping process is one of the most expensive step, with the finality of separate the cellulose of the other components of wood. One of this components is the lignin, a fenolic heteropolymer compound, in angiosperms, mostly by the monomers guaiacyl e syringyl. Much greather is the ratio of syringyl, more interesting is for paper industry because facilitates in the pulping process, lowering the pollution and the costs. This way, the monolignols pathway turned to be the focus of studies to identify candidate genes correlated to wood quality. With this candidate genes, were initiated studies to obtain their complete sequence and evaluate their nucleotide diversity to find the molecular basis for the wood quality. These studies were realized for the genes cinnamyl alcohol dehydrogenase 2 (*CAD2*) and caffeic acid 3-O-methyltransferase (*COMT2*). First, was realized a screening of bacterial artificial chromosome (BAC) library to find the clones that are carrying the target genes. After this identification, was realized a shotgun only for *CAD2* to obtain it's complete sequence. Were obtained the complete sequence for the gene *CAD2* in *E. grandis*, including it regulatory regions. Sequence studies of these genes demonstrated a high conservation between the eucalypt species, beyond a preferencial expression in xilem. Also was realized a nucleotide diversity analysis for regions of the genes *CAD2* e *COMT2*, being observed a high conservation in *E. globulus* individuals for the two genes, as different levels of linkage disequilibrium observed on all analysis. Just a few diversity was observed inside each species for the two studied genes, but between the species the diversity was higher. A courious observation was the distinction between the two groups of *E. grandis* studied, fact that can be explained by the microclimate variation and dispersion. New studies must be realized no analyze a greater region of these genes and the expression control of the lignification genes.

Keywords: *Eucalyptus*, lignin, *CAD* and *COMT*.

Índice Remissivo

Índice de Tabelas

Índice de Figuras

Lista de Abreviações

BAC – "Bacterial Artificial Chromosomes"

BLAST – "Basic Local Alignment Search Tool"

CAD – Cinamil álcool desidrogenase

CCR - Cinamoil CoA Redutase

cDNA – DNA complementar

COMT – Cafeico ácido-3-O-metiltransferase

D – coeficiente de desequilíbrio

D' – coeficiente de desequilíbrio normalizado

DNA – Ácido desoxirribonucléico

dNTP – desoxirribonucleotídeos

EC – "Enzyme Commission number"

ESTs – "Expressed Sequence Tag"

FAO – "Food and Agriculture Organization"

G – Guaiacil

GL – Grupo de E. globulus procedente de Victoria

GR – Grupo de E. grandis procedente de Atherton

GRR – Grupo de E. grandis procedente de Pine Creek

kb – Quilobase

DL – Desequilíbrio de ligação

M – Molar

ml – Mililitro

mM – Milimolar

mRNA – RNA mensageiro

NADP+ – Nicotinamida adenina dinucleotídeo fosfato

NADPH – Nicotinamida adenine dinucleotídeo fosfato reduzido

NCBI – “National Center for Biotechnology Information”

ng – Nanograma

pb – Pares de bases

PCR – “Polymerase Chain Reaction”

QTL – “Quantitative Trait Locus”

r^2 – Quadrado do coeficiente de correlação

RNA – Ácido ribonucléico

RNAse – Ribonuclease

S – Siringil

SNP – “Single Nucleotide Polymorphism”

UR – Grupo de E. urophylla procedente de Timor

µL – Microlitros

µM – Micromolar

X2 – Qui-quadrado

1 INTRODUÇÃO

1.1 O Eucalipto e a indústria da madeira

Por diversos séculos a madeira representa um dos insumos principais utilizados pela civilização humana, muito em função de se constituir no recurso biológico mais abundante do planeta. Verifica-se o emprego da madeira para os mais variados fins, como na construção civil, mobiliário, geração de energia, produção de aço e destacadamente a produção de papel e celulose. Interessantemente, o papel, apesar de ter sua origem datada ao redor do ano 100 DC na China, passou a ter a madeira como principal matéria-prima somente em meados do século XIX, pois originalmente era produzido a partir de outras fontes de fibra, como o cânhamo e tecidos.

Até poucos anos, o suprimento de madeira era proveniente da extração indiscriminada de florestas nativas, o que acarretou em danos irreversíveis a diversos ecossistemas. Atualmente grandes indústrias valem-se de extensivos projetos de reflorestamento para obtenção de madeira, o que garante uma maior produtividade e homogeneidade, além de resultar em significativos ganhos de escala em relação ao cenário extrativista. Inicialmente, o pinheiro (*Pinus sp.*) e o abeto (*Abies sp.*) foram utilizados em plantações comerciais, seguidos por árvores dos gêneros *Populus* e *Eucalyptus*, dentre outras.

Atualmente verifica-se uma ampla utilização de diversas espécies do gênero *Eucalyptus* em regiões de clima tropical e sub-tropical. Este fato justifica-se devido às suas características de rápido crescimento, produtividade, ampla diversidade de espécies, grande capacidade de adaptação e por ter aplicação para as mais diversas

finalidades. No Brasil, cerca de 66% das plantações florestais são de eucalipto (MOURA e GARCIA, 2000).

Em termos nacionais estima-se uma área total de 4,6 milhões de hectares de florestas plantadas, sendo 1,7 milhões com o gênero *Pinus* e 2,9 milhões do gênero *Eucalyptus*. Devido a sua robustez, adaptabilidade ao clima e solo brasileiros e ao seu curto ciclo de vida – cerca de sete anos – o eucalipto passou a ser a espécie florestal mais plantada no país a partir da segunda metade do século passado. O país tem destaque no cenário mundial sendo o segundo maior detentor de florestas plantadas de eucalipto, atrás apenas da Índia que contém 8,0 milhões de hectares dos cerca de 18 milhões de hectares plantados no mundo (FAO, 2000). Atualmente, o eucalipto é a espécie florestal tropical e sub-tropical de maior importância econômica mundialmente (Figura 1).

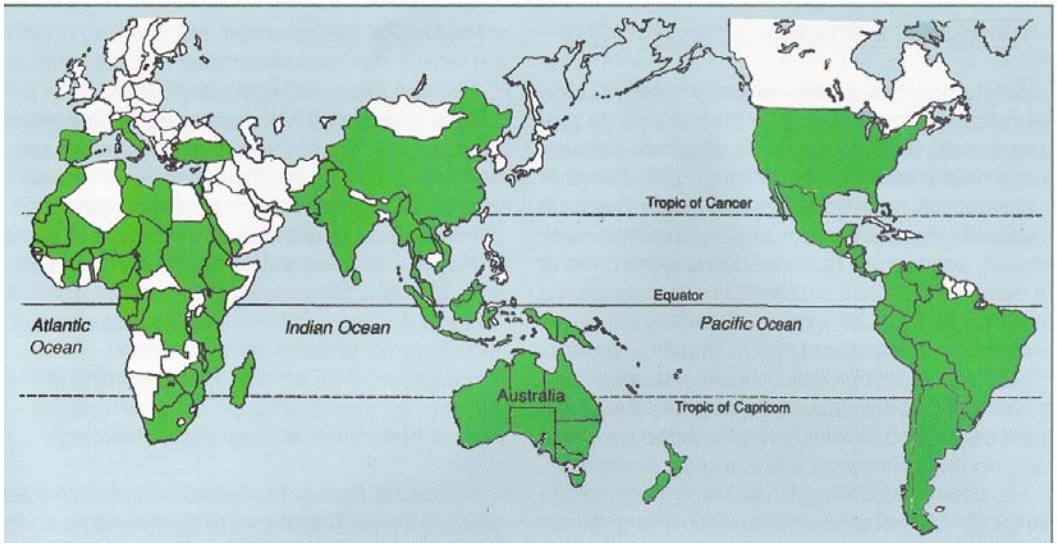


Figura 1: Distribuição dos países com plantações de eucalipto, de acordo com dados da FAO (1997).

Florestas de eucalipto de rápido crescimento suprem hoje a demanda por biomassa lenhosa com propriedades tecnológicas específicas para diversos setores industriais brasileiros, notadamente o de papel e celulose. Em 2000, na relação de produtos por fator agregado, o setor de papel e celulose supera produtos como automóveis, calçados e

semi-manufaturados de ferro e aço. Em 2003, as exportações do setor somaram US$ 3,1 bilhões (GAZETA MERCANTIL, 2003) e, em 2004, estimou-se que a produção de eucalipto movimenta R$ 12 bilhões por ano no país (VALOR ECONÔMICO, 2004), gerando 102.000 empregos diretos e 500.000 indiretos (MORA E GARCIA, 2000).

A madeira do eucalipto adequa-se bem à produção de papel em função de possuir fibras delgadas, curtas, de granulação reduzida, além de um alto número de fibras por grama. Observa-se também que, pelo fato das microfibrilas apresentarem uma pequena angulação em torno do eixo da fibra, esta é rígida, proporcionando uma estrutura volumosa de papel, além de alta opacidade.

A segunda maior utilização da madeira de eucalipto no Brasil, é na produção de carvão vegetal para a indústria siderúrgica. A lenha, é provavelmente, o energético mais antigo usado pelo homem e continua tendo grande importância na matriz energética brasileira, participando com cerca de 10% da produção de energia primária (BANCO DE DADOS DE BIOMASSA NO BRASIL). O Brasil é o maior produtor mundial desse insumo energético. No setor industrial (quase 85% do consumo), o ferro-gusa, aço e ferro-ligas são os principais consumidores do carvão de lenha, que funciona como redutor (coque vegetal) e energético ao mesmo tempo. Em termos globais, o volume de recursos proveniente da madeira ou de seus derivados representa mais de 1% da economia mundial (SEDEROFF, 1999).

A madeira de eucalipto no Brasil tem seu principal uso na indústria de papel e celulose. Nesta, o processo de polpagem constitui-se em uma etapa intermediária e bastante onerosa, visando a separação da celulose dos demais componentes da madeira. Dentre estes componentes indesejáveis encontra-se principalmente a lignina (JUNG, 1998). Para o processo de produção do papel, estima-se que a redução de 1% no teor de lignina da madeira representaria uma economia de 1 milhão de dólares por ano para

uma indústria com capacidade de produção de 300.000 toneladas de celulose anuais. Portanto, pequenas reduções no teor de lignina da madeira representariam grandes economias às indústrias papeleiras (ANTEROLA e LEWIS, 2002).

Para trazer benefícios práticos através da melhoria na qualidade do papel, nos custos de produção e na redução de resíduos poluentes desta indústria, esforços estão sendo concentrados com o objetivo de aprimorar ainda mais a qualidade da madeira do eucalipto procurando alterar a concentração ou composição da lignina. Uma pressão concreta para esta direção advém do aumento contínuo de demanda por madeira, previsto em mais de 50% até 2010 segundo a FAO (Food and Agriculture Organization). Melhorias na produtividade ou qualidade da madeira podem ser alcançados pelo melhoramento genético clássico e com pesquisas em biotecnologia visando, por exemplo, o melhoramento por seleção assistida focado em fatores determinantes de características da madeira. Portanto, existe um grande interesse em caracterizar os eventos principais na diferenciação do xilema uma vez que estes devem ser fatores decisivos na determinação das propriedades da madeira, refletindo no seu valor e desempenho industrial (PAUX *et al*., 2004).

Em termos gerais ocorreu pouca domesticação de árvores florestais quando comparado a culturas agrícolas tradicionais, tais como o milho e trigo. As razões principais são o tamanho das árvores e seu alto tempo médio de geração, que dificultam, mas não impossibilitam, tentativas coordenadas de melhoramento genético tradicional. De todo modo, diante de uma demanda crescente e de necessidades específicas dos diferentes setores industriais, observa-se uma grande pressão pelo estabelecimento de sistemas de produção eficientes capazes de suprir a matéria-prima, bem como deter a extração predatória que vem dizimando o meio-ambiente em nível mundial.

De certa forma a diminuição desta lacuna está ocorrendo através da utilização de técnicas biotecnológicas como a micropropagação, seleção assistida por marcadores, e recentemente a transgenia. Estudos utilizando marcadores também oferecem novos horizontes para a aquisição de novos conhecimentos, que eventualmente poderão ser utilizados para o aprimoramento de características de interesse.

1.2 As Espécies de Eucalipto

Foram identificadas aproximadamente 670 espécies do gênero *Eucalyptus* sendo apenas duas espécies não nativas da Austrália (*E. urophylla* e *E. deglupta*). Neste país formam-se extensas florestas naturais, com grande número de variedades e híbridos. A sua introdução na Europa deu-se por volta de 1773 e na Índia em 1856. Já na América do Sul, o Chile provavelmente foi o primeiro país a introduzir o eucalipto, enquanto que no Brasil admite-se que as primeiras mudas teriam sido plantadas em 1868 (MORA e GARCIA, 2000) com intuito meramente decorativo.

A maioria das espécies é composta por árvores altas (30 a 50 metros), sendo que existem também árvores menores (10 a 25 metros) e arbustivas (30 ou 40 espécies). Quanto ao plantio, a seleção da espécie depende das condições do clima e do solo, sendo que as melhores são aquelas que se aproximam mais do local de origem da espécie.

Uma das espécies mais utilizadas comercialmente é *E. grandis*. Esta ocorre naturalmente na Austrália em áreas com altitude variando desde o nível do mar até 600 metros e no caso da região de Atherton (estado de Queensland), entre 500 e 1100 metros de altitude. Sua madeira é considerada medianamente leve e é fácil de ser trabalhada em operações de usinagem. É considerada de baixa estabilidade, mas de elevada permeabilidade. Quando provinda de plantações de ciclo longo, a madeira é utilizada

intensivamente na Austrália, África do Sul, Brasil e Argentina como madeira de construção e matéria-prima na fabricação de móveis. Já quando oriunda de plantações em ciclos curtos é utilizada em caixotaria, paletes, carvão e mourões. É uma das espécies mais plantadas no mundo e considerada uma das mais versáteis e indicadas para múltiplos fins.

A espécie *E. urophylla* é uma das duas espécies de ocorrência natural fora do território australiano, ocorrendo, naturalmente, na ilha de Timor e outras ilhas a leste do arquipélago indonesiano em altitudes variando de 400 a 3.000 metros. A madeira é considerada medianamente leve, e as propriedades de resistência mecânica são moderadas. Inúmeros esforços são realizados para a introdução da espécie fora das condições de sua zona natural; os resultados são bastante satisfatórios, com a espécie apresentando alta plasticidade, adaptando-se a solos hidromórficos ou francamente arenosos, em diferentes altitudes. É considerada apta para regiões onde não ocorrem geadas e situações de déficits hídricos severos. No Brasil, a espécie tem sido plantada intensivamente em programas de melhoramento genético, principalmente de hibridação.

Outra espécie de importância é *E. globulus*, uma das espécies nativas mais cultivadas da Austrália. Caracteriza florestas de tamanho médio até bem alto (70 metros), mas normalmente é encontrada variando entre 15 e 25 metros. Sua madeira é forte, durável e possui anéis distintos. Uma característica importante desta espécie é a boa qualidade da madeira para a indústria papeleira, devido a sua maior quantidade de celulose e menor de lignina.

1.3 A Lignina

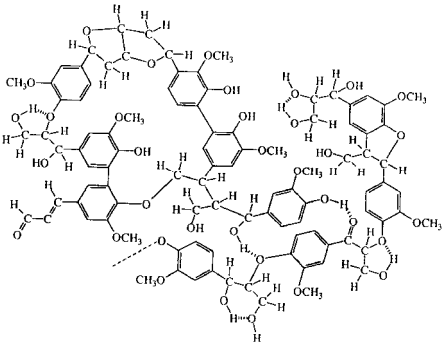
Nos últimos anos um esforço crescente tem sido feito de forma a investigar as bases moleculares da formação, estrutura e função da madeira. Suas propriedades físicas de

interesse compreendem sua densidade básica, o ângulo das microfibrilas, a espessura da parede celular, o diâmetro do lúmen e o comprimento das fibras. Quimicamente também existem propriedades de interesse, tais como abundância e composição dos três principais componentes da madeira: celulose, hemicelulose e lignina. Em termos histológicos, a madeira é composta pelo xilema secundário, o qual é derivado do crescimento lateral do câmbio vascular. As células derivadas deste meristema podem se diferenciar no xilema, que é um tecido de suporte e de condução de água; e no floema, responsável pelo transporte de nutrientes. As paredes secundárias são estruturas complexas formadas principalmente por moléculas como celulose, hemicelulose e lignina.

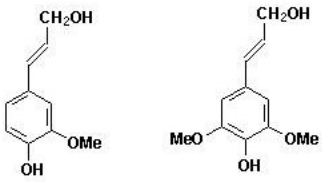
A celulose e a lignina são respectivamente os biopolímeros mais abundantes do planeta, e como visto anteriormente, são os principais componentes no processo de produção de papel. O cenário ideal para a indústria de produção de papel, seria uma engenharia metabólica onde se aumentasse o teor de celulose e, ao mesmo tempo, se diminuísse a quantidade de lignina. Para tanto é fundamental o entendimento da biossíntese e regulação destes componentes, o que efetivamente vem ocorrendo através de diversos estudos, principalmente tendo a via de lignificação como alvo (ANTEROLA e LEWIS, 2002; COSTA *et al*., 2003; RAES *et al*., 2003).

A lignina é um heteropolímero fenólico (Figura 2A) altamente hidrofóbico, com diversas ligações cruzadas (SEDEROFF, 1999), que tem função de suporte mecânico (esclerênquima), condução (xilema), e proteção e defesa (periderme) (BOUDET, 1998 & JUNG, 1998). O polímero de lignina é composto por unidades de fenilpropano acopladas oxidativamente por ligações éster e carbono-carbono. Em resposta a ataques microbianos, este polímero pode atuar como barreira física, e, por ter propriedades hidrofóbicas, é um componente fundamental do sistema vascular das plantas. Nas

angiospermas, a lignina é composta majoritariamente por monômeros de guaiacil (G) e siringil (S). A diferença entre as mesmas se encontra na metilação, enquanto guaiacil é unicamente metilado no grupo 3-hidroxil, o siringil é metilado tanto no grupo 3- quanto no 5-hidroxil (ZUBIETA, 2002) (Figura 2B). Quanto maior for a proporção de guaiacil, mais condensado é o composto de lignina. Por outro lado, quanto maior for a proporção de siringil, mais maleável será esse composto e mais interessante para a indústria papeleira será visto que a separação da celulose e a lignina se dá de forma mais fácil, reduzindo o processamento químico necessário.



(A)



(B)

Guaiacil (G) Siringil (S)

Figura 2: (A) Estrutura química do polímero da lignina. (B) Estrutura química dos monômeros formadores da lignina.

Para modificar a proporção (S/G) de forma a ficar mais atraente para a indústria, procura-se alterar a razão (S/G) de forma a cada unidade adicionada nesta taxa, seja duplicada a taxa de remoção de lignina (CHANG, 1973). Para se fazer esta alteração, deve-se influenciar de alguma forma algum passo de sua via biossintética.

1.4 A Via de Lignificação

A via de biossíntese dos precursores da lignina é complexa e não existe um consenso geral sobre sua arquitetura. Um dos desenhos mais aceitos encontra-se representado na Figura 3. O seu ponto de partida é o aminoácido fenilalanina, que sofre uma deaminação e sucessivas hidroxilações e metilações até formar os monômeros da lignina.

Esta via tem sido o foco de grande interesse experimental, e 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, 1998), álamo (*Populus sp*.; STERKY, 1998) e *Arabidopsis thaliana* (SIBOUT *et al*., 2003), dentre outros.

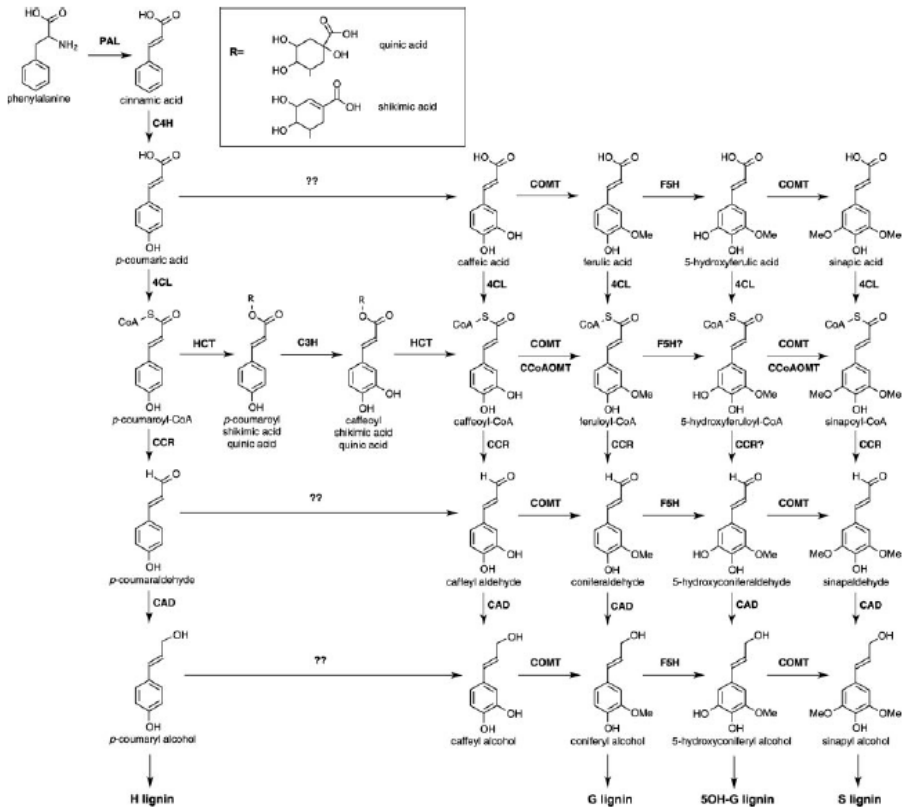


Figura 3: Via biossintética proposta para a formação de monolignóis (adaptação livre de Boerjan *et al*., 2003).

Para promover uma alteração da lignificação, modificações na via em determinados pontos poderão proporcionar o efeito desejado (HU, 1999). Estudos superexpressando (LI *et al*., 2003) e subexpressando (HU *et al*., 1999; PILATE *et al*., 2002; LI *et al*., 2003; SOMERVILLE *et al*., 1992; SEDEROFF *et al*., 1999) alguns genes da via de lignificação em diversos organismos têm demonstrado importantes alterações nas características finais da via de lignificação, como a diminuição de 45% da lignina e aumento de 15% da celulose observados por HU *et al*. (1999), assim como uma maior facilidade na deslignificação, menor utilização de químicos, e maior qualidade da polpagem observados por PILATE *et al*. (2002).

1.5 Enzimas Estudadas

Diante da complexidade da via de lignificação, é importante que o comportamento individual de seus componentes seja estudado em detalhes, abordando aspectos como a atividade bioquímica, variabilidade natural e regulação da expressão gênica. No contexto do presente trabalho, dois genes chave da via de lignificação foram estudados: *CAD* (cinamil-álcool desidrogenase) e *COMT2* (ácido cafeico 3-O-metiltransferase).

1.5.1 A enzima CAD

A enzima CAD (EC 1.1.1.195) catalisa o último passo de uma das rotas da síntese de fenilpropanóides (Figura 4) específico para a produção dos monômeros da lignina. Esta tem como substratos, os alcoois coniferil, sinapil, 4-coumaril e cinamil. A proporção desses monômeros precursores diretos do polímero da lignina varia de acordo com a espécie (HALPIN, 1992). Sua atividade catalítica compreende a catálise do álcool cinamílico mais NADP+ para cinamaldeído mais NADPH.

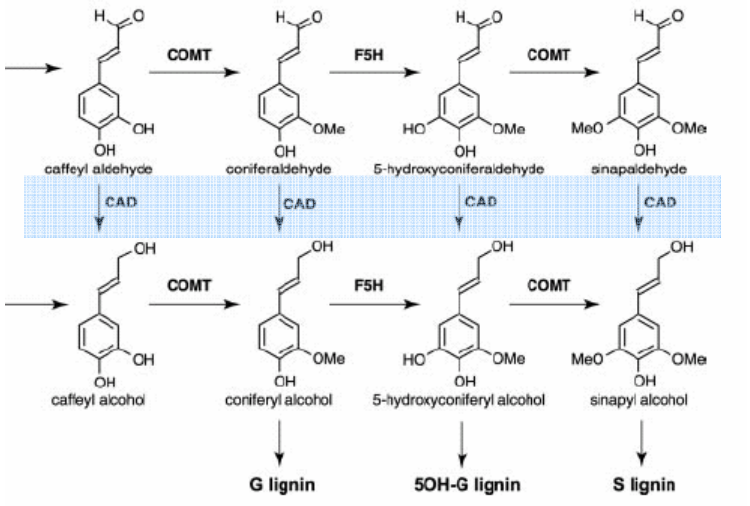


Figura 4: Reações catalisadas pela CAD (Cinamil-alcool desidrogenase) para a formação da lignina (Adaptado de BOERJAN *et al*., 2003).

A enzima CAD, segundo BOERJAN *et al*. (2003), é caracterizada por possuir três famílias. A primeira compreende aquelas que estão envolvidas na lignificação; a

segunda compreende várias álcool desidrogenases que possuem diversos substratos. Estas têm como característica serem diméricas ou tetraméricas, ligando-se a dois átomos de zinco por subunidade. Estas enzimas já foram encontradas em bactérias, mamíferos, plantas e fungos e, geralmente, está presente mais de uma isozima. Já a terceira família é composta por uma álcool desidrogenase de alfafa que é capaz de catalisar a redução de cinamaldeído, sinapaldeído e coniferaldeído, além de vários aldeídos alifáticos e benzaldeídos.

Em um experimento realizado por NI e JUNG, (1998), utilizando uma planta transgênica antisenso (introdução de porção da fita complementar ao gene objeto de estudo) para *CAD*, foi demonstrada uma redução maior de siringil em relação a guaiacil em tabaco. Analisando comparativamente seus resultados com os de RALPH *et al*. (1998), notamos que a deficiência de CAD provoca um acúmulo de aldeídos, sendo que *Pinus sp* (RALPH *et al*., 1998) acumulou menos aldeídos que o tabaco (NI e JUNG, 1998). Notou-se então que *Pinus* tem um componente adicional da lignina não usual, derivado do álcool dihidroconiferílico, não observado em tabaco. Notamos então, quão difícil poderá ser para trabalhos posteriores, a caracterização funcional dos genes desta via.

1.5.2 A enzima COMT

A proteína COMT (EC 2.1.1.68), também chamada CAOMT catalisa a conversão de ácido cafeico a ácido ferrúlico e ácido 5-hidroxiferrúlico para ácido sinápico (Figura 5). Os produtos resultantes são convertidos nos álcoois correspondentes que serão então incorporados à lignina. A COMT faz parte da superfamília das metiltransferases, da família do tipo 2 e da subfamília COMT.

Em seu estudo, BOERJAN *et al*., (1998) observou que o gene *COMT* é fortemente expresso em todos tecidos estudados, tais como plântulas, raiz, folha, flor e caule, dando ênfase à intensidade observada de sua expressão no caule da inflorescência.

Estudos utilizando a estratégia antisense têm mostrado que, mesmo reduzindo a expressão de COMT em 90%, o conteúdo de lignina não é afetado, mas sua estrutura química, em termos da compreensão dos monômeros, é dramaticamente alterada (LAPIERRE, 1999).

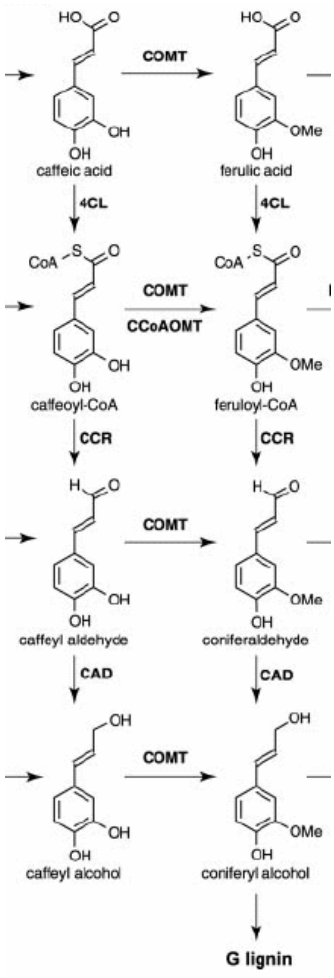


Figura 5: Reações catalisadas pela COMT (Ácido cafeico 3-O-metiltransferase) na via de lignificação (Adaptado de BOERJAN *et al*., 2003).

1.6 O Projeto Genolyptus

Recentemente foi estabelecida no Brasil uma rede cooperativa de estudos genômicos do eucalipto, denominada rede Genolyptus (Rede Brasileira de Pesquisa do

Genoma de *Eucalyptus*). Esta tem como objetivos o descobrimento, seqüenciamento, mapeamento e a determinação da função de genes de importância econômica de diferentes espécies de eucalipto, visando a incorporação de tecnologias de genética molecular nos programas de melhoramento e produção vegetal. O presente trabalho está inserido no projeto Genolyptus de forma que seus resultados sirvam de fomento para novas pesquisas por parte deste projeto.

Participam deste projeto 12 empresas de papel e celulose, 7 universidades, além da EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária) caracterizando-se como a maior e mais complexa rede de experimento florestal do mundo.

O Genolyptus produziu uma série de insumos que foram utilizados no presente estudo, tais como uma biblioteca de BACs ("Bacterial Artificial Chromosomes"), além de diversos bancos de cDNA para diversos tecidos, como xilema e folhas, de quatro espécies de *Eucalyptus*: *E. grandis*, *E. globulus*, *E. pellita* e *E. urophylla*.

Os cromossomos artificiais de bactéria (BACs) são vetores utilizados para clonar grandes fragmentos de DNA em células de *Escherichia coli*, baseados no fator F encontrado na mesma. Os BACs foram desenvolvidos para receber fragmentos de DNA muito maiores que os suportados por plasmídeos. Enquanto plasmídeos suportam insertos de até 10 kb, os BACs chegam a suportar 350 kb, valor maior que o do genoma de alguns vírus, como o da herpes (100 a 250 kb). Tais insertos são utilizados para os mais variados fins, como seqüenciamento e transformações genéticas. A biblioteca de BACs do Genolyptus foi desenvolvida a partir de um indivíduo de *E. grandis*, e com cobertura de quatro vezes o genoma. Com insertos de 150 Kb em média, esta biblioteca pode ser utilizada para se obter a sequência completa dos genes da via de lignificação.

Também no contexto do Genolyptus, gerou-se a partir dos bancos de cDNA, ESTs, que se caracterizam por ser pequenos fragmentos de seqüência gênica (geralmente entre

200 e 600 nucleotídeos), que são gerados por seqüenciamento único de uma ou ambas pontas do gene expresso (RNA mensageiro – mRNA), resultando em um fragmento de qualidade de seqüência relativamente baixa.

Tanto a biblioteca de BACs quanto o banco de ESTs são fontes bastante úteis para os mais variados estudos genômicos, principalmente para a descoberta de genes. Uma outra utilidade dos mesmos seria o estudo de diversidade nucleotídica de genes em que se ainda tem pouca informação de seqüência.

1.7 Diversidade nucleotídica

A própria variabilidade genética natural existente entre as diferentes espécies de eucalipto gera variações no conteúdo e composição de lignina, além de outros aspectos relevantes na produção da celulose como o comprimento de fibras. Do ponto de vista da produtividade de celulose, essa variação é interessante, pois algumas espécies são naturalmente mais produtivas e requerem menor gasto no processamento químico.

A fonte desta variabilidade é uma alteração súbita e herdável no material genético, caracterizada como mutação. Quando estas são localizadas e alteram apenas poucas bases, são chamadas de mutações de ponto, ou mutações gênicas. Por outro lado, se as mesmas alterarem muitas bases ou até cromossomos inteiros, são chamadas de mutações ou aberrações cromossômicas.

No ponto de vista do DNA, existem dois tipos de mutações pontuais, as substituições e as inserções ou deleções de bases (indels). As substituições podem ser classificadas de acordo com suas características. Aquelas que modificam uma base por outra de mesmas características químicas (uma purina [A e G] por outra purina, e uma pirimidina [C e T] por outra pirimidina) são chamadas de transições. Já aquelas que

substituem bases de características químicas distintas (purina substituída por uma pirimidina, e vice versa), são chamadas de transversões.

Quando ocorre uma indel dentro da região codificadora de um gene, o quadro de leitura pode ser alterado (*frameshift*), acarretando na mudança dos aminoácidos do restante da proteína. Por outro lado, devido à redundância do código genético, nem todas as alterações de pares de base acarretarão a alteração dos aminoácidos da proteína, principalmente quando ocorrem na terceira base do códon.

Estas alterações de bases que não acarretam modificação do aminoácido são chamadas de silenciosas. Como este tipo de mutação não tem nenhum efeito positivo ou negativo na descendência do indivíduo, acaba por ser considerada uma mutação neutra. Aquelas mutações que acarretam uma alteração do aminoácido, mas este sendo com características químicas similares, são consideradas substituições conservativas, isto porque provavelmente afetarão pouco ou nada na estrutura e funcionamento da proteína.

Por outro lado, as substituições que geram aminoácidos com características químicas diferentes, são chamadas de não conservativas, e poderão afetar, em diferentes intensidades, a estrutura final da proteína. Existem também as substituições que acabam por gerar códons de terminação. Estas podem inviabilizar a funcionalidade da proteína por interromper a seqüência da mesma, possivelmente impossibilitando a atividade final da mesma.

Tais mutações podem ser causadas por substituição de uma base por outra, assim como visto acima (ação do agente mutagênico ácido nitroso, por exemplo); por análogos de bases, como a 5-bromouracila (5-BU) que é um análogo da timina e que forma par com a adenina. Este 5-BU pode perder um átomo de hidrogênio (seu estado raro), passando a parear com a guanina; outra causa de mutações seriam adições e

remoções de bases por agentes mutagênicos como a acridina; além de poderem ser causadas por modificações tautoméricas.

Estes tautômeros são estados raros das bases comuns (A,C,T e G) que ocorrem a partir do rearranjo na distribuição dos prótons e elétrons. No DNA, é encontrada normalmente a forma ceto de cada base, enquanto que a forma enol destas bases é rara. Seus pareamentos são distintos e, desta forma, tais enóis são uma fonte de mutações raras. O processo de substituição das bases acontece semelhantemente ao de 5-BU, com o pareamento do tautômero com uma base não usual (Figura 6), provocando, na próxima replicação, uma alteração das bases.

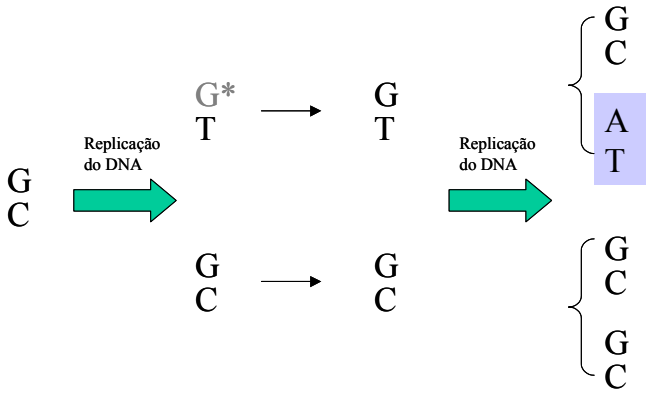


Figura 6: Exemplificação de mutação por tautomerismo de base no DNA (modificado livremente de GRIFFITHS *et al*., 2001).

1.7.1 SNPs (Single Nucleotide Polymorphism)

A disponibilidade crescente de seqüências parciais de genes, promoveu a descoberta de uma grande quantidade de polimorfismos de uma base entre indivíduos de uma mesma espécie, tais polimorfismos de base individual são chamados de SNPs ("Single Nucleotide Polymorphism") (Figura 7).

SNPs ocorrem aleatoriamente a cada 1250 bases aproximadamente em humanos (LEWIS, 2002) alternando os nucleotídeos A, C, T e G com uma freqüência maior que 1% na população (BROOKES, 1999).

Os SNPs que ocorrem em exons e que são responsáveis por uma variação em um aminoácido da proteína são chamados de não sinônimos (um exemplo de variação não sinônima é uma mutação na hemoglobina provocando a anemia falciforme). Os que não variam a seqüência de aminoácidos são chamados de sinônimos. Mesmo não alterando a proteína, os SNPs sinônimos podem alterar a estrutura do RNA mensageiro e desestabilizá-lo, afetando a concentração final de proteína (GRIFFITHS *et al*., 2001). SNPs podem também afetar o splicing alternativo; gerar alterações no padrão de expressão de genes quando ocorrem em regiões promotoras; gerar ou suprimir códons de terminação ou poliadenilação na molécula de RNA mensageiro; e alterar códons de iniciação de tradução.

Em milho, SNPs no gene *dwarf8* foram associados com florescimento (THORNSBERRY, 2001). Estas associações podem resultar no desenvolvimento de marcadores moleculares para seleção assistida em plantas baseados na variabilidade de seqüência de genes e não somente em marcadores microssatélites ligados (MORGANTE, 2003). Marcadores moleculares são todo e qualquer fenótipo molecular oriundo de um gene expresso ou de um segmento específico do DNA (FERREIRA e GRATTAPAGLIA, 1998). Em situações ideais, o uso de genes como marcadores para seleção de árvores permitiria 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.

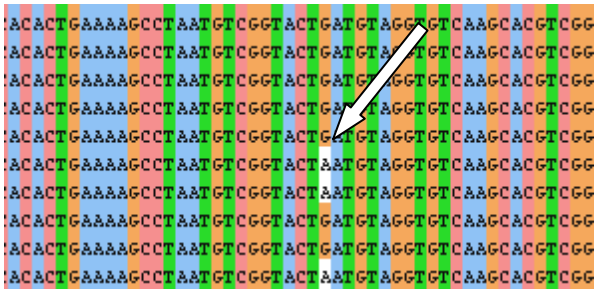


Figura 7: Exemplificação de polimorfismos identificados através de alinhamento de sequências. Cada linha representa a seqüência de um indivíduo para uma determinada região gênica. A seta indica a posição onde se encontrou o polimorfismo.

Através da análise de SNPs, pode-se constatar a presença de blocos de DNA incluindo diferentes locus que tendem a ser herdados em conjunto, os quais são denominados haplótipos. Por terem esta característica e por estarem sendo descritos em larga escala (aproximadamente 32.000.000 de submissões no banco de dados de SNPs atualmente[GenBank]), os haplótipos tornaram-se alvos para estudos como marcadores em estudos de associação (HIRSCHHORN, 2005) de genes relacionados a doenças, estudos de demografia genética e de evolução cromossomal. Estudos têm mostrado que, no genoma humano, cerca de 11 milhões de SNPs têm grupos de vizinhos estreitamente relacionados entre si (HIRSCHHORN, 2005, THE INTERNATIONAL HAPMAP CONSORTIUM, 2003). Em variedades elite de milho, ocorre um SNP a cada 60 pb (CHING *et al*., 2002) assim como em *Populus* (INGVARSSON, 2004), já no gene *CCR* de *Eucalyptus nitens*, THUMMA *et al*. (2005) observou um SNP a cada 94pb.

Sabendo então do padrão de desequilíbrio de ligação para determinada região, um conjunto de marcadores pode ser desenvolvido de forma a se tentar associar a presença dos mesmos a possíveis fenótipos de interesse.

1.7.2 Haplótipos e Desequilíbrio de Ligação

Os haplótipos são constatados quando dois SNPs segregam conjuntamente. Desta maneira, estarão então caracterizando um DL (Desequilíbrio de ligação) que é a falta de independência, de forma estatística, entre os alelos em dois loci. O DL existe entre dois loci ligados quando alelos nestes ocorrem no mesmo haplótipo de maneira mais freqüente do que o esperado (BANSAL, 2003; RAFALSKI, 2002; REMINGTON, 2001). Isto acaba por caracterizar uma segregação dependente entre os dois ou mais loci analisados, ou mais (LONG, 2004). O DL, segundo REMINGTON (2001), é geralmente dependente da história de recombinações entre os polimorfismos. Fatores tais como deriva genética, seleção entre populações, migração (miscigenação) e a redução do tamanho populacional (*bottleneck*), podem modificar o DL entre os marcadores e as características relacionadas. Fatores que aumentam o DL são o endocruzamento, tamanho pequeno da população, isolamento genético entre linhagens, subdivisão entre populações, baixa taxa de recombinação, seleção natural e artificial. Por outro lado, alguns fatores que diminuem o DL são a fecundação cruzada, altas taxas de recombinação e mutação.

A extensão do desequilíbrio de ligação varia entre 50 e 100 kb em humanos (RAFALSKI, 2002; THUMMA *et al*., 2005; BOREVITZ e NORDBORG, 2003), entre 150 e 250 kb em *Arabidopsis* (BOREVITZ e NORDBORG, 2003; THUMMA *et al*., 2005), aproximadamente 100 kb em arroz (BOREVITZ e NORDBORG, 2003), entre 100 e 500 kb em linhagens elite de milho com efeito fundador (RAFALSKI, 2002; CHING *et al*., 2002), e é cerca de 10 cM em cevada (THUMMA *et al*., 2005).

Uma associação entre SNPs individuais ou haplótipos de vários SNPs com algum fenótipo não necessariamente indica uma relação direta entre o polimorfismo de seqüência e a perda ou ganho de função (GRATTAPAGLIA, 2004). Falsas associações

podem ser detectadas onde o DL é resultado de uma subestruturação da população e não de uma associação verdadeira (CARDON, 2001).

De forma a avaliar se os dados estão de acordo com a hipótese de independência (Ho), ou seja, avaliar se os locos estão ou não segregando independentemente, a utilização de uma medida estatística deve ser empregada. O teste do qui-quadrado (χ^2) procura ajudar no julgamento dos dados de forma a avaliar se os mesmos são ou não significativos, isto porque correspondências muito próximas ou diferenças muito grandes normalmente não são de difícil análise, mas existem áreas que inevitavelmente a correspondência avaliada será duvidosa (GRIFFITHS *et al*., 2001). Desta forma, o mesmo é utilizado para ajudar a tomar a decisão de manter ou rejeitar a hipótese proposta.

Como geralmente não se têm a distância exata de ligação para se usar na hipótese, um teste de ausência de ligação deve ser realizado. Desta forma, se os resultados indicarem uma rejeição da hipótese de não ligação, pode-se deduzir que existe a ligação. Esse teste permite a estimativa da probabilidade de ocorrência dos desvios, dada a condição de independência. Normalmente, utiliza-se como ponto crítico para a rejeição da hipótese de independência (Ho), a probabilidade de 5% (P<0,05) ou a de 1% (P<0,01). Neste trabalho será utilizada P<0,05. Valores abaixo de 5% significam rejeitar Ho e, portanto, assumir que os locus não estão segregando independentemente. A partir de uma tabela de contingência gerada de forma a ilustrar as proporções alélicas dos loci, é utilizada a fórmula

$$\chi^2 = \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}}$$

A extensão do DL pode ser medida pelo valor de *D* (coeficiente de desequilíbrio). O coeficiente de desequilíbrio é a diferença entre a frequência haplotípica observada e a esperada na condição de equilíbrio de ligação (BANSAL *et al*., 2003; DEVLIN e RISCH, 1995).

$$D = \pi_{11}\,\pi_{22} - \pi_{12}\,\pi_{21}$$

Sua forma normalizada (*D'*) foi demonstrada por LEWONTIN (1964), onde o numerador é o valor de *D* e o denominador igual ao máximo absoluto de *D*.

O denominador típico de *D'* é $\pi_{+1}\pi_{2+}$ para características (doenças) raras e amostras aleatórias porque ele é o mínimo se π_{12}-π_{21} = π_{1+}-π_{+1} >0. Esta condição será encontrada sempre que outro alelo associado ao marcador for mais comum que o alelo da característica de interesse. Caso contrário, o denominador deverá ser alterado.

$$D' = \begin{cases} \dfrac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{min(\pi_{1+}\pi_{+2},\ \pi_{+1}\pi_{2+})} & D > 0 \\ \dfrac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{max(\pi_{1+}\pi_{+1},\ \pi_{+2}\pi_{2+})} & D < 0 \end{cases}$$

Outra medida comum de desequilíbrio é o quadrado do coeficiente de correlação entre os loci (r^2 ou Δ^2) (DEVLIN e RISCH, 1995; WEISS e CLARK, 2002).

$$\Delta = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})^{1/2}}$$

1.8 Estudos de associação

A partir da identificação de SNPs e do seu padrão de DL, estes podem ser utilizados em estudos de associação de forma a identificar QTLs ("Quantitative Trait Locus")

(BOREVITZ e CHORY, 2004; BOREVITZ e NORDBORG, 2003; DEVLIN e RISCH, 1995). Segundo BOREVITZ (2003), o método básico para se identificar loci responsáveis por variações em características complexas era o mapeamento genético, já o mapeamento de ligação utilizava pedigrees e cruzamentos, mas uma nova tecnologia capaz de analisar populações naturais de indivíduos não relacionados se mostra bastante promissora, o mapeamento por desequilíbrio de ligação. Este estudo utiliza a recombinação para delimitar a região relacionada ao fenótipo e a medida de associação alélica para detectar o desequilíbrio de ligação.

Como, em termos estatísticos, um alelo é típicamente responsável por uma parte muito pequena da variação fenotípica, poucos QTL tem sido molecularmente identificados (BOREVITZ e NORDBORG, 2003). Além disso, ZONDERVAN e CARDON (2004), em seu estudo, demonstraram ser mais eficientes e poderosos estudos que detectam alelos de susceptibilidade a doenças onde a diferença na freqüência dos alelos é pequena. Se a diferença das freqüências for muito grande, especialmente em casos em que o SNP causador é raro (CRAWFORD *et al*., 2005), provavelmente somente estudos procurando determinantes genéticos de grandes efeitos poderão obter sucesso (ZONDERVAN e CARDON, 2004). Outros problemas são os falsos positivos (diferenças entre freqüências de alelos) e os falsos negativos (limitações estatísticas) (BOREVITZ e NORDBORG, 2003).

Para análises de ligação, os haplótipos são mais informativos que alelos simples, especialmente para estudos de desequilíbrio de ligação e de análise de associação. Vários métodos já foram descritos para determinar haplótipos de SNPs, como híbridos de células somáticas, seqüências clonadas para estudos de baixa escala, e PCR (Reação de Polimerase em Cadeia) alelo específica em estudos em larga escala de diplóides.

Neste último, cada SNP pode ser utilizado como parte da seqüência para iniciadores alelo específicos (MIR, 2000).

Aproveitando o incrível aumento do número de entradas de seqüências nos bancos de dados (JANSSEN, 2003; SAIER, 1998), SNPs e indels são fontes inesgotáveis de marcadores polimórficos para uso no mapeamento genético de alta resolução, assim como para estudos de associação baseados em genes candidatos ou, possivelmente, em todo o genoma (RAFALSKI, 2002). Por outro lado, REMINGTON (2001) sugere que somente polimorfismos com ligação muito forte com um locus associado a um efeito fenotípico parecem ser significantemente associados à característica em uma população aleatória, provendo uma resolução muito mais fina que o mapeamento genético.

2 OBJETIVOS

Detectar e avaliar a diversidade nucleotídica intra e interespecífica para dois genes da via de lignificação, além de obter a seqüência completa para um deles a partir da biblioteca de BACs do projeto Genolyptus.

3 MATERIAIS E MÉTODOS

3.1 Material genético da análise de diversidade

Para o estudo de diversidade nucleotídica, foram utilizadas folhas de 100 indivíduos de duas populações naturais de *Eucalyptus grandis*, além de 50 folhas de indivíduos de *E. globulus* e 50 de *E. urophylla* geneticamente não relacionados e com sementes provindas de árvores da Austrália. Estas sementes foram coletadas de plantas posicionadas a, no mínimo, 300 metros de distância, trazidas ao Brasil e plantadas por empresas florestais. Para realizar estudos intraespecíficos, foram coletadas folhas de *E. grandis* de duas procedências distintas de forma a aumentar a quantidade de amostras para verificar se há diversidade dentro de uma mesma espécie.

As sementes do primeiro grupo de *E. grandis* (GR) são procedentes de Atherton na Austrália, enquanto que as folhas oriundas do plantio de tais sementes são provenientes da Klabin Floresta, Paraná – Brasil. O outro grupo de *E. grandis* (GRR) tem suas sementes procedentes de Pine Creek (estado de New South Wales), também na Austrália. Tais indivíduos possuem sua população base no Brasil, na Fazenda Estrela, proveniente da Suzano Celulose e Papel (Suzano – SP).

Os indivíduos de *E. globulus* analisados neste trabalho têm procedência de Victoria – Austrália. As folhas para as análises são provenientes da Aracruz Celulose, Guaíba – RS.

Por fim, os indivíduos de *E. urophylla* têm como procedência o Timor. Sua população base encontra-se na fazenda Ribeirão Grande – MG, proveniente da Suzano Celulose e Papel, Suzano – SP.

A partir das folhas frescas dos indivíduos, para os estudos com o gene *CAD2*, foi isolado o DNA utilizando o método proposto por FERREIRA e GRATTAPAGLIA (1998) com modificações, para serem analisados 74 indivíduos de *E. grandis*, 10 de *E. urophylla* e 8 de *E. globulus*. Para *COMT2*, foram analisados 68 indivíduos de *E. grandis*, 29 indivíduos de *E. urophylla*, e 32 indivíduos de *E. globulus*.

3.2 PCR (Reação de polimerase em cadeia)

Cada reação de PCR foi realizada em volume de 25 µL, contendo 10 mM Tris-HCl (pH 8,4), 50 mM KCl, 2 mM $MgCl_2$, 200 nM de cada dNTP, 0,1 ng de BSA, 240 nM de cada primer, 2,5 u de Taq polimerase (Phoneutria, Biotecnologia e Serviços) e cerca de 15 ng de DNA. A desnaturação inicial foi realizada durante 4 min a 94°C. A partir desta desnaturação, se seguiam 30 ciclos compreendendo, 45 seg a 92°C, 45 seg à temperatura de anelamento (de acordo com o iniciador) e 1min e 30 seg a 72°C. A extensão final era realizada por 20 min a 72°C.

As reações foram então aplicadas em gel de agarose 1% e visualizadas em luz UV (ultra violeta) com coloração por brometo de etídio.

3.3 Biblioteca de BACs e triagem para genes candidatos

3.3.1 Material genético da triagem

No âmbito do projeto Genolyptus, foi desenvolvida uma biblioteca de BAC (Bacterial Artificial Chromosome) a partir de DNA genômico de *E. grandis* e disponibilizada pelo Dr. Sérgio Brommonschenkel (Universidade Federal de Viçosa).

Esta biblioteca possui 20.160 clones, com um tamanho médio de inserto na ordem de 150 kb, o que perfaz uma cobertura de aproximadamente quatro vezes o genoma. Adicionalmente, utilizou-se informações derivadas das ESTs do projeto Genolyptus (cerca de 100 mil entradas) para desenho de iniciadores, obtenção de seqüências e inferência de padrões de expressão.

3.3.2 Triagem de genes candidatos

A triagem dos clones de BAC foi desenvolvida objetivando identificar quais são os clones, dentre os 20.160 da biblioteca, que contêm o gene de interesse, além de utilizá-los, em um segundo momento, para o seqüenciamento completo do gene. A triagem foi realizada de forma a detectar clones positivos para os genes *CAD2* e *COMT2*.

3.3.3 Desenho dos iniciadores para a triagem

Procurando realizar a triagem dos clones de BAC, procuramos na literatura por trabalhos desenvolvidos utilizando genes da via de lignificação, além de desenvolvermos iniciadores através de estudos próprios. Estes últimos foram baseados em seqüências de *Eucalyptus* e *Arabidopsis* provenientes do banco de dados público Genbank (www.ncbi.nlm.nih.gov), além de seqüências do banco de cDNA (DNA complementar) do projeto Genolyptus. Para o desenho destes iniciadores, foi utilizado o programa Primer 3 (ROZEN & SKALETSKY, 1996) e parâmetros como tamanho entre 18 e 22 nucleotídeos e uma temperatura média de anelamento de 60°C. Estes

iniciadores não são demonstrados na Tabela 1 porque não foram eficientes na amplificação.

Para esta triagem, tanto para *COMT2* (Figura 8) quanto para *CAD2* (Figura 9), foram utilizados iniciadores (Tabela 1) desenvolvidos por Gion *et al*. (2000) (G-CAD F, G-CAD R, G-COMT F, G-COMT R) a partir de seqüências de *E. gunnii*. No caso de *CAD2*, a partir de um clone genômico e para *COMT2*, a partir de mRNA juntamente com o alinhamento múltiplo de clones genômicos de outras espécies. Os iniciadores desenvolvidos por POKE *et al*., (2003) foram utilizados nas análises de diversidade nucleotídica, relatadas mais adiante (3.4.1).

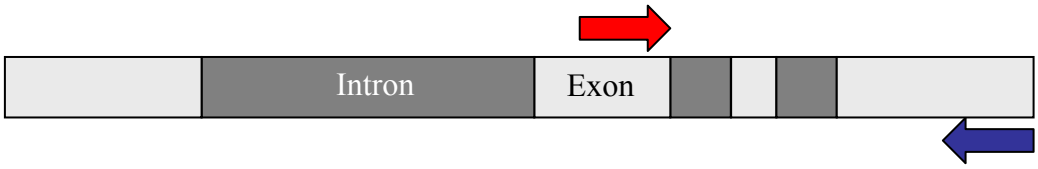


Figura 8: Esquema do anelamento do par de iniciadores (G-COMT F e G-COMT R) para *COMT2*. Em cinza, encontram-se os exons e, em preto, os introns.

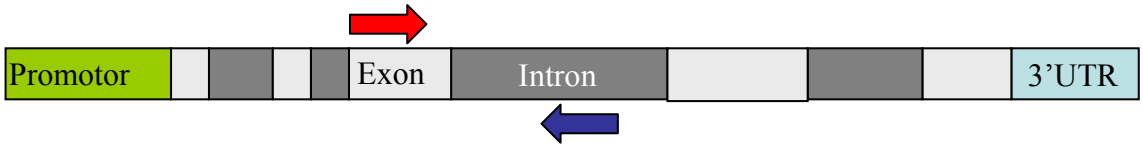


Figura 9: Esquema de anelamento do par de iniciadores (G-CAD F e G-CAD R) utilizado para a triagem dos clones de BAC contendo o gene *CAD2*. À esquerda encontra-se a região promotora, em cinza os exons, em preto os introns e à direita a região 3' UTR.

Tabela 1: Iniciadores utilizados neste trabalho para a triagem de BACs e para os estudos intra e interespecíficos. Ta – Temperatura de anelamento. Tm – "melting temperature".

Iniciador	Seqüência 5'-3'	Tm (°C)	Ta otimizada em PCR (°C)
G-CAD-F (Gion *et al.,* 2000)	TTgAgCAAAAATggggAgTCTT		
G-CAD-R (Gion *et al.,* 2000)	ATCTTCTggTCCCgTgTTTCTg		**65**
CAD 2386 F (Poke *et al*, 2003)	gCCAAAgTCACgTCTCTCTTTTggATTCTT		
CAD 5189 R (Poke *et al*, 2003)	CTTAggAAAAgCTggTgAgCACCCTATCAA		**65**
CAD 4578 F (Poke *et al*, 2003)	TCTTggATTgACggTCTTTCA	58,66	**50**
CAD 2423 F (Poke *et al*, 2003)	gTCCgTCTCTTTCCTCgTTg		
CAD 5081 R (Poke *et al*, 2003)	ACACAgCACAACCCAATTCA		**60**
G-COMT-F (Gion *et al.,* 2000)	CgCTCCACCCCTTCCT		
G-COMT-R (Gion *et al.,* 2000)	GGCTCCTTCACgACCTTTC		**69**

3.3.4 Formação dos grupos e supergrupos

Os clones referentes a biblioteca genômica de BAC foram crescidos em uma placa "deep well" contendo meio "Circle Grow" mais clorafenicol (concentração final de 12,5 μg/ml), incubados a 37°C a 280 rpm durante 16 horas para as minipreparações.

De forma a tentar facilitar a triagem, dada sua complexidade, foi montado um esquema de grupos e supergrupos destes clones de BACs (Figura 10) procurando aumentar a eficiência de detecção dos clones e diminuir os custos da detecção dos clones positivos. A partir das 210 placas que contêm os 20.160 clones, foram formados os 210 grupos, sendo que cada grupo representa uma placa, e a partir de alíquotas de seis grupos, formados cada um dos 35 supergrupos, ou seja, foram necessárias 35 PCRs para identificar qual o supergrupo positivo. A partir do supergrupo positivo, mais 6 PCRs para definir o grupo positivo. Na análise do grupo, 96 PCRs para a identificação do clone positivo.

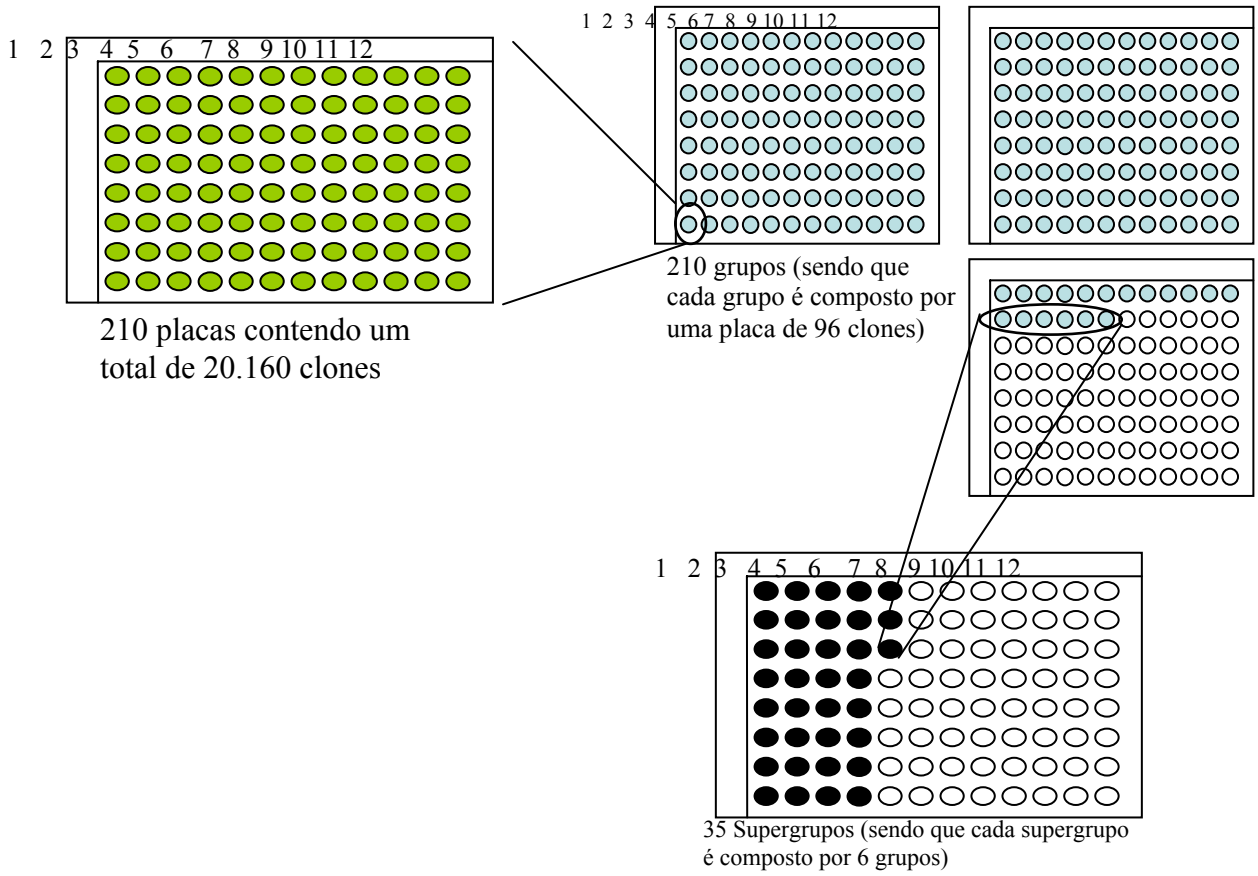


Figura 10: Esquema da construção dos grupos e supergrupos. A partir dos 20.160 clones que estavam dispostos em 210 placas, foram montados 210 grupos, sendo que cada um corresponde à uma placa inicial. Cada supergrupo foi então montado a partir de seis destes grupos.

A partir de amostras de tais clones e de sua disposição em grupos e supergrupos, foram feitas minipreparações de DNA (EMBRAPA, PLATAFORMA DE SEQUENCIAMENTO DE DNA, com alterações). Posteriormente, partiu-se para a análise via PCR primeiramente dos supergrupos, para somente depois analisar os grupos positivos, e então a placa positiva. Os iniciadores utilizados para estas reações encontram-se em negrito na Tabela 1.

3.3.5 Análise do clone de CAD2

Visando seqüenciar completamente um clone contendo o gene de *CAD2*, foi adotada a estratégia de seqüenciamento por "shotgun" (GREEN, 2001). Após a fragmentação mecânica e aleatória do inserto original que, neste caso, continha 30 Kb, uma sub-biblioteca foi criada com insertos de 1,0 Kb em média. A partir deste material foram seqüenciadas seis placas de 96 clones resultando em 768 leituras válidas de seqüências, o que se constitui em uma cobertura de aproximadamente dez vezes do inserto original. A partir destas seqüências foi realizada a montagem da seqüência através da utilização do programa CAP3 (HUANG e MADAN, 1999) que visa reunir as seqüências parciais que se sobrepõe para formar regiões contíguas (contigs) representando seqüências consenso.

Para proceder a montagem da estrutura gênica de *CAD2*, foi utilizado o programa est2genome (RICE *et al*., 2000) que prediz os genes por homologia de seqüência. Este programa alinha seqüências nucleotídicas processadas (EST, cDNA ou mRNA) com uma seqüência de DNA genômico, procurando obter a estrutura dos introns e exons.

Análises filogenéticas foram realizadas utilizando o programa PHYLIP (FELSENSTEIN, 1989). A árvore filogenética foi construída pelo método da máxima parcimônia. utilizando 1000 amostragens ("bootstrap") entre todas as seqüências e a árvore final representa o consenso destas. O comprimento dos ramos foi calculado por máxima verossimilhança da árvore consenso utilizando o modelo padrão de substituição de aminoácidos.

3.4 Diversidade nucleotídica

3.4.1 Desenho dos iniciadores para o estudo de diversidade

Visando realizar a análise da diversidade nucleotídica dos genes *CAD2* e *COMT2*, assim como descrito mais detalhadamente (ver em 3.3.1), foram desenvolvidos iniciadores a partir de estudos próprios, além de terem sido utilizados alguns provenientes de estudos de GION *et al*., (2000) e POKE *et al*. (2003) que podem ser observados na Tabela 1.

Nos estudos de diversidade nucleotídica intra e interespecífica de *CAD2*, foram utilizados iniciadores (Figura 11) desenvolvidos por POKE *et al*. (2003) enquanto que para *COMT2* foram utilizados os mesmos da triagem (G-COMT F e G-COMT R, apresentados na Tabela 1). A amplificação para o sequenciamento de *CAD2* foi realizada através de duas reações, sendo que a primeira utilizava os iniciadores 2386 F, 5189 R, e a segunda os 2423 F, e 5081 R. Quanto ao seqüenciamento, foram utilizados os iniciadores 4578 F e 5081 R.

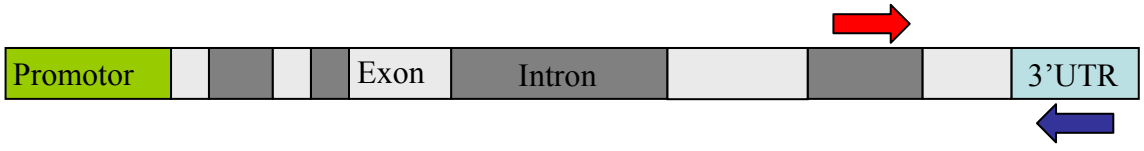


Figura 11: Esquema de anelamento do par de iniciadores (CAD 4578 F e CAD 5081 R) utilizados para o sequenciamento de *CAD2* para a análise de diversidade. Em verde encontra-se a região promotora, em cinza os exons, em preto os introns e em azul a região 3'UTR.

3.4.2 Seqüenciamento de DNA

Os produtos da PCR, após a análise em gel de agarose, inicialmente foram submetidos a uma purificação por precipitação com isopropanol 65% e etanol 60%. Posteriormente notamos que esta fase de purificação fazia-se por desnecessária, bastando apenas adicionar 1µl da reação de PCR diretamente na reação de seqüenciamento. É importante salientar que para proceder este seqüenciamento sem a purificação é necessário utilizar o mínimo possível de iniciador na reação de PCR.

Tais produtos foram então seqüenciados nas plataformas ABI 3100, utilizando o Big Dye Terminator Kit v3 (Applied Biosystems), e ABI PRISM 377 utilizando o DYEnamic™ ET Terminator Cycle Sequencing Kit (Amersham). Para tal, foi utilizada uma reação de seqüenciamento iniciada com 2 min a 94°C, trinta ciclos ou mais de 10 seg a 94°C, 10 seg a 50°C e 4 min a 60°C. Finalizando a reação, utilizou-se 20 min a 72°C.

As concentrações utilizadas na reação de seqüenciamento no ABI 3100 eram de 3.2 pmol de primer, 2 µl de tampão "save money", 1 µl de "Big Dye Terminator", 1 µl do produto da reação de PCR em um volume final de 10 µl. Para as reações no ABI PRISM 377, foram utilizados 5 pmol de iniciador, 2 µl do mix em um volume final de 10 µl.

3.4.3 Análise das seqüências

Após a obtenção das seqüências dos indivíduos, procederam-se buscas utilizando o programa BLAST (ALTSCHUL *et al*., 1997) para garantir a correta identificação dos genes estudados. O BLAST ("Basic Local Alignment Search Tool") é

um programa para análise de similaridade de seqüências, desenvolvido pelo NCBI ("National Center for Biotechnology Information"), servindo como instrumento de análise de identificação de genes. O BLAST pode executar uma busca de seqüências em bancos de dados públicos de DNA em poucos segundos fornecendo indicadores para a inferência funcional de genes desconhecidos. Posteriormente, analisou-se a qualidade dos dados brutos e identificou-se os polimorfismos utilizando-se o programa SeqScape v2.1 ("Applied Biosystems") (Figura 12).

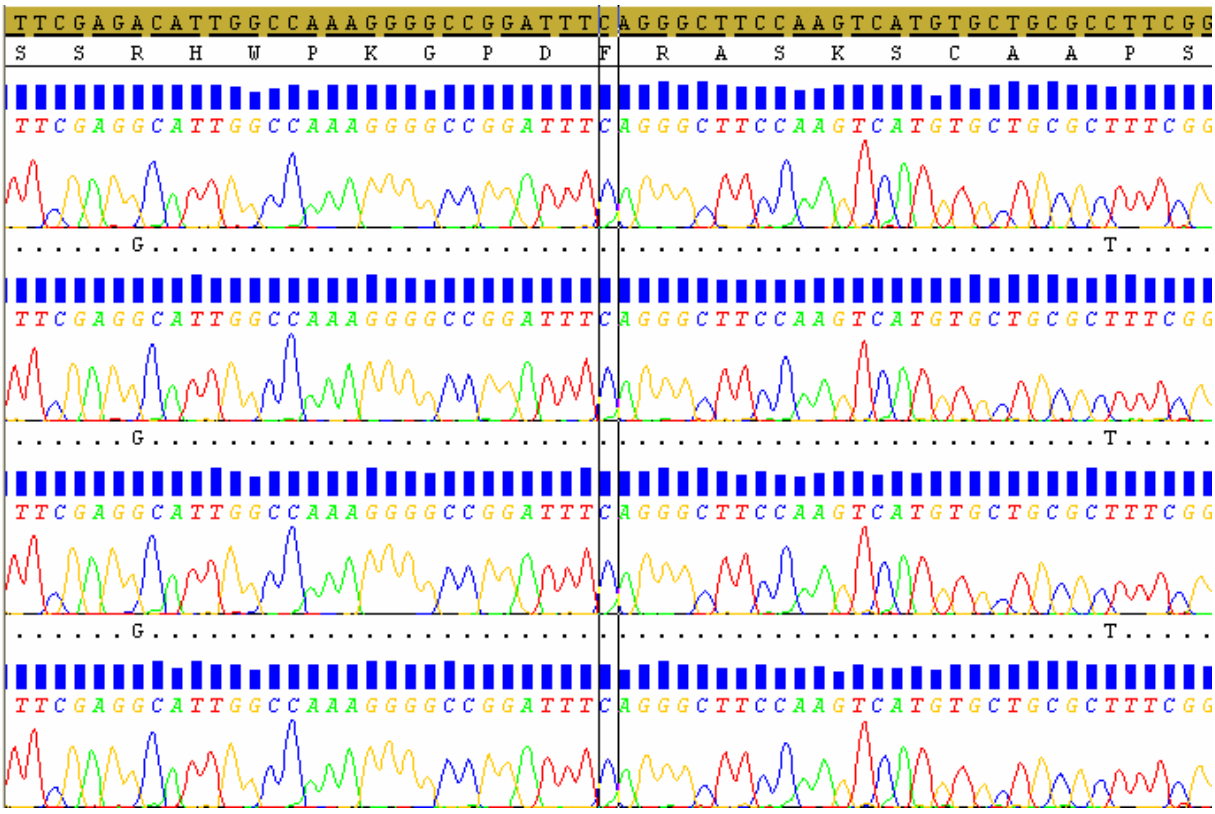


Figura 12: Análise dos dados pelo programa SeqScape v2.1. As barras sinalizam a qualidade da seqüência dada pelo programa Phred (EWING e GREEN, 1998).

Após análise das seqüências pelo software SeqScape v2.1 os alinhamentos foram inspecionados visualmente. As posições polimórficas foram utilizadas para gerar os arquivos de entrada nos programas LDA (DING *et al*., 2003) e Haploview (referência), respectivamente para análise de diversidade nucleotídica e análise de desequilíbrio de

ligação. Para análise de desequilíbrio de ligação foram utilizados somente aqueles SNPs com alelo de menor freqüência igual ou superior a 5%.

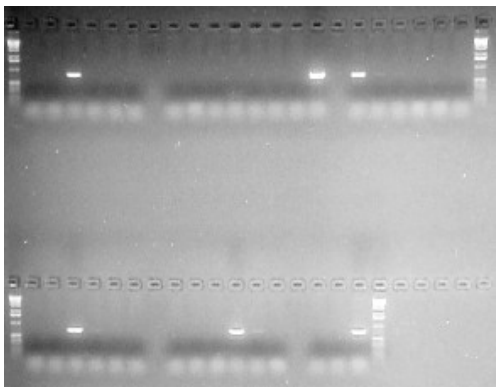
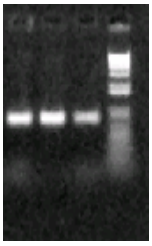
4 RESULTADOS

De forma a se iniciar uma busca por genes da via de lignificação para *E. grandis*, foi realizada a seleção de clones genômicos que contivessem os genes *CAD* e *COMT* na biblioteca de BACs do Genolyptus. Posteriormente, foi iniciado o estudo da diversidade nucleotídica dos mesmos genes de forma a identificar polimorfismos que possam potencialmente ser utilizados como marcadores do tipo SNP em estudos de associação.

4.1 Triagem dos clones de BAC

Visando-se identificar clones genômicos dos genes *CAD2* e *COMT2*, adotou-se a estratégia de criação de supergrupos para a e identificação dos genes de interesse.

Através do processo de triagem de clones de BAC descrito na metodologia (ver em 3.3), foi possível detectar vários clones que contêm os genes de nosso interesse. Este processo de triagem por etapas está ilustrado na Figura 13 e seu efeito líquido é identificar um ou mais clones de BAC que contém a seqüência desejada.

A)  B) 

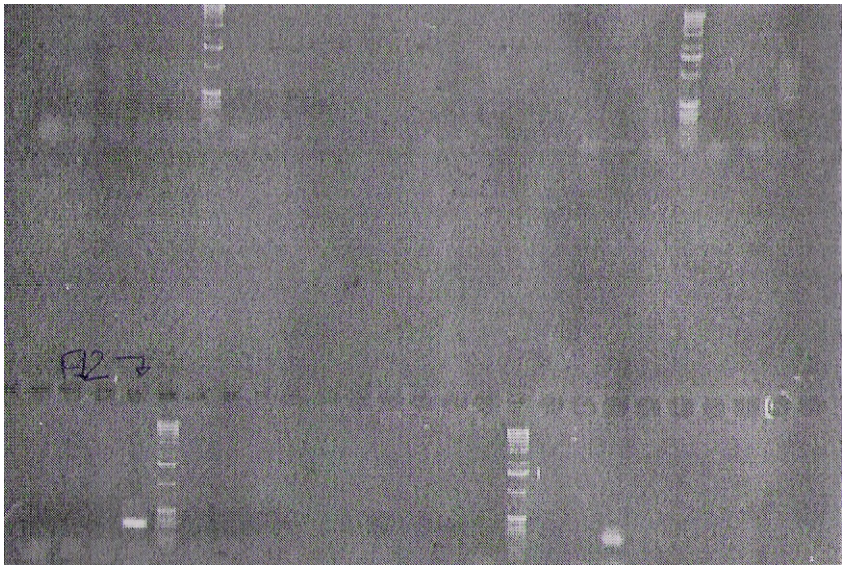
C) 

Figura 13: Triagem dos clones de BAC. A: Análise dos supergrupos para verificar quais tem o gene *COMT2*; em B observamos uma PCR confirmando os grupos positivos para *COMT2*; e em C temos a identificação de um clone de *CAD* (indicação à esquerda) através da PCR de uma placa derivada de um grupo positivo. A última banda deste gel é o controle positivo (indicação à direita).

Para cada um dos dois genes estudados, foram detectados pelo menos dois clones possuindo o fragmento de interesse (Tabela 2). Cada um destes clones, assim como os grupos e supergrupos, foi confirmado por pelo menos três PCRs distintas, além de uma reconfirmação por uma nova minipreparação (EMBRAPA, PLATAFORMA DE SEQUENCIAMENTO DE DNA, com modificações) do clone analisado.

Tabela 2: Triagem dos grupos e supergrupos de clones.

Genes	Super grupo	Grupo / Placas positivas	Clones
CAD	A2	8	C1,B3
	H3	182	A1
COMT	C2	33	B5
	D3	134	-
	E2	55	D7
	G1	75	-
	H2	94	-

4.2 Clone com o gene completo de *CAD2*

4.2.1 Montagem do gene de *CAD2*

Utilizando a identificação do clone B3 realizada na triagem (Tabela 2), contendo o gene *CAD2*, partimos para seu sequenciamento completo utilizando a estratégia do "shotgun". Após a fragmentação mecânica dos 30 Kb do inserto do clone, foi criada uma sub-biblioteca e, a partir desta, seqüenciadas 6 placas de contendo 96 clones cada, resultando em 768 leituras válidas de seqüência de forma a se obter uma cobertura de dez vezes o inserto. A montagem das seqüências foi realizada com o programa CAP3 (HUANG e MADAN, 1999).

Durante a montagem do BAC, apesar da grande cobertura (10 vezes o inserto original), não foi possível montar completamente as 30 Kb do inserto. Entretanto, o maior contig obtido possuía 9.785 bases. Através de buscas por similaridade utilizando o programa BLAST foi possível identificar que este continha o gene *CAD2* completo, cuja seqüência completa encontra-se no apêndice (9.1).

4.2.2 A estrutura gênica de CAD2

Com a seqüência completa do gene *CAD2* procedeu-se uma busca no banco de dados de ESTs do Genolyptus a fim de se identificar seqüências expressas de mRNAs referentes a este gene. Este tipo de análise, além de auxiliar na elucidação da estrutura gênica (identificação de exons e introns), ao mesmo tempo permite realizar inferências sobre o padrão de expressão deste gene em relação às diversas bibliotecas de cDNA disponíveis.

Inicialmente foi feita uma busca por similaridade do fragmento genômico contra o banco de seqüências consenso de ESTs utilizando o programa BLASTN. Um consenso de EST contendo 1.500 bases (CL12Contig1) foi recuperado por fornecer a maior similaridade em relação ao clone genômico de *CAD2*.

De posse desta seqüência consenso de mRNA, procedeu-se o alinhamento entre o segmento genômico (BAC) e a mesma utilizando o programa est2genome pertencente ao pacote EMBOSS (RICE *et al*., 2000). Um fragmento deste alinhamento encontra-se na Figura 14.

```
BAC_CAD             3816 CTACCCGGCAACTTCCCACTACGATAAGCAGCAAGTCTACGGCTCTGTCG   3865
                         |||||||||||||||||||||||||||||||||||||||||||||||||||
consenso_mRNA_CAD     57 CTACCCGGCAACTTCCCACTACGATAAGCAGCAAGTCTACGGCTCTGTCG    106

BAC_CAD             3866 AATCTCTCTCCGAGCACCACTTTGAAAAAAGCTTGGATCTTTGAGCAAAA   3915
                         |||||||||||||||||||||||||||||||||||||||||||||||||||
consenso_mRNA_CAD    107 AATCTCTCTCCGAGCACCACTTTGAAAAAAGCTTGGATCTTTGAGCAAAA    156

BAC_CAD             3916 ATGGGCAGTCTTGAGAAGGAGAGGACTACCACGGGTTGGGCTGCAAGGGA   3965
                         ||||||||||||||||||||||||| |||||||||||||||||||||||||
consenso_mRNA_CAD    157 ATGGGCAGTCTTGAGAAGGAGAGGACCACCACGGGTTGGGCTGCAAGGGA    206

BAC_CAD             3966 CCCGTCTGGCGTTCTCTCTCCTTACACTTATAGCCTCAGgtaga.....t   4004
                         |||||||||||||||||||||||||||||||||||||>>>>> 194 >
consenso_mRNA_CAD    207 CCCGTCTGGCGTTCTCTCTCCTTACACTTATAGCCTCAG...........    245

BAC_CAD             4004 gcagAAACACGGGACCAGAAGATCTTTACATCAAGGTGTTGAGCTGCGGA   4244
                         >>>>|||||||||||||||||||||||||||||||||||||||||||||||||
consenso_mRNA_CAD    245 ....AAACACGGGACCAGAAGATCTTTACATCAAGGTGTTGAGCTGCGGG    291

BAC_CAD             4245 ATTTGCCACAGTGACATTCACCAGATCAAGAATGATCTTGGCATGTCCCA   4294
                         |||||||||||||||||||||||||||||||||||||||||||||||||||
consenso_mRNA_CAD    292 ATTTGCCACAGTGACATTCACCAGATCAAGAATGATCTTGGCATGTCCCA    341

BAC_CAD             4295 CTACCCTATGGTTCCTGGgtagg....ttcagGCATGAAGTGGTGGGTGA   4416
                         ||||||||||||||||>>>>> 86 >>>>>||||||||||||||||||||
consenso_mRNA_CAD    342 CTACCCTATGGTTCCTGG.............GCATGAAGTGGTGGGTGA    377
```

Figura 14: Alinhamento de sequência resultante do programa est2genome para o gene cad entre o segmento genômico derivado do BAC (BAC_CAD) e o mRNA consenso (consenso_mRNA_CAD) inferido pelas ESTs do projeto Genolyptus. As regiões não alinhadas do mRNA (entre os símbolos >>>>>) representam os introns.

Primeiramente, observa-se que o consenso das ESTs realmente refere-se ao clone genômico obtido. Além disso, este tipo de análise permite a correta identificação das coordenadas dos introns na seqüência genômica (identificada por BAC_CAD), e resulta na estrutura gênica representada na Figura 15.

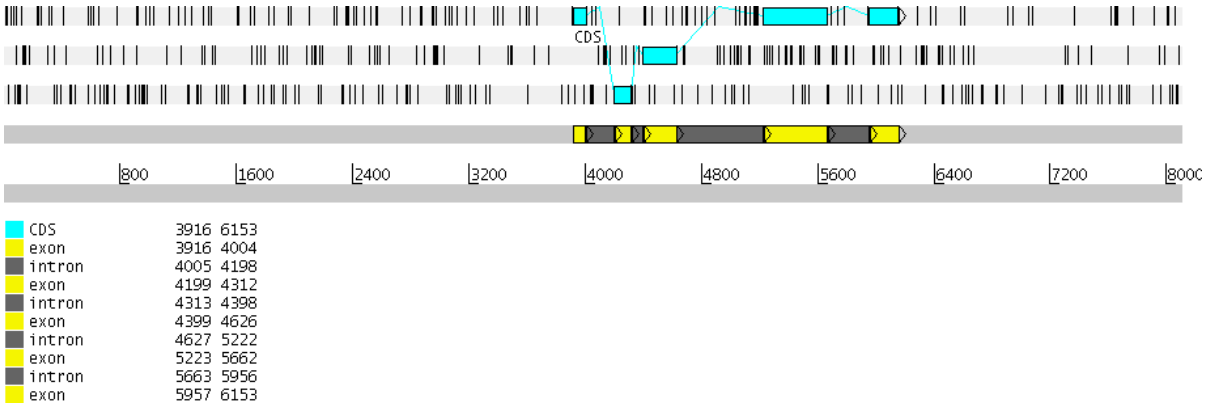


Figura 15: Representação esquemática da estrutura gênica de *CAD2* de *E. grandis* no contexto do contig obtido pelo seqüênciamento do clone de BAC para este gene. A imagem foi criada pelo programa Artemis (RUTHERFORD et al., 2000).

Verifica-se que o gene possui 5 exons que codificam uma proteína de 356 aminoácidos, exatamente como observado para os ortólogos deste gene em *E. gunnii* e *E. saligna* que se encontram depositados no GenBank com números de acesso, X75480 e AF294793, respectivamente.

Outra importante observação é que se dispõe da região promotora deste gene (Figura 16), que notadamente tem grande influência sobre a regulação da expressão gênica. Recentemente, foi demonstrado que os promotores dos genes *CAD* e *CCR* (Cinamoil CoA Redutase) contêm regiões regulatórias em *cis* que são sítios de ligação para o fator de transcrição EgMYB2 (GOICOECHEA *et al*., 2005). No mesmo estudo foi possível identificar em *E. gunnii* a porção da seqüência reconhecida pelo fator de transcrição. Esta é demonstrada abaixo, sendo a numeração +1 referente ao ponto de iniciação da transcrição:

-203 TGACCCTTAACCCACCCCACTGGTTCACCTACCGCACCTCGGGTTAGGTATTGC -129

ACTGGGAATTGGGTGGGGTGACCAAGTGGATGGCGTGGAGCCCAATCCATAACG

Figura 16: Sítio reconhecido pelo fator de transcrição EgMYB2 no gene CAD2.

Encontram-se sublinhadas na Figura 16 as regiões conhecidas como elementos AC (pela prevalência das bases adenina e citosina) e que também são encontradas na região promotora de outros genes da via de lignificação. Buscou-se localizar este motivo na seqüência de *E. grandis* e constatou-se uma conservação de 100% nesta região, indicando a sua relevância funcional.

4.2.3 Similaridade entre CAD2 de eucaliptos

Diante da indicação de que as seqüências são conservadas na região não-codificadora questionou-se qual seria a extensão de similaridade global entre genes ortólogos de eucalipto. Mais uma vez utilizou-se a seqüência de *CAD2* de *E. saligna* (AF294793) e procedeu-se um alinhamento local (ver em 9.2) entre nucleotídeos contra a seqüência genômica de *E. grandis*. Obteve-se um grau de identidade de 99,2% ressaltando ainda mais o fato de que existe uma conservação muito grande ao longo de todo o gene *CAD2*, mesmo em espécies diferentes.

Um outro ponto a ser notado é que a seqüência de *E. saligna* refere-se a cópia denominada *CAD2*. Ao se alinhar o gene da cópia *CAD1* de *E. gunnii* (X75480) e a seqüência de *E. grandis* (ver em 9.3) obtivemos uma similaridade entre as seqüências de 67,9% sendo que tal valor é decorrente principalmente de divergências nas regiões de introns. Na região exônica, a similaridade é muito alta, fazendo com que os indivíduos sejam diferenciados pela relevante presença de alguns "indels" nos primeiros nucleotídeos do intron 4. No caso do alinhamento entre *CAD1* (*E. gunnii*) e *CAD2* (*E. saligna*) (ver em 9.4), identificamos novamente a presença dos "indels", fortalecendo a nossa identificação do clone completo de *E. grandis* como sendo de *CAD2*. Na região

codificante, a similaridade entre as seqüências chega a 96,6%, o que é bastante expressivo, considerando que se tem uma comparação de nucleotídeos.

4.2.4 Análise de expressão do gene *CAD2*

A biblioteca de ESTs do Genolyptus fornece também informação qualitativa sobre o nível relativo de expressão deste gene. O grupo de ESTs do gene *CAD2* contém 55 seqüências oriundas de diversas bibliotecas, que por sua vez foram construídas a partir de diferentes tecidos e espécies. Visto que as ESTs foram geradas tanto para a extremidade 5' quanto para a 3', foi possível montar o consenso de 1.500 bases (utilizado nas análises anteriores), que representaria o mensageiro completo para este gene, com todo o quadro de leitura representado. A Figura 17 ilustra o processo de montagem do grupo de ESTs, demonstrando a cobertura total do transcrito por diversas ESTs. Nesta pode-se observar que ocorreu uma distribuição de seqüências de ESTs ao longo do mRNA, o que viabilizou sua montagem por completo.

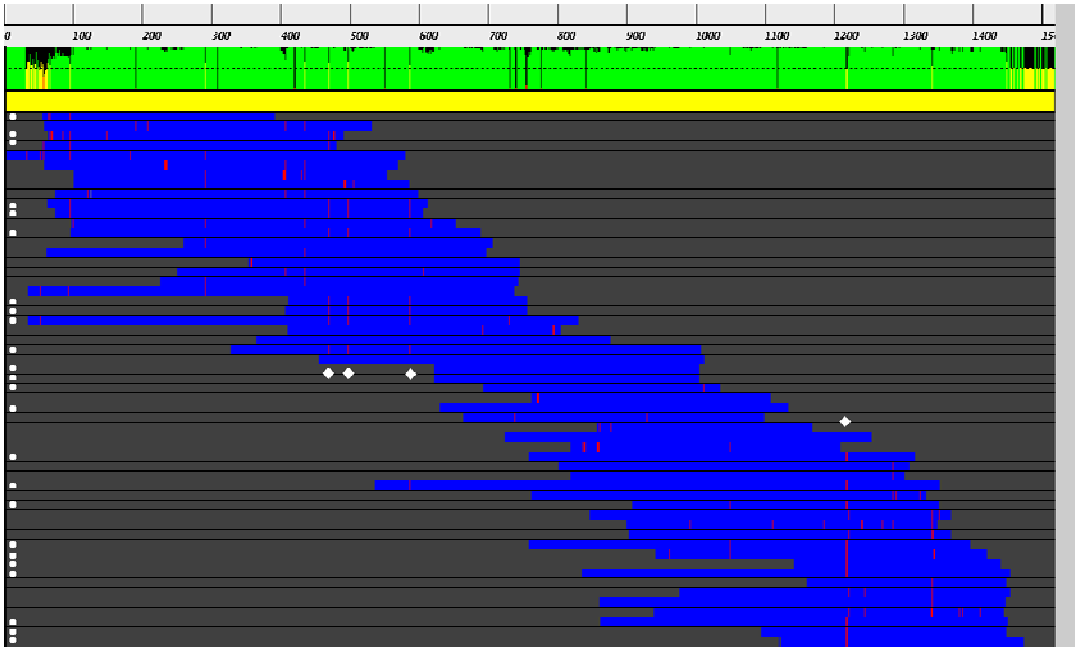


Figura 17: Representação esquemática da montagem das ESTs do projeto Genolyptus que foram identificadas com sendo do gene *CAD*. As linhas verticais nas seqüências representam bases polimórficas em relação ao consenso.

4.2.5 Análise de expressão tecido-específica do gene *CAD2*

No que tange ao nível de expressão do mRNA de *CAD2* podem ser extraídas algumas informações a partir da contagem de ESTs amostradas nas diversas bibliotecas analisadas. A Tabela 3 mostra a distribuição de ESTs do gene *CAD2* entre as bibliotecas de cDNA do projeto Genolyptus. Pode-se perceber que existe uma maior representação das ESTs provenientes de xilema quando comparado a outros tecidos.

Tabela 3: Número de ocorrências de ESTs de CAD2 identificadas nas bibliotecas de cDNA do projeto Genolyptus.

Biblioteca	Espécie	Tecido	Total de ESTs da biblioteca	ESTs de *CAD*
GL-XY	*E. globulus*	xilema	13549	25
PE-XY	*E. pelitta*	xilema	8380	7
UR-XY	*E. urophylla*	xilema	6624	12
GR-XY	*E. grandis*	xilema	641	1
SP-FX	*Várias*	floema	10434	2
GR-TS	*E. grandis*	plântulas	10275	3
GR-ML	*E. grandis*	folhas	5399	3
GR-SE	*E. grandis*	plântulas	10296	1
GR-PU	*E. grandis*	plântulas	4133	1
total			69731	55

Para verificar se existe uma diferença significativa de expressão deste gene no xilema em relação a outros tecidos, foi montada abaixo uma tabela de contingência e realizado um teste de qui-quadrado (χ^2) para verificar a hipótese nula de que não existe diferença de expressão de *CAD2* entre os tecidos.

	xilema	não-xilema
ESTs de *CAD2*	45	10
outras ESTs	29149	40527

$$\chi^2 = 36{,}1;\ P<0{,}0001$$

Pelo valor obtido no teste acima, rejeita-se a hipótese nula e conclui-se que este gene é significativamente mais expresso no xilema. Foram testadas outras hipóteses para comparação entre bibliotecas, tal como a expressão em xilema de *E. globulus* ser

maior que em xilema de *E. pellita*, sendo que nenhuma destas resultou em diferenças significativas que indicassem expressão diferencial.

1.2.6 Análise de variabilidade de seqüências a partir das ESTs de CAD2

Além de exibir a consistência do grupo em termos de identidade das seqüências que o compõe, a Figura 17 também permite que se visualize a presença de polimorfismos intra e inter-específicos no gene *CAD2*. Deve-se ressaltar que as ESTs são seqüenciadas apenas uma vez, e portanto, existe uma maior probabilidade de incidência de erros de seqüenciamento, principalmente nas extremidades das leituras. Locais onde existe um polimorfismo não compartilhado por nenhuma outra seqüência têm alta probabilidade de serem artefatos. Assim, quando encontrados, os polimorfismos em ESTs devem ser analisados com reservas, mas isso não impede que este tipo de dado seja utilizado para a identificação de variantes alélicos, visto que as regiões de menor qualidade foram previamente removidas das extremidades.

De maneira geral observa-se na Figura 17 e Figura 18 que mesmo com seqüências das 4 espécies existe uma grande conservação por toda a extensão do transcrito reconstruído. Não obstante, alguns padrões de substituição emergem através de uma inspeção visual, particularmente nas colunas marcadas com o símbolo ♦. Nestas, as seqüências com o marcador vertical têm a mesma alteração de base quando comparadas à maioria das seqüências. Mais especificamente, estes padrões de polimorfismos estão presentes apenas em seqüências oriundas de *E. globulus*, formando um haplótipo reconhecível para esta espécie. Um detalhamento da região onde este haplótipo é observável pode ser visualizado na Figura 18. As colunas marcadas com o símbolo ♦ correspondem as posições 469, 498 e 587 do mensageiro a partir do codon de iniciação.

Nestas, todas as seqüências com bases em cor diferente pertencem a *E. globulus*, sendo as outras pertencentes as demais espécies estudadas. Não foi possível identificar nenhum outro haplótipo espécie-específico nesta análise.

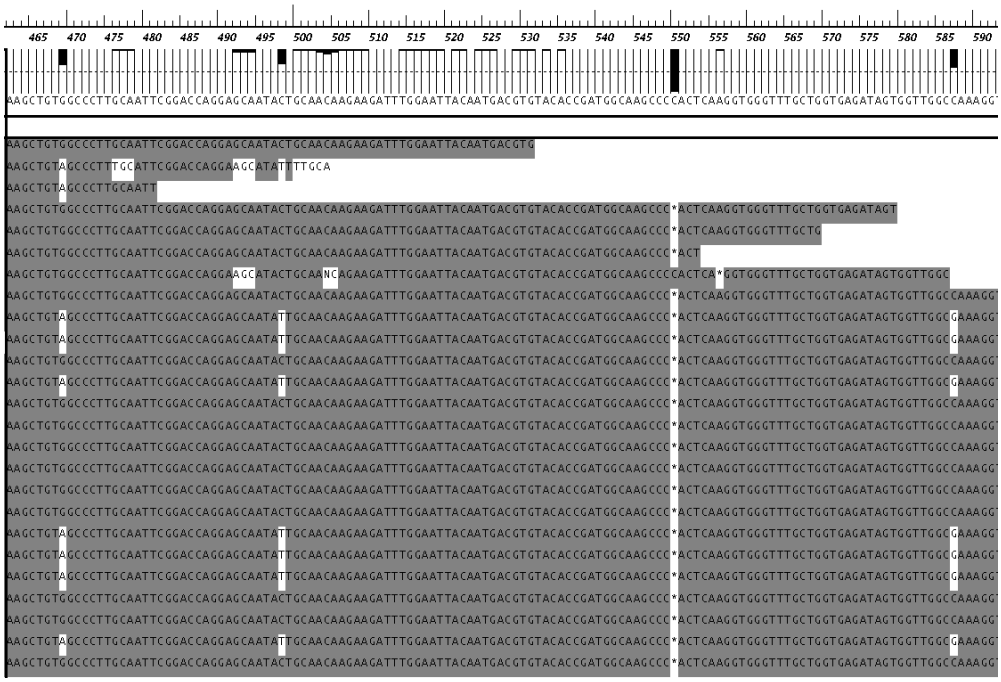


Figura 18: Detalhamento de uma região do alinhamento de ESTs do projeto Genolyptus para o gene *CAD2*. As bases coloridas em cor de fundo diferente são polimórficas.

4.2.7 Análise da seqüência de aminoácidos da enzima *CAD2* de *E. grandis*

De acordo com a estrutura gênica determinada anteriormente para o gene *CAD2*, realizou-se a tradução conceitual de seus nucleotídeos obtendo a sua seqüência completa de 356 aminoácidos. Inicialmente, desejou-se mensurar o grau de similaridade entre esta nova seqüência em relação a seqüências do mesmo gene, porém de outras espécies de eucalipto depositadas no GenBank. No total foram encontradas cinco seqüências de outras quatro espécies. Alinhamentos globais, em termos de aminoácidos (diferentemente de alinhamentos de DNA do item 4.2.3), foram realizados para todos os pares de seqüências utilizando o programa “needle” do pacote EMBOSS, resultando em

uma matriz de distância expressa em porcentagem de identidade de seqüência. Os resultados apresentados na Tabela 4, demonstram o elevado grau de similaridade entre as seqüências deste gene, variando de 99,72% entre *CAD2* de *E. gunnii* e *CAD* de *E. globulus*, até 97,73% entre *CAD1* de *E. gunnii* e *CAD* de *E. botryoides*. Este nível de identidade reflete que existem, no pior cenário, oito alterações de aminoácidos ao longo dos 356 resíduos da proteína. Apesar de não existirem estudos bioquímicos para a análise dos parâmetros cinéticos, pode-se projetar que estes não sejam significativamente diferentes para a *CAD* entre as espécies de *Eucalyptus*.

Tabela 4: Percentual de identidade entre seqüências protéicas de *CAD*. Os códigos oriundos do banco SwissProt são: CADH_EUCGR (E. grandis), CADH_EUCSA (E. saligna), CADH_EUCGL (E. globulus), CAD1_EUCGU e CAD2_EUCGU (E. gunnii) e CADH_EUCBO (E. botryoides).

	CADH_EUCGR	CADH_EUCSA	CADH_EUCGL	CAD2_EUCGU	CAD1_EUCGU
CADH_EUCGR					
CADH_EUCSA	99,44				
CADH_EUCGL	99,44	98,88			
CAD2_EUCGU	99,16	98,61	99,72		
CAD1_EUCGU	98,59	98,02	99,15	98,87	
CADH_EUCBO	99,15	99,15	98,59	98,31	97,73

Para contextualizar a relevância da conservação das seqüências protéicas de CAD2 de eucalipto, realizou-se uma análise filogenética. Para proceder tal análise, foram coletadas seqüências de CAD de distintas espécies de plantas, e também a de uma álcool desidrogenase da bactéria *Sinorhizobium meliloti* (CAC47271), que serve como seqüência fora do grupo ("outgroup") utilizada para enraizar a árvore filogenética.

A árvore filogenética resultante (Figura 19) revela que as seqüências de eucalipto são realmente bastante conservadas formando um grupo monofilético bem destacado.

Este fato reforça a visão de que o gene CAD2 sofre uma intensa pressão de seleção, ou que o evento de especiação em *eucalyptus* foi recente. Em *Arabidopsis thaliana* não ocorre o mesmo, visto que os parálogos (CADH1_ARATH e CADH2_ARATH) aparecem em ramos diferentes, resultado esperado para loci distintos que divergem aleatoriamente (parálogos). Este comportamento contrasta marcadamente com as cópias de *E. gunnii* (CAD1_EUCGU e CAD2_EUCGU).

Outras observações relevantes são que as seqüências de monocotiledôneas também formam um grupo monofilético, assim como as seqüências de gimnospermas, indicando gargalos evolutivos específicos.

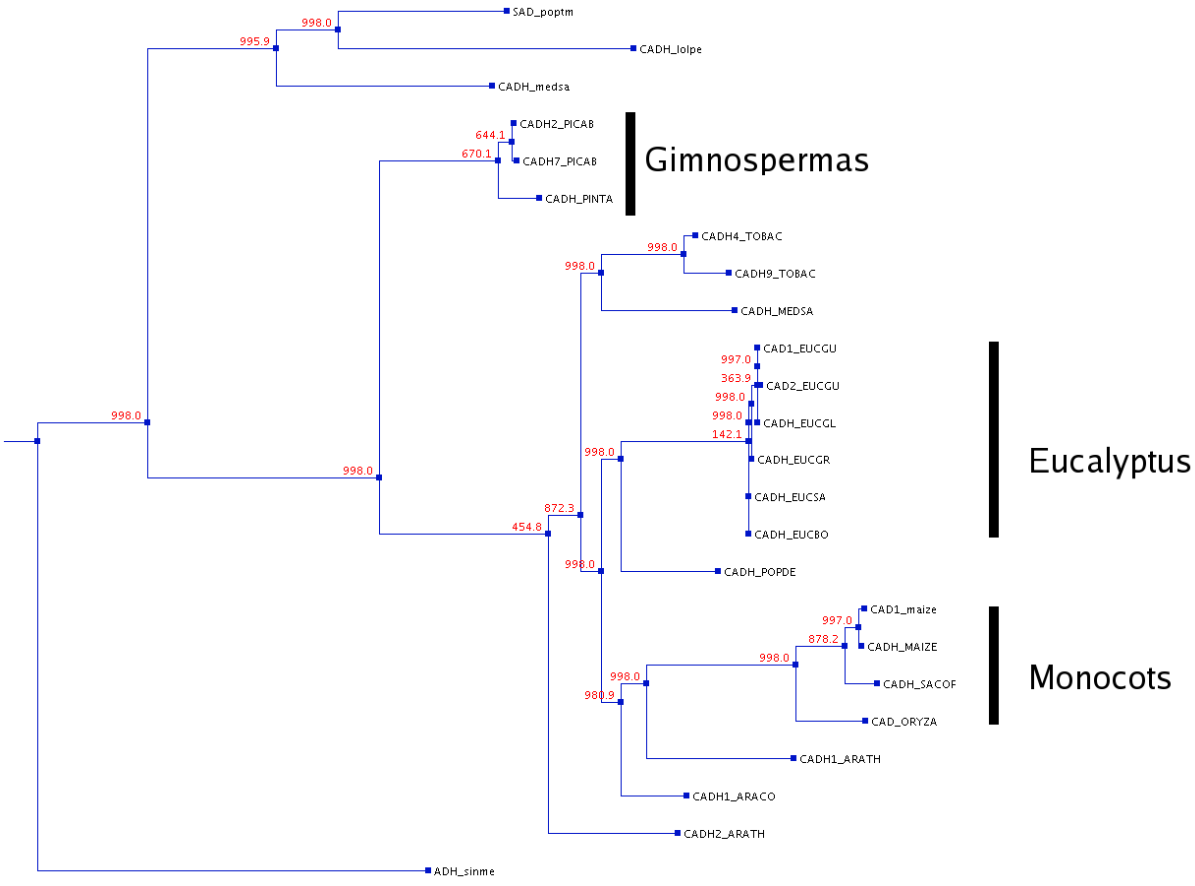


Figura 19: Árvore filogenética construída pelo programa PHYLIP para seqüências de CAD de diversas plantas.

A seqüência da sinapil álcool desidrogenase de *Populus tremuloides* (SAD_poptm; AF273256) que se encontra em um grupo distinto no topo da árvore, representa uma álcool desidrogenase que se acredita atuar apenas sobre o álcool sinapílico (KIN *et al*., 2004). Esta foi incluída para salientar a classificação funcional da seqüência de *E. grandis* como sendo realmente CAD.

4.3 Análise da variabilidade nucleotídica de genes da via de lignificação

Após proceder o seqüenciamento completo de um gene chave para a via de lignificação, o *CAD2*, passamos então a avaliar sua diversidade nucleotídica entre indivíduos de três espécies de eucalipto. Em função do tempo e por dificuldades técnicas no sequenciamento, apenas uma região foi estudada.

4.4 Análise da variabilidade nucleotídica do gene CAD2

Para as análises de polimorfismos do gene *CAD2*, foram utilizadas no total 78 amostras de DNA distribuídas da seguinte forma: 19 da classe GR (ver em 3.1.2) e 46 da classe GRR de *E. grandis* (total de 65 indivíduos), 6 de *E. urophylla* e 7 de *E. globulus*. Para cada um destes indivíduos, foi seqüenciada uma região de 358 nucleotídeos sendo que destes, 55 pertencem ao intron 4, 194 nucleotídeos ao exon 5, e 109 a região 3' UTR. Os iniciadores usados para amplificar e sequenciar esta região encontram-se listados na seção 3.4.1.

A análise de 358 nucleotídeos nas três espécies levou a identificação de 18 sítios polimórficos (Figura 20), sendo que um destes (4808), é polimórfico em relação apenas à seqüência de referência provinda de *E. gunnii*, não apresentando polimorfismo dentro

dos indivíduos analisados neste trabalho e não sendo considerado para as análises posteriores.

	4666	4770	4814	4820	4822	4830	4846	4886	4887	4888	4889	4904	4934	4943	4944	4946	4966
E.gunnii	T	C	G	C	T	A	T	C	T	C	T	C	A	G	C	A	T
GR22	.	.	.	.	.	.	G	.	.	.	.	.	.	A	T	.	.
GR20	.	.	.	T	.	.	G	.	.	.	.	.	G	A	T	.	.
GR18	.	.	.	T	.	.	G	.	.	.	.	.	G	A	T	.	.
GR17	.	.	.	.	.	.	G	.	.	.	.	.	.	A	T	.	.
GR10	.	T	.	T	.	.	G	.	.	.	.	.	.	A	T	.	.
GR09	.	Y	.	Y	.	.	G	.	.	.	.	.	.	A	T	.	.
GR01	.	Y	.	Y	.	.	G	.	.	.	.	.	.	A	T	.	.
GR26	.	.	.	T	.	.	G	.	.	.	.	.	G	A	T	.	.
GR24	.	.	.	T	C	.	.	.	.	.	.	.	.	A	T	.	.
GR27	.	.	.	.	.	.	G	.	.	.	.	.	.	A	T	.	.
GR25	-	.	.	T	.	.	G	.	.	.	.	.	G	A	T	.	.
GR16	.	Y	.	Y	.	.	G	.	.	.	.	.	.	A	T	.	.
GR08	.	.	.	.	.	.	G	.	.	.	.	.	.	A	T	.	.
GR19	.	.	.	T	.	.	G	.	.	.	.	.	G	A	T	.	.
GR28	.	.	.	.	.	.	G	.	.	.	.	.	.	A	T	.	.
GR46	.	.	.	T	.	.	G	.	.	.	.	.	R	A	T	.	.
GR43	.	T	.	T	.	.	G	.	.	.	.	.	.	A	T	.	.
GR40	.	Y	.	Y	.	.	G	.	.	.	.	.	.	A	T	.	.
GR02	.	.	.	.	.	.	G	.	.	.	.	.	.	A	T	.	.
GRR01	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR50	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR49	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR48	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR47	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	W	.
GRR46	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR45	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR44	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR43	.	.	.	Y	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR42	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR41	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR40	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR38	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR37	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR36	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	W	.
GRR35	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR33	.	.	.	Y	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR32	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.

Continuação...

	4666	4770	4814	4820	4822	4830	4846	4886	4887	4888	4889	4904	4934	4943	4944	4946	4966
E.gunnii	T	C	G	C	T	A	T	C	T	C	T	C	A	G	C	A	T
GRR31	.	.	.	Y	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR30	.	.	.	.	.	.	.	.	-	-	-	.	.	.	.	.	.
GRR29	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR28	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR27	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR26	.	.	.	.	.	.	.	.	-	-	-	.	.	.	.	W	.
GRR25	.	.	.	T	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR22	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	W	.
GRR21	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR20	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR18	.	.	.	Y	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR17	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR16	.	.	.	.	.	.	.	.	-	-	-	.	.	.	.	W	.
GRR15	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR14	.	.	.	.	.	R	G	.	.	.	.	.	.	.	.	.	.
GRR13	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR12	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR11	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR09	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR08	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR07	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR06	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR04	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR03	.	.	.	.	.	.	G	.	.	.	.	M	.	.	.	.	.
GRR02	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR39	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR34	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GRR23	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
UR07	.	.	.	.	.	.	G	.	.	.	.	.	.	A	T	.	.
UR05	.	.	.	T	.	.	G	T	.	.	.	.	.	A	T	.	-
UR09	.	.	.	T	.	.	G	T	.	.	.	.	.	A	T	.	.
UR03	.	.	R	.	.	.	G	.	.	.	.	.	.	A	T	.	.
UR02	.	.	.	Y	.	.	G	Y	.	.	.	.	.	A	T	.	.
UR01	.	.	.	.	.	.	G	.	.	.	.	.	.	A	T	.	.
GL09	.	.	.	.	.	.	.	.	-	-	-	.	.	.	.	.	.
GL07	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GL06	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GL05	.	.	.	Y	.	.	G	.	.	.	.	.	.	.	.	.	.
GL03	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GL02	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
GL01	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.

Figura 20: Sítios polimórficos observados na análise do gene *CAD2*. As setas indicam a posição de cada sítio.

4.4.1 Análise interespecífica do gene *CAD2*

No estudo conjunto das três espécies (*E. grandis*, *E. globulus* e *E. urophylla*), só foi detectado um polimorfismo no íntron 4, sendo esse uma indel (inserção/deleção) na posição 4666. Uma outra indel foi encontrada no nucleotídeo 4887 que se localiza no exon 5. Estes dois sítios polimórficos fazem parte do total de 17 identificados nos 358 nucleotídeos seqüenciados, caracterizando uma média de um polimorfismo a cada 21 nucleotídeos (interespecificamente) e 95,3% de similaridade entre os 78 indivíduos amostrados.

De um total de 15 substituições, 8 ocorrem em exon e 7 no 3'UTR. Nove são transições (modificação de uma base por outra de mesmas características químicas, purina-purina, pirimidina-pirimidina) e 6 são transversões (substituição entre bases de características distintas, purina - pirimidina). Dentre as substituições em exon, duas são silenciosas (não acarretam modificação do aminoácido) e seis não silenciosas (acarretam modificação do aminoácido). Dentre estas não silenciosas, uma gerou aminoácido de característica química similar (conservativa) ao modificar uma fenilalanina para uma valina, ambos não polares.

Por outro lado, cinco substituições foram não conservativas (substituição que gera um aminoácido com características químicas diferentes das do aminoácido inicial) sendo 4 transições e uma transversão. A transversão gerou a modificação de uma serina (polar sem carga) para um código de terminação (em GRR, sendo que este já está na posição do código de terminação do exon 5, não se tornando informativo). Já nas transições, houve a substituição de uma serina (polar sem carga) por uma leucina (apolar); de um triptofano (apolar) por uma arginina (polar positivo); de uma serina (polar sem carga) por uma fenilalanina (apolar); além da substituição de um triptofano (apolar) para gerar outro códon de terminação, no indivíduo 3 de *E. urophylla*. Este

código de terminação acarreta uma possível perda de 31 aminoácidos (distância até o código de terminação do exon 5), fato que pode ser crucial para a atividade desta enzima.

Estas modificações de aminoácidos são bastante pertinentes já que as não conservativas podem alterar a atividade da proteína, podendo até torná-la inviável funcionalmente. A modificação que gerou o códon de terminação também é bastante importante, já que a proteína seria truncada no seu carboxi-terminal e, dependendo do tamanho do truncamento, isto pode mudar bastante a atividade da enzima.

Apesar da incidência alta de polimorfismos, o índice geral de similaridade entre as espécies se manteve alto, e, para nossa surpresa, os indivíduos de *E. grandis* procedentes de Atherton (GR) apresentaram mais similaridade com os de *E. urophylla*, procedentes do Timor. Da mesma forma, os indivíduos de *E. grandis* procedentes de Pine Creek (GRR) se assemelharam mais com os indivíduos de *E. globulus* de Victoria.

Através de análises utilizando os programas Arlequin (COX, 2001), DNAsp (ROZAS, 1999), Phase (STEPHENS *et al*., 2001 e STEPHENS and DONNELLY, 2003) e Haploview (BARRETT *et al*., 2005) obtemos o valor de Pi que corresponde à diversidade nucleotídica dos indivíduos amostrados. Outro valor calculado é o de h que diz respeito ao número de haplótipos. Obtivemos como resultado para todos os indivíduos de *CAD2* um valor de Pi equivalente a 0,78% (Tabela 16), indicando uma alta diversidade nucleotídica em termos interespecíficos, principalmente comparando aos 0,02% encontrados em *Pinus sylvestris* (GARCÍA-GIL *et al*., 2003), 0,077%, 0,087% e 0,134% encontrados em bonobos, humanos e chimpanzés, respectivamente (YU *et al*., 2004). Por outro lado, não chega aos 2,93% da região intrônica do gene *ACL5* encontrado em *Arabidopsis*, diferentemente de sua região exônica, que apresenta 0,42% (YOSHIDA *et al*., 2003).

Analisando todos os indivíduos de *CAD2* em conjunto, conseguimos identificar 11 haplótipos (Tabela 5), uma quantidade muito maior que a observada quando analisado cada grupo isoladamente.

Tabela 5: Estrutura dos haplótipos observados em *CAD2* em uma análise interespecífica.

haplótipo	E(freq)
CCGCAATA	0.147404
CCGCAGCA	0.564103
CCGCAGCT	0.019231
CCTCAGCA	0.038462
CCTCAGCT	0.012821
CTGCAATA	0.006442
CTGCAGCA	0.044872
CTGCGATA	0.070513
CTGTAATA	0.032039
CTTCAATA	0.012821
TTGCAATA	0.051263

A partir de uma análise dos polimorfismos, conseguimos detectar variáveis níveis de desequilíbrio de ligação, como podemos ver na Tabela 6.

Tabela 6: Análise de desequilíbrio de ligação para todos indivíduos de *CAD2*.

Sítios		D'	r^2
sitio1	sitio2	1.0	0.083
sitio1	sitio3	1.0	0.037
sitio1	sitio4	1.0	0.01
sitio1	sitio5	1.0	0.01
sitio1	sitio6	1.0	0.083
sitio1	sitio7	1.0	0.083
sitio1	sitio8	1.0	0.022
sitio2	sitio3	0.389	0.068
sitio2	sitio4	1.0	0.083
sitio2	sitio5	1.0	0.083
sitio2	sitio6	0.633	0.401
sitio2	sitio7	0.633	0.401
sitio2	sitio8	1.0	0.267
sitio3	sitio4	1.0	0.037
sitio3	sitio5	1.0	0.037
sitio3	sitio6	0.389	0.068
sitio3	sitio7	0.389	0.068
sitio3	sitio8	0.312	0.058
sitio4	sitio5	1.0	0.01
sitio4	sitio6	1.0	0.083
sitio4	sitio7	1.0	0.083
sitio4	sitio8	1.0	0.022
sitio5	sitio6	1.0	0.083
sitio5	sitio7	1.0	0.083
sitio5	sitio8	1.0	0.022
sitio6	sitio7	1.0	1.0
sitio6	sitio8	1.0	0.267
sitio7	sitio8	1.0	0.267

4.4.2 Análise intraespecífica do gene *CAD2*

4.4.2.1 Análise de *E. urophylla*

Quanto a *E. urophylla*, os dados devem ser confirmados devido ao pequeno número de indivíduos amostrados (10 indivíduos). Foram observados 2 sítios polimórficos sendo os dois informativos. A partir destes, foram identificados 4 haplótipos (Tabela 7). Sua diversidade nucleotídica foi a maior dentre as análises das três espécies em *CAD2*,

(Pi=0,882%) (Tabela 16). Seu padrão de polimorfismos é bastante semelhante aos de *E. grandis* proveniente de Atherton (GR).

Tabela 7: Estrutura dos haplótipos observados em E. urophylla para CAD2 em uma análise intraespecífica.

haplótipo	E(freq)
CCGCAATA	0.582452
CCGTAATA	0.000882
CTGCAATA	0.000882
CTGTAATA	0.415785

Analisando quanto à presença de desequilíbrio de ligação, como pode ser observado na Tabela 8, não foi detectado desequilíbrio entre tais polimorfismos.

Tabela 8: Análise de desequilíbrio de ligação para *CAD2* entre os indivíduos de *E. urophylla*.

Sítios		D'	r^2
sitio2	sitio4	0.0	0.0

4.4.2.2 Análise de *E. globulus*

A análise de *E. globulus* indicou somente 2 sítios polimórficos, sendo esses singletons. Desta forma, apenas 3 haplótipos foram identificados (Tabela 9). A diversidade haplotípica de *E. globulus* em *CAD2* de 0,42, uma das menores detectadas neste trabalho. Entretanto, deve-se ressaltar a necessidade de um número maior de indivíduos para resultados mais confiáveis. De qualquer forma, foi observado um baixo número de polimorfismos intraespecíficos em *E. globulus*.

Tabela 9: Estrutura dos haplótipos observados em *E. globulus* para *CAD2* em uma análise intraespecífica.

haplótipo	E(freq)
CCTCAGCA	0.142857
CCGCAGCA	0.785714
CTGCAGCA	0.071429

Foram realizados estudos visando detectar desequilíbrio de ligação entre os polimorfismos detectados, resultando nos dados da Tabela 10. Tais dados indicam desequilíbrio de ligação presente.

Tabela 10: Análise de desequilíbrio de ligação para *CAD2* entre os indivíduos de *E. globulus*.

Sítios		D'	r^2
sitio2	sitio3	1.0	0.25

4.4.2.3 Análise de *E. grandis*

A análise dos dois grupos de *E. grandis* (GR e GRR) acabou por demonstrar uma diferença pequena, mas facilmente identificável quanto aos polimorfismos. Cada grupo possui seus próprios sítios polimórficos, sendo que em apenas dois casos (sítios 4820 e 4846) o polimorfismo foi compartilhado entre estes grupos. Em uma análise conjunta, foram observados 6 sítios polimórficos, sendo destes 2 "singletons" e 4 informativos. Foi obtido também um número de 10 haplótipos (Tabela 11). Já a diversidade nucleotídica desta comparação de grupos intraespecífica foi de 0,672%.

Tabela 11: Estrutura de haplótipos observados em uma análise conjunta dos indivíduos de *E. grandis*.

haplótipo	E(freq)
CCGCAATA	0.123077
CCGCAGCA	0.592308
CCGCAGCT	0.023077
CCTCAGCA	0.030769
CCTCAGCT	0.015385
CTGCAATA	0.007692
CTGCAGCA	0.046154
CTGCGATA	0.084615
CTTCAATA	0.015385
TTGCAATA	0.061538

Avaliando cada grupo de *E. grandis* separadamente, observamos que GR possui uma diversidade nucleotídica maior que GRR (P_i=0,612%), por outro lado, o número de haplótipos de GRR (Tabela 12 e Tabela 16) é o mesmo em relação a GR (h=5) (Tabela 14 e Tabela 16).

Tabela 12: Estrutura dos haplótipos observados em E. grandis (GRR) para CAD2 em uma análise intraespecífica.

haplótipo	E(freq)
CCGCAGCA	0.836957
CCGCAGCT	0.032609
CCTCAGCA	0.043478
CCTCAGCT	0.021739
CTGCAGCA	0.065217

Quanto a uma análise referente a presença de desequilíbrio de ligação, pudemos verificar baixos valores, assim como mostra a Tabela 13.

Tabela 13: Análise de desequilíbrio de ligação para *CAD2* entre o grupo GRR de *E. grandis*.

Sítios		D'	r^2
sitio2	sitio3	1.0	0.167
sitio2	sitio8	1.0	0.167
sitio3	sitio8	0.167	0.028

A quantidade de haplótipos observada em GR foi a mesma em relação a GRR. A Tabela 14 mostra os haplótipos observados.

Tabela 14: Estrutura dos haplótipos observados em *E. grandis* (GR) para *CAD2* em uma análise intraespecífica.

haplótipo	E(freq)
CCGCAATA	0.420816
CTGCAATA	0.026552
CTGCGATA	0.289474
CTTCAATA	0.052632
TTGCAATA	0.210290

Em uma análise de desequilíbrio de ligação, pudemos detectar baixos valores de r^2, apesar de altos valores de D'.

Tabela 15: Análise de desequilíbrio de ligação para *CAD2* entre o grupo GR de *E. grandis*.

Sítios		D'	r^2
Sitio1	sitio2	1.0	0.062
Sitio1	sitio3	1.0	0.062
Sitio1	sitio5	1.0	0.062
sitio2	sitio3	1.0	0.062
sitio2	sitio5	1.0	0.062
sitio3	sitio5	1.0	0.062

De forma a resumir os dados anteriormente relatados, a Tabela 16 os consolida de forma a facilitar uma comparação entre os grupos analisados.

Tabela 16: Dados resultantes da análise da diversidade nucleotídica de *CAD2*. Os grupos são identificados inicialmente pelo gene estudado (*CAD2*), e posteriormente pelo grupo (gr – GR; grr – GRR; gr_grr – análise conjunta destes dois grupos de *E. grandis*; gl – *E. globulus*; ur – *E. urophylla*; e todos – análise conjunta de todos os indivíduos analisados.

	Sítios polimórficos	Singletons	Sítios informativos	h	Pi
cad_gr	4	0	4	5	0,00612
cad_grr	5	3	2	7	0,00431
cad_gr_grr	6	2	4	8	0,00672
cad_gl	2	2	0	3	0,00164
cad_ur	2	0	2	3	0,00882
cad_todos	3	1	2	4	0,00781

4.5 Análise da variabilidade nucleotídica do gene *COMT2*

Para enriquecer o estudo de variabilidade nucleotídica em *Eucalyptus*, realizou-se o mesmo estudo para o gene *COMT2*. Para a análise do gene *COMT2*, foram utilizados 27 indivíduos de *E. globulus*, 25 de *E. urophylla*, 26 de GR (*E. grandis*) e 38 de GRR (*E. grandis*), totalizando 64 indivíduos de *E. grandis*. Para todos estes, foram seqüenciados 623 nucleotídeos sendo que destes, 388 pertencem a introns e 235 a exons. Esta região

encontra-se entre os iniciadores G-COMT F e G-COMT R descritos na Tabela 1. Deve-se ressaltar que ainda não existe nenhuma seqüência genômica completa para *COMT2* no gênero *Eucalyptus*, o que limitou as regiões amostradas.

Primeiramente, como o gene *COMT* possui três cópias conhecidas, foi realizado um alinhamento entre a seqüência de um indivíduo aleatório de nossas amostras e seqüências de mRNA de *COMT* de *E. globulus* disponíveis no GenBank para os parálogos COMT1 (gi/5739364) e COMT2 (gi/5739366). Os resultados encontram-se na Figura 21 e permitem afirmar que a região amplificada em nossos estudos refere-se ao gene *COMT2*.

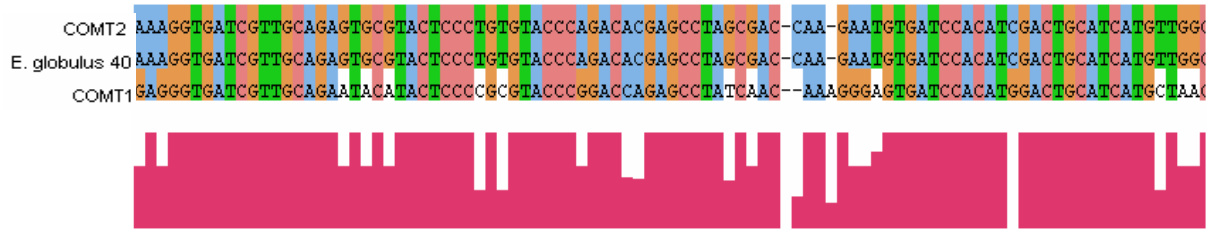


Figura 21: Alinhamento para identificação da cópia estudada para o gene *COMT2*. A sequência utilizada para este alinhamento de forma a ilustrar *COMT2* em *E. globulus* foi gi[5739366] enquanto que para *COMT1* foi gi[5739364], também de *E. globulus*. O indivíduo *E globulus* 40 demonstrado foi colhido aleatoriamente dentre os analisados neste estudo.

Foram observados 28 sítios polimórficos ao longo da região analisada do gene *COMT2*. Assim como era de se esperar, a maior parte destes polimorfismos encontra-se nas regiões intrônicas, assim como pode ser observado na Figura 22. Entretanto, existe uma razoável incidência em exons. Na Figura 22 são mostrados todos os polimorfismos, independente de sua frequência. A partir destes 28 sítios, selecionamos para análises haplotípicas e de desequilíbrio de ligação somente aqueles com polimorfismos com freqüência igual ou acima de 5% dentro dos indivíduos analisados, resultando então em 13 sítios polimórficos, os quais podem ser analisados na Figura 23.

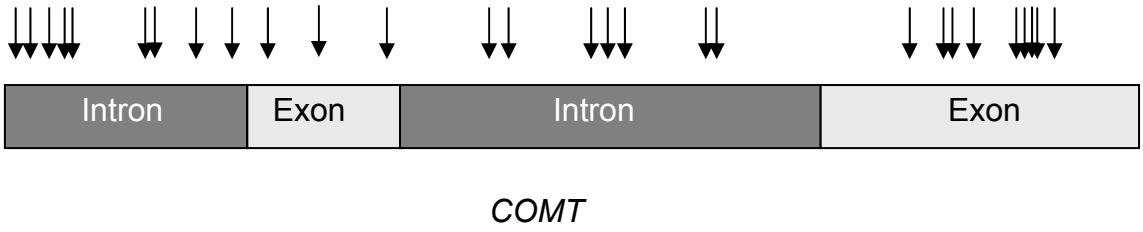


Figura 22: Distribuição dos polimorfismos observados na região analisada do gene *COMT2*.

	52	87	186	303	308	322	357	359	507	538	576	585	600
	T	A	T	A	G	C	C	C	C	A	C	C	C
GR01	.	.	.	.	.	.	.	.	.	G	.	.	T
GR02	.	.	.	.	T	A	.	.	T	.	.	.	.
GR03	.	.	.	.	.	.	.	.	.	G	.	.	T
GR05	.	.	.	.	T	A	.	.	T	.	.	.	.
GR06	.	.	C	.	.	.	.	.	.	G	.	.	.
GR08	.	.	.	.	.	.	.	.	.	G	.	.	Y
GR10	.	.	.	.	.	.	.	.	.	G	.	.	Y
GR11	.	.	.	.	.	.	.	.	.	G	.	.	Y
GR12	.	.	.	.	.	.	.	.	.	G	.	Y	Y
GR13	.	.	.	.	K	M	.	.	Y	R	.	.	Y
GR14	.	.	.	.	.	.	.	.	.	G	.	.	T
GR15	.	.	.	.	.	.	.	.	.	G	.	.	.
GR16	.	.	.	.	.	.	.	.	.	G	.	.	Y
GR17	.	.	.	.	.	.	.	.	.	G	.	.	.
GR22	.	.	.	.	K	M	.	.	Y	R	.	.	.
GR24	.	.	.	.	.	.	.	.	.	G	.	.	T
GR25	.	.	.	.	.	.	.	.	.	G	.	.	T
GR26	.	.	.	.	.	.	.	.	.	G	.	.	.
GR27	.	.	.	.	.	.	.	.	.	.	.	.	.
GR28	.	.	.	.	.	.	.	.	.	G	.	.	T
GR29	.	.	.	.	.	.	.	.	.	G	.	Y	Y
GR34	.	.	.	.	K	M	.	.	Y	R	.	Y	.
GR35	.	.	.	.	.	.	.	.	.	G	.	.	T
GR38	.	.	.	.	.	.	.	.	.	G	.	.	Y
GR40	.	.	.	.	.	.	.	.	.	G	.	.	T
GR43	.	.	.	.	.	.	.	.	.	G	.	.	T
GRR01	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR02	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR03	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR04	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR05	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR06	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR07	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR08	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR09	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR11	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR12	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR13	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR14	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR15	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR16	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR17	C	G	C	.	.	.	T	.	.	.	K	.	.
GRR18	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR19	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR20	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR21	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR22	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR28	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR29	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR30	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR31	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR32	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR33	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR34	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR35	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR36	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR37	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR40	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR41	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR42	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR43	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR44	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR46	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR48	C	G	C	.	.	.	T	.	.	.	G	.	.
GRR49	C	G	C	.	.	.	T	.	.	.	G	.	.

Continuação...

	52	87	186	303	308	322	357	359	507	538	576	585	600
	T	A	T	A	G	C	C	C	C	A	C	C	C
UR01	.	.	.	.	.	.	.	.	Y	.	.	.	.
UR02	.	.	.	M	.	.	.	Y	.	G	.	Y	.
UR03	.	.	.	.	.	.	.	T	.	G	.	T	.
UR04	.	.	.	.	.	.	.	.	.	.	.	-	.
UR06	.	.	.	.	.	.	.	T	.	G	.	T	.
UR07	.	.	.	.	.	.	.	T	.	G	.	T	.
UR08	.	.	.	.	.	.	.	T	.	G	.	T	.
UR09	.	.	.	.	.	.	.	T	.	G	.	T	.
UR10	.	.	.	M	.	.	.	Y	.	G	.	Y	.
UR11	.	.	.	.	.	.	.	T	.	G	.	T	.
UR12	.	.	.	.	.	.	.	T	.	G	.	T	.
UR13	.	.	.	.	.	.	.	T	.	G	.	T	.
UR16	.	.	.	.	.	.	.	T	.	G	.	T	.
UR21	.	.	.	C	.	.	.	.	.	G	.	.	.
UR22	.	.	.	C	.	.	.	.	.	G	.	.	.
UR23	.	.	.	C	.	.	.	Y	.	G	.	Y	-
UR25	.	.	.	.	.	.	.	T	.	G	.	T	.
UR26	.	.	.	M	.	.	.	Y	.	G	.	Y	.
UR28	.	.	.	.	.	.	.	T	.	G	.	T	.
UR37	.	.	.	.	.	.	.	T	.	G	.	T	.
UR39	.	.	.	.	.	.	.	T	.	G	.	T	.
UR40	.	.	.	.	.	.	.	T	.	G	.	T	.
UR41	.	.	.	.	.	.	.	T	.	G	.	T	.
UR42	.	.	.	.	.	.	.	T	.	G	.	T	.
UR43	.	.	.	.	.	.	.	T	.	G	.	T	.
GL01	C	G	C	.	.	.	T	.	.	.	G	.	.
GL02	C	G	C	.	.	.	T	.	.	.	G	.	.
GL03	C	G	C	.	.	.	T	.	.	.	G	.	.
GL04	C	G	C	.	.	.	T	.	.	.	G	.	.
GL05	C	G	C	.	.	.	T	.	.	.	G	.	.
GL06	C	G	C	.	.	.	T	.	.	.	G	.	.
GL07	C	G	C	.	.	.	T	.	.	.	G	.	.
GL08	C	G	C	.	.	.	T	.	.	.	G	.	.
GL09	C	G	C	.	.	.	T	.	.	.	G	.	.
GL10	C	G	C	.	.	.	T	.	.	.	G	.	.
GL11	C	G	C	.	.	.	T	.	.	.	G	.	.
GL12	C	G	C	.	.	.	T	.	.	.	G	.	.
GL15	C	G	C	.	.	.	T	.	.	.	G	.	.
GL17	C	G	C	.	.	.	T	.	.	.	G	.	.
GL19	C	G	C	.	.	.	T	.	.	.	G	.	.
GL20	C	G	C	.	.	.	T	.	.	.	G	.	.
GL21	C	G	C	.	.	.	T	.	.	.	G	.	.
GL23	C	G	C	.	.	.	T	.	.	.	G	.	.
GL25	C	G	C	.	.	.	T	.	.	.	G	.	.
GL27	C	G	C	.	.	.	T	.	.	.	G	.	.
GL29	C	G	C	.	.	.	T	.	.	.	G	.	.
GL30	C	G	C	.	.	.	T	.	.	.	G	.	.
GL31	C	G	C	.	.	.	T	.	.	.	G	.	.
GL32	C	G	C	.	.	.	T	.	.	.	G	.	.
GL36	C	G	C	.	.	.	T	.	.	.	G	.	.
GL38	C	G	C	.	.	.	T	.	.	.	G	.	.
GL40	C	G	C	.	.	.	T	.	.	.	G	.	.

Figura 23: Polimorfismos detectados para *COMT2* com frequência igual ou superior a 5%.

4.5.1 Análise interespecífica do gene *COMT2*

Na análise conjunta de todos indivíduos para *COMT2*, encontramos 7 inserções/deleções. Três delas ocorrendo isoladamente (um único indivíduo), enquanto que as outras ocorrem em vários indivíduos das três espécies, sendo que para o grupo de *E. grandis* (GR) proveniente de Atherton, nenhum indivíduo apresentou tal polimorfismo. Esta particularidade só poderá ser afirmada a partir de um estudo com uma quantidade maior de indivíduos, de forma que sejam encontradas mais inserções para que se torne estatisticamente relevante. Para o total de 24 polimorfismos identificados em 623 nucleotídeos seqüenciados, obtemos uma média de um polimorfismo a cada 26 bases.

Além das indels relatadas, *COMT2* apresentou 24 substituições, sendo 10 exônicas e 14 intrônicas. Dentre estas substituições, 14 substituem uma base por outra de mesmas características químicas (transição) enquanto que 10 substituem por outra de características químicas distintas (transversão).

Analisando o total de 10 substituições em exon, observamos que 6 delas são silenciosas enquanto que 4 são não silenciosas. Dentre estas 10 substituições em exon, 7 são transições enquanto que 3 são transversões.

Observando somente as substituições não silenciosas, 2 são conservativas (fenilalanina para leucina e isoleucina para metionina). Adicionalmente, duas substituições acarretaram a mudança para aminoácidos com características distintas (não conservativa). Uma substitui uma treonina (polar sem carga) por uma alanina (apolar) enquanto que a outra substitui uma lisina (polar positivo) por uma treonina (polar sem carga). Diferentemente de *CAD2*, não foi observado nenhum código de terminação na análise de *COMT2*.

Assim como em *CAD2*, *COMT2* tem diversidade nucleotídica alta (comt_todos na Tabela 17) entre as espécies (Pi=0,629%), apesar de a similaridade também ser alta (95,02%). A maioria dos sítios polimórficos acabou por ser encontrada em *E. urophylla* e *E. grandis* (GR). Por outro lado, entre *E. grandis* (GRR) e *E. globulus* praticamente não houve polimorfismo. Desta forma, observa-se um predomínio de um grupo de alelos entre GR e *E. urophylla* e de outro para GRR e *E. globulus*.

Tabela 17: Dados resultantes da análise da diversidade nucleotídica de COMT2. Os grupos são identificados inicialmente pelo gene estudado (COMT2), e posteriormente pelo grupo (gr – GR; grr – GRR; gr_grr – análise conjunta destes dois grupos; gl – E. globulus; ur – E. urophylla; e comt todos – análise conjunta de todos os indivíduos analisados).

	Sítios polimórficos	Singletons	Sítios informativos	h	Pi
comt_gr	4	0	4	8	0,00306
comt_grr	15	8	8	15	0,00411
comt_gr_grr	11	2	9	13	0,00576
comt_gl	2	2	0	3	0,00056
comt_ur	2	0	2	4	0,00255
comt_todos	15	4	11	18	0,00629

Assim como pode ser avaliado na Tabela 17, uma análise detalhada foi realizada utilizando o programa DNAsp, onde foram detectados 15 sítios polimórficos sendo, dentre estes, 4 singletons e 11 informativos. Foram relacionados 12 haplótipos (Tabela 18) em uma análise utilizando o programa Phase.

Tabela 18: Estrutura dos haplótipos observados para *COMT2* em uma análise interespecífica.

haplótipo	E(freq)
TATAGCCCCGCCT	0.112000
TATAGCCCCGCCC	0.052404
TATAGCCCCGCTT	0.000069
TATAGCCCCGCTC	0.012828
TATAGCCCCACCC	0.021543
TATAGCCCTACCC	0.004310
TATAGCCTCGCTC	0.158907
TATATACCTACCC	0.030138
TATCGCCCCGCCC	0.033877
TATCGCCTCGCTC	0.004882
TACAGCCCCGCCC	0.008621
CGCAGCTCCAGCC	0.560345

Foram também realizadas análises de desequilíbrio de ligação utilizando o programa Haploview, demonstrando vários níveis do mesmo.

Tabela 19: Análise de desequilíbrio de ligação para *COMT2* utilizando todos os indivíduos.

Sítios		D'	r^2
sitio1	sitio2	1.0	1.0
sitio1	sitio3	1.0	0.455
sitio1	sitio4	1.0	0.018
sitio1	sitio5	1.0	0.0080
sitio1	sitio6	1.0	0.0080
sitio1	sitio7	1.0	1.0
sitio1	sitio8	1.0	0.018
sitio1	sitio9	1.0	0.018
sitio1	sitio10	1.0	0.182
sitio1	sitio11	1.0	1.0
sitio1	sitio12	1.0	0.045
sitio1	sitio13	1.0	0.018
sitio2	sitio3	1.0	0.455
sitio2	sitio4	1.0	0.018
sitio2	sitio5	1.0	0.0080
sitio2	sitio6	1.0	0.0080
sitio2	sitio7	1.0	1.0
sitio2	sitio8	1.0	0.018
sitio2	sitio9	1.0	0.018
sitio2	sitio10	1.0	0.182
sitio2	sitio11	1.0	1.0
sitio2	sitio12	1.0	0.045
sitio2	sitio13	1.0	0.018
sitio3	sitio4	1.0	0.04
sitio3	sitio5	1.0	0.018
sitio3	sitio6	1.0	0.018
sitio3	sitio7	1.0	0.455
sitio3	sitio8	1.0	0.04
sitio3	sitio9	1.0	0.04
sitio3	sitio10	0.25	0.025
sitio3	sitio11	1.0	0.455
sitio3	sitio12	1.0	0.1
sitio3	sitio13	1.0	0.04
sitio4	sitio5	1.0	0.018
sitio4	sitio6	1.0	0.018
sitio4	sitio7	1.0	0.018
sitio4	sitio8	0.4	0.16
sitio4	sitio9	1.0	0.04
sitio4	sitio10	1.0	0.1
sitio4	sitio11	1.0	0.018
sitio4	sitio12	0.25	0.025
sitio4	sitio13	1.0	0.04
sitio5	sitio6	1.0	1.0

sitio5	sitio7	1.0	0.0080
sitio5	sitio8	1.0	0.018
sitio5	sitio9	1.0	0.455
sitio5	sitio10	1.0	0.182
sitio5	sitio11	1.0	0.0080
sitio5	sitio12	1.0	0.045
sitio5	sitio13	1.0	0.018
sitio6	sitio7	1.0	0.0080
sitio6	sitio8	1.0	0.018
sitio6	sitio9	1.0	0.455
sitio6	sitio10	1.0	0.182
sitio6	sitio11	1.0	0.0080
sitio6	sitio12	1.0	0.045
sitio6	sitio13	1.0	0.018
sitio7	sitio8	1.0	0.018
sitio7	sitio9	1.0	0.018
sitio7	sitio10	1.0	0.182
sitio7	sitio11	1.0	1.0
sitio7	sitio12	1.0	0.045
sitio7	sitio13	1.0	0.018
sitio8	sitio9	1.0	0.04
sitio8	sitio10	1.0	0.1
sitio8	sitio11	1.0	0.018
sitio8	sitio12	1.0	0.4
sitio8	sitio13	1.0	0.04
sitio9	sitio10	1.0	0.4
sitio9	sitio11	1.0	0.018
sitio9	sitio12	1.0	0.1
sitio9	sitio13	1.0	0.04
sitio10	sitio11	1.0	0.182
sitio10	sitio12	1.0	0.25
sitio10	sitio13	1.0	0.1
sitio11	sitio12	1.0	0.045
sitio11	sitio13	1.0	0.018
sitio12	sitio13	0.25	0.025

4.5.2 Análise intraespecífica do gene *COMT2*

4.5.2.1 Análise de *E. urophylla*

O gene *COMT2* em *E. urophylla* mostrou pouca diversidade intraespecífica (Pi=0,255%), além de ter demonstrado um polimorfismo exclusivo para a espécie.

Possui dois sítios polimórficos (303 e 359) que, para esta análise, estão presentes somente nesta espécie, assim como pode ser analisado na Figura 23.

Os dois sítios polimórficos encontrados são informativos. Foram encontrados 7 haplótipos (Tabela 20). A diversidade nucleotídica (Pi) calculada para *E. urophylla* foi de 0,00255, confirmando a baixa variabilidade intraespecífica.

Tabela 20: Estrutura dos haplótipos observados em *E. urophylla* para *COMT2* em uma análise intraespecífica.

haplótipo	E(freq)
TATAGCCCCACCC	0.059546
TATAGCCCCACTC	0.000454
TATAGCCCTACCC	0.020000
TATAGCCTCGCTC	0.739999
TATCGCCCCGCCC	0.159899
TATCGCCTCGCTC	0.019743
TATCGCCTCGCTD	0.000257

Como mostra a Tabela 21, foi analisada a presença de desequilíbrio de ligação para *E. urophylla*.

Tabela 21: Análise de desequilíbrio de ligação para *COMT2* em indivíduos de *E. urophylla*.

Sítios		D'	r^2
sitio4	sitio8	0.25	0.062
sitio4	sitio9	1.0	0.1
sitio4	sitio10	1.0	0.5
sitio4	sitio12	0.0	0.0
sitio8	sitio9	1.0	0.1
sitio8	sitio10	1.0	0.5
sitio8	sitio12	1.0	0.5
sitio9	sitio10	1.0	0.2
sitio9	sitio12	1.0	0.2
sitio10	sitio12	0.333	0.111

4.5.2.2 Análise de *E. globulus*

Dentre todas as análises em *COMT2*, *E. globulus* demonstrou a maior conservação intraespecífica, com apenas duas substituições encontradas. Estas substituições são 2

singletons. Foi encontrado apenas 1 haplótipo (Tabela 22), demonstrando a baixa diversidade intraespecífica de *COMT2* em *E. globullus*. A diversidade nucleotídica também foi bastante baixa, e a menor observada dentre todas neste trabalho (Pi=0,056%).

Tabela 22: Estrutura do haplótipo observado em *E. globulus* para *COMT2* em uma análise intraespecífica.

haplótipo	E(freq)
CGCAGCTCCAGCC	1.000.000

4.5.2.3 Análise de *E. grandis*

Quanto a *E. grandis*, intraespecificamente foi possível notar que o nível médio de polimorfismos é pequeno, mas existe uma distinção clara entre os dois grupos. Enquanto que GR (Atherton) possui vários indivíduos por sítio polimórfico, GRR (Pine Creek) possui muito poucos indivíduos compartilhando os polimorfismos (uma média de dois indivíduos por polimorfismo). Além disso, poucos são os polimorfismos coincidentes, assim como visto na Figura 23.

Na análise entre os dois grupos (GR e GRR), foram encontrados 2 “singletons” e 9 sítios informativos, totalizando 11 sítios polimórficos, Além de 13 haplótipos. Já a diversidade nucleotídica (Pi), só não é maior que a obtida para o conjunto de todos indivíduos (Pi=0,576%). Analisando isoladamente, avaliamos que GRR apresentou uma pequena quantidade de haplótipos (h=1) (Tabela 23).

Tabela 23: Estrutura dos haplótipos observados em E. grandis (GRR) para COMT2 em uma análise intraespecífica.

haplótipo	E(freq)
CGCAGCTCCAGCC	1.000.000

GR, por sua vez, apresentou uma quantidade maior de haplótipos (h=8) (Tabela 24).

Tabela 24: Estrutura dos haplótipos observados em *E. grandis* (GR) para *COMT2* em uma análise intraespecífica.

haplótipo	E(freq)
TATAGCCCCGCCC	0.234847
TATAGCCCCGCTT	0.003010
TATAGCCCCGCTC	0.053615
TATAGCCCCACCC	0.038462
TATATACCTACCC	0.133548
TATATACCTACTC	0.001067
TACAGCCCCGCCC	0.038462

Tabela 25: Análise de desequilíbrio de ligação para *COMT2* em *E. grandis* (GR).

Sítios		D'	r^2
sitio3	sitio5	1.0	0.048
sitio3	sitio6	1.0	0.048
sitio3	sitio9	1.0	0.048
sitio3	sitio10	1.0	0.086
sitio3	sitio12	1.0	0.086
sitio3	sitio13	1.0	0.048
sitio5	sitio6	1.0	1.0
sitio5	sitio9	1.0	1.0
sitio5	sitio10	1.0	0.556
sitio5	sitio12	0.2	0.022
sitio5	sitio13	1.0	0.111
sitio6	sitio9	1.0	1.0
sitio6	sitio10	1.0	0.556
sitio6	sitio12	0.2	0.022
sitio6	sitio13	1.0	0.111
sitio9	sitio10	1.0	0.556
sitio9	sitio12	0.2	0.022
sitio9	sitio13	1.0	0.111
sitio10	sitio12	0.111	0.0040
sitio10	sitio13	1.0	0.2
sitio12	sitio13	0.2	0.022

5 DISCUSSÃO

5.1 Triagem dos clones de BAC

No contexto deste trabalho os genes da via de lignificação são os alvos preferenciais devido estarmos focando nas bases moleculares da qualidade da madeira. Para desenvolver tais estudos, uma ferramenta bastante útil é a biblioteca de BAC desenvolvida no âmbito do projeto Genolyptus, sendo uma fonte confiável para se obter genes completos, já que o BAC é muito menos complexo que o DNA genômico, e o mesmo é suficientemente grande para conter toda a seqüência do gene.

Para seqüenciar o clone que contem um gene em particular, foi preciso identificar inicialmente qual dos 20.160 clones continha o gene de interesse. Para este fim foi realizada a triagem dos mesmos através da estratégia dos grupos e supergrupos. Para todos os clones confirmados por pelo menos três PCRs distintas, foi feita uma nova minipreparação de forma a anular o risco de contaminação na minipreparação de DNA. Apesar deste cuidado, percebemos a presença de dois clones positivos para *CAD2* em uma mesma placa, algo improvável, sugerindo contaminação, principalmente por serem poços não muito distantes na placa.

Para o seqüenciamento dos genes após a seleção de seus respectivos clones de BAC, foram inicialmente traçadas duas estratégias. A primeira consistia em realizar o "primer walking", onde desenhou-se iniciadores baseados em seqüências de ESTs do banco do Genolyptus tentando utilizá-los para o seqüenciamento direto do BAC. Entretanto, tal estratégia acabou por se tornar inviável visto que o rendimento do seqüenciamento dos clones mostrou-se reduzido.

Após verificarmos que a estratégia de "primer walking" não foi bem sucedida, partimos então para a utilização da técnica de "shotgun" para gerar seqüências aleatórias do clone de BAC e utilizar técnicas de bioinformática para a montagem destas.

5.2 Seqüência completa de *CAD2*

A seqüência completa para o gene *CAD2* foi obtida a partir do seqüenciamento de uma biblioteca "shotgun" (GREEN, 2001) alcançando-se uma cobertura de dez vezes em relação ao tamanho do inserto do clone de BAC que, neste caso, era de aproximadamente 30 Kb. A montagem gerou uma seqüência contígua de 9.785 bases, a qual continha o gene completo de *CAD2*. Apesar do sucesso no seqüenciamento do gene de interesse, vale a pena ressaltar que era esperado que todo o inserto de 30 Kb pudesse ser montado, em função do número obtido de leituras de seqüência. A razão para isto pode ser a alta incidência de seqüências repetitivas em regiões extra-gênicas que podem afetar a eficiência da montagem.

De qualquer maneira, foi obtida a primeira seqüência completa de *CAD2* em *Eucalyptus grandis*. A confirmação em relação a qual cópia obtivemos através do alinhamento contra as seqüências de *CAD1* e *CAD2* de *E. botryoides*, que mostrou maior identidade com a última.

A obtenção desta seqüência completa proporcionará estudos mais completos deste gene, assim como uma análise total de sua diversidade nucleotídica através da utilização de iniciadores que possam cobrir tanto regiões codificadoras quanto de controle (promotor). Com isso, será possível uma melhor caracterização da variação nucleotídica entre espécies, permitindo melhores estimativas da extensão do DL e estrutura haplotípica para este gene.

A análise de seqüência demonstrou que *CAD* é significativamente conservada em várias espécies de eucalipto, tanto em termos de nucleotídeos quanto em termos de aminoácidos, resultados coincidentes com os de ENDT *et al*. (2000). Contrastando com estes dados, foi encontrada uma distinção entre as cópias de *CAD*, indels nos últimos nucleotídeos do íntron 4, fator que pode estar gerando alteração no padrão de splicing na cópia *CAD1* do gene.

Adicionalmente à comparação de nucleotídeos e de aminoácidos, os resultados obtidos a partir de uma análise filogenética mais uma vez demonstram uma grande pressão de seleção ocorrendo em *CAD* de *Eucalyptus*, visto que todas as cópias de *CAD* de diferentes espécies de *Eucalyptus* formaram um grupo monofilético bastante distinto.

Esta possível pressão de seleção também pode ser observada pela conservação estrita da região regulatória que é alvo de ligação do fator de transcrição EgMYB2 (GOICOECHEA *et al*., 2005) entre *E. gunnii* e *E. grandis*. Tal região regulatória, além e ser bastante conservada entre espécies e ter expressão preferencial em xilema, revelou ter bastante influência na atividade de enzimas da via de lignificação, aumentando em cinco vezes a quantidade de transcritos de *CAD* e quarenta vezes a de *COMT* quando expressa de maneira heteróloga em tabaco.

5.3 Análise de diversidade nucleotídica dos genes *CAD2* e *COMT2*

Muitas foram as dificuldades de se amplificar com um mesmo par de iniciadores os fragmentos para as três espécies analisadas. Os melhores foram aqueles desenvolvidos por GION *et al*. (2000) enquanto que outros se limitavam a amplificar apenas uma espécie, poucos indivíduos, ou na maioria dos casos, não amplificavam na reação de seqüenciamento. Este impedimento foi observado para os dois genes estudados, principalmente para as espécies *E. urophylla* e *E. globulus*. Estes indivíduos não

analisados podem estar apresentando, na região de anelamento dos iniciadores, algum polimorfismo que dificulte o anelamento dos iniciadores.

A partir de tais fatos, deve-se avaliar com precaução a análise de diversidade nucleotídica realizada, visto que uma parcela de amostras que possivelmente possui alguns polimorfismos distintos deixou de ser analisada. Tais indivíduos podem apresentar uma estrutura de haplótipos própria, o que não pôde ser detectado, já que os mesmos não foram amplificados.

Diferentemente deste trabalho, POKE *et al*. (2003), autor do trabalho de onde retiramos os iniciadores para o seqüenciamento de *CAD2*, relatou um certo sucesso na amplificação e seqüenciamento de *CAD2* em *E. globulus*, atingindo 45% do seqüenciamento completo na região analisada. No presente trabalho, os iniciadores tinham dificuldade de amplificação em *E. globulus* e *E. urophylla* tanto na PCR quanto no sequenciamento.

Para o processo de desenho dos iniciadores desenvolvidos por estudos próprios, vários foram desenhados procurando-se regiões conservadas entre as três espécies estudadas, de forma que um mesmo iniciador tivesse sucesso para todas amostras. Devido ao baixo número de seqüências depositadas no banco público (GenBank), para alguns iniciadores foram utilizadas também seqüências de *Arabidopsis thaliana* e *Pinus spp*, o que pode ser um dos motivos para a ineficiência. Alguns iniciadores foram também desenhados a partir de alinhamentos das seqüências de ESTs do Genolyptus, mas seu resultado deixou a desejar visto que, em alguns casos, detectávamos posteriormente uma indel na posição de anelamento dos iniciadores.

Portanto, novos estudos devem ser realizados com a finalidade de gerar iniciadores que sejam capazes de amplificar estes genes a partir do DNA de várias espécies diferentes. Tais estudos serão facilitados quando novas seqüências sejam depositadas no

banco de dados público "GenBank", de forma a obtermos mais informações de polimorfismos espécie-específicos.

5.4 Análise interespecífica do gene *CAD2*

Tanto para a análise intraespecífica quanto para a interespecífica, a região genômica seqüenciada envolvia 358 nucleotídeos: 55 pertencentes ao último intron, 194 ao último exon, além de 109 nucleotídeos da região 3' UTR, que geralmente tem um papel importante na estabilidade do mRNA, influenciando fortemente na expressão de alguns genes (ALBERTS *et al*., 2002).

Ao contrário do esperado, o valor bruto das substituições exônicas foi maior que o das observadas no 3' UTR, mas isto se deve primariamente ao tamanho das regiões estudadas para ambos. Foram analisados praticamente o dobro de nucleotídeos em exon em relação a 3' UTR, sendo a quantidade relativa de substituições quase igual. Mesmo tendo metade do tamanho do exon (no fragmento analisado), a região 3' UTR demonstrou ser bastante polimórfica, o que era previsível.

Nesta análise de *CAD2*, foi observado um padrão de polimorfismos mais freqüente na região 3' UTR que em exon ou no intron. Nenhuma substituição foi observada na região intrônica analisada, 7 na região 3' UTR para os 109 pb analisados (caracterizando uma substituição a cada 15 pb), e 8 substituições encontradas nos 194 pb exônicos, caracterizando uma a cada 24 pb.

O número de transições no gene *CAD2* entre as três espécies foi maior que o de transversões. As transições se concentraram mais no exon enquanto que as transversões no 3' UTR. Das substituições encontradas em exon, 5 são não conservativas. Este dado é bastante significativo já que destas 5, duas são "singletons" que acabam por codificar um códon de terminação. Um desses códons de terminação é o do próprio exon 5,

enquanto que o outro foi originado de uma transição no indivíduo 3 de *E. urophylla*. A partir destes dados, mais indivíduos devem ser avaliados para se confirmar tais dados já que a presença de um códon de terminação em uma região exônica, pode vir a ser determinante para a funcionalidade da proteína e acarretar em fenótipos característicos.

Mesmo sendo apresentada uma alta similaridade interespecífica, distinções bastante claras foram observadas, sendo que algumas delas até mesmo entre os dois grupos de E. grandis.

Interespecificamente, *CAD2* demonstrou uma diversidade nucleotídica alta em relação às análises intraespecíficas, fato decorrente da utilização dos dados dos indivíduos de todas espécies conjuntamente. Observou-se uma diferenciação entre um grupo formado por GR (Atherton) e *E. urophylla* (Timor) em relação a GRR (Pine Creek) e *E. globulus* (Victoria). Esta diferenciação não era esperada, visto que os dois grupos de *E. grandis* acabaram por tomar direções opostas, podendo, em algum momento, ter sofrido algum tipo de seleção por algum agente externo, como o microclima da região. Outro fator que pode ter influenciado tais padrões á a latitude de cada grupo, já que os mais semelhantes geneticamente são os mais próximos quanto à latitude (10º N para GR e *E. urophylla* e 10º S para GRR e *E. globulus*). De qualquer forma, uma análise mais ampla dos genes deve ser realizada, além de uma análise do histórico evolutivo dos grupos analisados de forma a embasar futuras discussões.

Foram encontrados 11 haplótipos interespecíficos em *CAD2*, demonstrando o grau de diversidade que ocorre entre os quatro grupos analisados. Apenas 4 dos 11 haplótipos correspondem a cerca de 83% dos casos, indicando que apesar dos polimorfismos, não são muitos aqueles que destoam do padrão haplotípico principal. Todos os haplótipos interespecíficos foram encontrados nas análises intraespecíficas, fortalecendo o montante de dados analisados.

Foi detectada a presença de desequilíbrio de ligação entre tais sítios polimórficos, algo esperado devido à pequena distância entre os mesmos.

5.5 Análise intraespecífica do gene *CAD2*

Foi observada em *E. urophylla* a maior diversidade nucleotídica dentre os quatro grupos analisados para *CAD2*. Tais dados devem ser avaliados com ressalvas, já que apenas 10 indivíduos foram analisados devido a problemas de amplificação.

Apesar da alta diversidade nucleotídica, o número de haplótipos encontrados em *E. urophylla* foi de apenas quatro, número que pode refletir o baixo número de indivíduos analisados. Tal número pode demonstrar um alto grau de similaridade intraespecífica, assim como pôde ser observado em GR (Atherton) para *CAD2*, que apresentou apenas cinco haplótipos, apesar do número de indivíduos ser quase três vezes maior. Dois dos haplótipos observados em *E. urophylla* também puderam ser observados no grupo GR (Atherton) de *E. grandis*, ou seja, metade dos haplótipos observados em *E. urophylla* foram também observados neste grupo (GR), indicando que ambos devem ser bastante próximos geneticamente. Para indivíduos de *E. urophylla*, a freqüência de sítios polimórficos foi de 51 pb, em média. Na análise de desequilíbrio de ligação, nada foi detectado para *E. urophylla* em *CAD2*.

O código de terminação observado no indivíduo 3 de *E. urophylla* pode-se tornar bastante informativo, visto que podem ter sido perdidos 31 aminoácidos no gene *CAD2* deste indivíduo, sendo um mutante natural para este gene, e interessante para uma análise fenotípica.

Os estudos de *CAD2* em *E. globulus*, assim como para *E. urophylla* não puderam ser conclusivos visto a quantidade de indivíduos analisados, mas de qualquer maneira, estes dados servem como base para estudos posteriores. De acordo com a atual análise,

CAD2 em *E. globulus* foi onde observamos o menor valor de diversidade nucleotídica (P_i=0,164%), o que acabou por ser refletido na ausência de sítios informativos e presença de apenas 2 "singletons" (1 polimorfismo a cada 119 pb intraespecificamente, assim como em *E. urophylla*). Com tal conservação, como conseqüência, o número de haplótipos observados foi baixo (h=3). Os haplótipos observados em *E. globulus* foram totalmente distintos dos de GR e *E. urophylla*, mas todos os mesmos foram observados também na análise do grupo GRR de *E. grandis*. Desta forma, podemos indicar que estes dois grupos são bastante semelhantes geneticamente e que uma eventual dispersão que os separou deve ter ocorrido recentemente.

Esta diferenciação quanto às outras espécies também foi confirmada na detecção dos haplótipos específicos para *E. globulus*, detectados nas ESTs. Tais haplótipos podem ser bastante úteis já que a seqüência está disponível para o desenho de iniciadores que poderão ser desenhados justamente nas posições dos haplótipos, tornando-se iniciadores espécie específicos.

Em *E. grandis*, a diversidade intraespecífica foi mais alta que o esperado (P_i=0,672%), refletindo em um maior número total de haplótipos (h=10). Resultados semelhantes puderam ser observados também individualmente para cada um dos dois grupos de *E. grandis* analisados (Tabela 12 e Tabela 14). A presença desta diferenciação entre os dois grupos de *E. grandis*, mesmo que pequena, acaba por indicar que tais grupos acabaram por sofrer pressões seletivas distintas, gerando esta pequena, mas consistente diferença entre estes. Assim como demonstrado nas tabelas 13 e 15, foi detectado desequilíbrio de ligação entre os sítios polimórficos analisados. Tal desequilíbrio não é muito forte provavelmente pelas limitações causadas pelo baixo número de indivíduos amostrados, já que, pela pequena distância física, os mesmos deveriam estar em total desequilíbrio de ligação.

Nas análises de desequilíbrio de ligação de todos os grupos, observa-se uma grande discrepância entre os valores de D' e r^2 mas sabe-se que o valor de r^2 passa a não ser muito confiável em uma análise de um pequeno número de amostras (GARCIA, 2003; DEVLIN e RISCH, 1995). Além disso, DEVLIN e RISCH (1995) e LIN *et al*. (2002) afirmam que freqüências alélicas altas em sítios próximos geram valores de r^2 baixos (o que pode ter ocorrido na análise). Outro ponto a ser avaliado é outra afirmação de DEVLIN e RISCH (1995), afirmando que o coeficiente de correlação pode ser de difícil interpretação quando as freqüências alélicas variam entre os loci. Já CRAWFORD *et al*. (2005) afirma que r^2 possui relação inversa com a força para detectar a associação, dados que indicam que os valores de r^2 baixos observados acompanhados de um alto valor de *D'* provavelmente indicam um forte desequilíbrio de ligação.

Avaliando a estrutura dos haplótipos gerada para cada grupo de indivíduos de *E. grandis*, observa-se que nenhum deles é coincidente em relação aos do outro grupo, diferindo totalmente do esperado que seria uma alta similaridade em todos os aspectos entre estes dois grupos de *E. grandis*. Em cada grupo foram observados 5 haplótipos, sendo todos os 5 de cada grupo distintos dos haplótipos do outro grupo. Tal diferenciação pode estar relacionada à posição geográfica das mesmas, já que os grupos com latitudes semelhantes se apresentaram muito mais semelhantes entre si que em relação aos outros. Quanto à freqüência de sítios polimórficos, em GR foi observado 1 a cada 60 pb enquanto que em GRR, 1 a cada 36 pb, em média, uma diferença bastante significativa, visto serem indivíduos da mesma espécie. Esta diferenciação pode ter se originado da forma de dispersão ocorrida para tais grupos, de forma que a cada geração separados, cada grupo flutuaria suas freqüências alélicas irregularmente, se dispersando e tornando-se progressivamente diferenciados (FALCONER, 1987).

5.6 Análise interespecífica do gene *COMT2*

Um ponto positivo obtido na análise de *COMT2* diz respeito ao número de indivíduos analisados. Diferentemente de *CAD2*, obteve-se para *COMT2* um número homogêneo de indivíduos para cada grupo analisado. Isto reforça a hipótese de problemas de anelamento no caso de *CAD2*, dada que a qualidade do DNA utilizado ter sido satisfatória para *COMT2*.

Assim como ilustrado na Figura 22, conseguimos amplificar dois introns, sendo um com a seqüência completa e outro parcial, além de dois exons, um com seqüência completa e outro com seqüência parcial. Apesar de o esperado ser a presença muito maior de substituições intrônicas, esta proporção não ocorreu para a presente análise de *COMT2*. Por outro lado, a maioria das substituições em exon são silenciosas (substituição em que não há modificação do aminoácido).

Foram encontradas inserções (5 inserções para um total de 126 indivíduos analisados) na posição 581 para indivíduos de GRR, *E. urophylla* e *E. globulus*, mas não de GR. Por outro lado, o número de indivíduos amostrados ainda é pequeno para se fazer algum tipo de afirmação neste aspecto, sendo que uma caracterização genética de mais indivíduos faz-se necessária. De qualquer forma, futuramente poderá se identificar se tal polimorfismo é ou não candidato para uso como marcador molecular, ou como sítio de anelamento para iniciadores voltados para os indivíduos possuidores ou não, de tal inserção (espécie específico).

Em termos de polimorfismos no gene *COMT2*, foram descobertas 24 substituições nestes 623 nucleotídeos (1 polimorfismo a cada 26 pb interespecificamente), sendo que 10 aconteceram em exon e 14 em intron. Foi observado um padrão de substituições distinto do esperado, onde a frequência de substituições exônicas se apresentou maior que as intrônicas. Foi observada uma substituição a cada 28 pb em intron (um total de

14 substituições) e uma a cada 23 pb em exon (um total de 10 substituições), demonstrando um padrão diferente do esperado, que seria uma frequência menor de polimorfismos na região exônica em relação à intrônica.

As transições se dividiram entre o exon e o intron enquanto que as transversões (substituição de uma base por outra de características químicas diferentes) foram mais freqüentes em introns. Como esperado, a maioria das substituições foram silenciosas nos exons, visto que uma substituição não conservativa poderia inviabilizar a atividade da proteína.

Assim como foi observado em *CAD2*, a diversidade nucleotídica interespecífica é grande, sendo mais claramente percebida quando se compara GR (Atherton) e *E. urophylla* (Timor) com GRR (Pine Creek) e *E. globulus* (Victoria). Enquanto o primeiro grupo tem poucos polimorfismos, o segundo acaba por demonstrar uma grande similaridade já que a maioria dos polimorfismos detectados neste segundo grupo são conservados para a maioria dos indivíduos.

Uma grande quantidade de haplótipos foi observada (12 haplótipos), indicando uma falta de correlação genética entre os mesmos, além de indicar que a população, em sua história genética, não sofreu recentemente um efeito gargalo ou efeito fundador. Assim como foi observado para *CAD2*, os haplótipos interespecíficos foram os mesmos que foram observados em todos os grupos e, assim como em *CAD2*, pôde-se observar que pelo número de haplótipos, os indivíduos não têm uma similaridade muito alta.

Diversos níveis de desequilíbrio de ligação foram detectados entre tais polimorfismos, reforçando a idéia de que estes grupos têm histórias genéticas distintas.

5.7 Análise intraespecífica do gene *COMT2*

Foi observado um baixo valor de diversidade nucleotídica para *E. urophylla* (Timor) em relação a *E. grandis* (GR). Este baixo valor deve-se ao fato de seus polimorfismos serem muito frequentes na grande maioria dos indivíduos analisados, diferentemente de GR em que os mesmos são pouco freqüentes e dispersos. *E. urophylla* apresentou um sítio polimórfico exclusivo (posição 303), dado que pode ser reconfirmado em estudos utilizando mais indivíduos. De qualquer forma, a partir destes dados, este polimorfismo se caracteriza como espécie-específico.

Apesar dos 7 haplótipos detectados para *E. urophylla*, a maioria apresenta freqüências muito baixas e desequilíbrio de ligação de variadas intensidades entre os polimorfismos. Provavelmente essas diferenças de intensidade são derivadas do baixo número de indivíduos analisados e da imprecisão dos resultados nestes casos. Em uma análise comparativa, os haplótipos de *E. urophylla* se mostraram totalmente distintos do grupo de *E. globulus* e do grupo GRR de *E. grandis*. Por outro lado, mostraram alguma semelhança em relação ao grupo GR de *E. grandis* ao compartilhar com o mesmo um haplótipo em comum.

Foi também observada uma freqüência de sítios polimórficos, em média a cada 44 pb, diferentemente de genes de milho, onde esta média sobe para cerca de 60,8 pb em linhagens elite (CHING, 2002), e 130 pb na região codificante de linhagens cultivadas (RAFALSKI, 2002).

No caso de *E. globulus*, a freqüência de sítios polimórficos foi de 1 a cada 156 pb, uma freqüência bastante baixa. Valor confirmado pela baixa diversidade intraespecífica encontrada para seus indivíduos (P_i=0,056%). Esta baixa diversidade foi também apresentada no presente estudo para o gene *CAD2* em *E. globulus* (com um número de indivíduos analisados 3 vezes menor), demonstrando que talvez *E. globulus* se destaque das outras espécies pela baixa taxa de polimorfismos nestes genes.

Refletindo a sua baixa diversidade nucleotídica, foi observado também apenas um haplótipo em todos os indivíduos em *COMT2*, caracterizando total desequilíbrio de ligação na região analisada. Tais valores só vêm a comprovar os dados obtidos anteriormente, indicando que *E. globulus* provavelmente sofreu pressões seletivas um pouco distintas ou sua base genética é estreita, principalmente comparado a GR (Atherton) e *E. urophylla*. Comparando os haplótipos do grupo de *E. globulus* e do grupo GRR de *E. grandis*, observamos que ambos apresentaram apenas um haplótipo, sendo este exatamente igual. Desta forma, podemos inferir que tais grupos são extremamente semelhantes geneticamente no que se refere a esta região gênica.

Quanto aos estudos realizados para os dois grupos de *E. grandis*, assim como já descrito, o número médio de polimorfismos não é muito grande, mas aqueles polimorfismos que foram observados acabaram por demonstrar que estes dois grupos de *E. grandis* devem ter tomado a algum tempo rotas evolutivas distintas. Enquanto que polimorfismos em GRR (Pine Creek) são escassos e pouco freqüentes, em GR (Atherton) tal padrão é quebrado com vários indivíduos polimórficos para cada sítio.

Isoladamente, GRR (Pine Creek) apresentou um pequeno número de haplótipos (h=1), assim como uma diversidade nucleotídica relativamente baixa (P_i=0,411%) comparada a encontrada em Arabidopsis (P_i=1,63% generalizado, P_i=0,42% em exon, e P_i=2,93% em intron ;YOSHIDA *et al*., 2003). Além disso, seu haplótipo, asssim como descrito acima, é igual ao único haplótipo observado no grupo de *E. globulus,* demonstrando uma extrema afinidade genética nesta região gênica entre tais grupos. Já GR (Atherton) apresentou 8 haplótipos de freqüências variadas, indicando bastante diversidade entre os indivíduos, contrastando bastante com o grupo GRR (Pine Creek) e se aproximando do grupo de *E. urophylla*, compartilhando um haplótipo, dentre os 8 observados no grupo GR. Isto indica que existe semelhança entre estes grupos, mas

ainda são bastante distintos, sugerindo uma dispersão longíngua, dando tempo para os mesmos se distanciarem geneticamente.

Na análise do desequilíbrio de ligação, foram observados variados níveis de desequilíbrio em GR (Atherton), demonstrando a diversidade dos mesmos, por outro lado, todos os indivíduos de GRR (Pine Creek) se apresentaram idênticos, em total desequilíbrio de ligação. Tal diferenciação pode estar relacionada com gargalos genéticos ou ações externas.

6 CONCLUSÕES

Este trabalho visou contribuir para a busca da diversidade nucleotídica em genes da via de lignificação, como uma etapa preliminar para se entender os fatores genéticos ligados à deposição de lignina, e para a realização de estudos de associação para esta característica complexa.

Primeiramente otimizou-se a tecnologia de seleção de genes específicos em uma biblioteca de BACs obtendo-se dois genes para a via de lignificação. A partir desta triagem selecionou-se o BAC contendo o gene *CAD2* e obteve-se a sua seqüência completa, até então inédita. A análise de seqüência permitiu vislumbrar o grau de conservação deste gene em diversas espécies de eucaliptos, identificar regiões de controle de expressão na sua região promotora, identificar um haplótipo específico para *E. globulus* neste gene, e até o tecido preferencialmente expresso.

Passando então ao estudo da variabilidade nucleotídica em regiões dos genes *CAD2* e *COMT2*, de uma forma geral, era esperada uma diversidade nucleotídica menor dentro das espécies que entre as mesmas, justamente o que ocorreu. Observou-se também que as seqüências de *E. globulus* para ambos os genes apresentaram uma alta taxa de conservação.

Outro ponto que vale ser ressaltado, diz respeito à similaridade observada entre os grupos mais próximos geograficamente. Os indivíduos de GR (Atherton, estado de Queensland) e *E. urophylla* (Timor) se mostraram muito semelhantes geneticamente nas regiões amostradas, assim como os de GRR (Pine Creek, estado de "New South Wales") e *E. globulus* (Victoria). Tais semelhanças podem ter alguma relação com a latitude em que tais grupos encontram-se, já que grupos de latitudes semelhantes

apresentaram padrões genéticos semelhantes. Além desta hipótese, o que pode ser avaliado é de que região cada grupo migrou para as regiões atuais, ou se os mesmos foram introduzidos. De qualquer forma, novos estudos devem ser realizados de forma a caracterizar mais áreas destes genes para se confirmar tais padrões genéticos.

Para o gene *CAD2*, foi observado um baixo número de polimorfismos intraespecificamente, acompanhado de uma pequena diversidade nucleotídica para os grupos analisados enquanto que o gene *COMT2*, devido ao maior tamanho da região analisada, apresentou uma quantidade ligeiramente maior de polimorfismos. O que deve ser lembrado é que esta análise foi realizada para apenas 358 pb para *CAD2* e 623 pb para *COMT2*. Desta forma, temos apenas uma indicação de que estes genes são bastante conservados, nada podendo garantir quanto a diversidade nucleotídica das regiões adjacentes.

No caso da análise de desequilíbrio de ligação, apesar de terem sido detectados variados desequilíbrios entre sítios polimórficos para ambos os genes, estudos com mais indivíduos devem ser realizados de forma a melhorar a confiabilidade estatística dos dados.

7 PERSPECTIVAS FUTURAS

Este trabalho permitiu a identificação de clones da biblioteca de BAC para diversos genes da via de lignificação através de uma estratégia relativamente rápida e custo-efetiva. Estes são alvos imediatos para um seqüenciamento completo via a estratégia de "shotgun" ou por sub-clonagem após Southern blots. Seqüências completas destes genes oferecem uma perspectiva mais ampla, para que estudos futuros, busquem variações nucleotídicas entre as espécies, que potencialmente expliquem variações fenotípicas na quantidade e qualidade da lignina. É importante ressaltar que estes estudos devem ser realizados com um número maior de indivíduos e estarem associados com medidas fenotípicas confiáveis para as características de interesse. Adicionalmente, foi demonstrado neste estudo que o gene *COMT2* possui uma quantidade de haplótipos significativamente menor que a encontrada para *CAD2* em relação ao tamanho da região analisada, portanto, seria fundamental continuar a mesma linha de estudos com mais regiões deste gene.

Outro estudo que pode ser realizado é uma análise mais aprofundada do indivíduo 3 de *E. urophylla* que apresentou um código de terminação no exon 5, à 31 aminoácidos do código de terminação do exon 5, podendo estar alterando a atividade da proteína.

Seria também importante estender a gama de genes utilizados além da via de lignificação propriamente dita. Por exemplo, a triagem e reseqüenciamento do fator de transcrição EgMYB2 seria bastante útil em função deste ser um fator trans, que além de controlar a expressão de diversos genes da via de lignificação, se colocaliza com um QTL explicando 4,5% da variação fenotípica no conteúdo de lignina (GOICOECHEA *et al*., 2005).

8 BIBLIOGRAFIA

ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K., e WALTER, P. (2002) Molecular Biology of The Cell. 4th ed. Garland Publishing, New York. p.1661.

ALLONA, I., QUINN, M., SHOOP, E., SWOPE, K., ST CYR, S., CARLIS, J., RIEDL, J., RETZEL, E., CAMPBELL, M. M., SEDEROFF, R. e WHETTEN, R. W. (1998) Analysis of xylem formation in pine by cDNA sequencing. Proc Natl Acad Sci U S A. 95 (16): 9693-9698.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. e LIPMAN, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17): 3389-3402.

ANTEROLA, A. M. e LEWIS, N. G. (2002) Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochemistry. 61 (3): 221-294.

BANSAL, A., BOYD, P.R., e MCGINNIS, R. Tools for statistical analysis of genetic data. In: BARNES, M.R. e GRAY, I.C. Bioinformatics for Geneticists. John Wiley & Sons, Ltd, 2003. p. 217-245.

BATLEY, J., BARKER, G., O'SULLIVAN, H., EDWARDS, K. J. e EDWARDS, D. (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol. 132 (1): 84-91.

BELL, J. I. (2003) New hope for haplotype mapping. Arthritis Res Ther. 5 (2): 51-53.

BOREVITZ, J. O. e NORDBORG, M. (2003) The impact of genomics on the study of natural variation in Arabidopsis. Plant Physiol. 132 (2): 718-725.

BOREVITZ, J. O. e CHORY, J. (2004) Genomics tools for QTL analysis and gene discovery. Curr Opin Plant Biol. 7 (2): 132-136.

BOUDET, A.M. (1998) A new view of lignification. Trends in plant science 3(2): 67-71

BOUDET,A.M. (2000) Lignins and lignification: Selected issues. Plant Physiol. Biochem. 38 (1/2): 81-96.

BROOKES, A. J. (1999) The essence of SNPs. Gene. 234 (2): 177-186.

BROUILLETTE, J. A., ANDREW, J. R. e VENTA, P. J. (2000) Estimate of nucleotide diversity in dogs with a pool-and-sequence method. Mamm Genome. 11 (12): 1079-1086.

BROWN, G. R., GILL, G. P., KUNTZ, R. J., LANGLEY, C. H. e NEALE, D. B. (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U S A. 101 (42): 15255-15260.

BUCKLER, E. S. T. e THORNSBERRY, J. M. (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol. 5 (2): 107-111.

CARDON, L. R. e BELL, J. I. (2001) Association study designs for complex diseases. Nat Rev Genet. 2 (2): 91-99.

CARDON, L. R. e ABECASIS, G. R. (2003) Using haplotype blocks to map human complex trait loci. Trends Genet. 19 (3): 135-140.

CHANG, H.M. e SARKANEN, K.V. (1973) Techn. Assoc. Pulp Pap. Ind. 56:132-143

CHAPPLE, C. C., VOGT, T., ELLIS, B. E. e SOMERVILLE, C. R. (1992) An Arabidopsis mutant defective in the general phenylpropanoid pathway. Plant Cell. 4 (11): 1413-1424.

CHING, A., CALDWELL, K. S., JUNG, M., DOLAN, M., SMITH, O. S., TINGEY, S., MORGANTE, M. e RAFALSKI, A. J. (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3 (1): 19.

COSTA, M. A., COLLINS, R. E., ANTEROLA, A. M., COCHRANE, F. C., DAVIN, L. B. e LEWIS, N. G. (2003) An in silico assessment of gene function and organization of the phenylpropanoid pathway metabolic networks in Arabidopsis thaliana and limitations thereof. Phytochemistry. 64 (6): 1097-1112.

COX, D. G. e CANZIAN, F. (2001) Genotype transposer: automated genotype manipulation for linkage disequilibrium analysis. Bioinformatics. 17 (8): 738-739.

CRAWFORD, D. C., AKEY, D. T. e NICKERSON, D. A. (2005) The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet. 6: 287-312.

DA ROSA, C.A.B., CARDOSO, G.V., FOELKEL, C.E.B., FRIZZO, S.M.B., DE ASSIS, T.F., e OLIVEIRA, P. (2002) Comportamento da madeira de Eucalyptus globulus com diferentes teores de lignina para produção de celulose kraft. 35° Congresso e Exposição anual de Celulose e Papel. São Paulo, Brasil.

DEAN, J.F.D. (2005) Synthesis of Lignin in Transgenic and Mutant Plants. Biotechnology of Biopolymers. From Synthesis to Patents. Weinheim. P. 26.

DEVLIN, B. e RISCH, N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics. 29 (2): 311-322.

DING, K., ZHOU, K., HE, F. e SHEN, Y. (2003) LDA--a java-based linkage disequilibrium analyzer. Bioinformatics. 19 (16): 2147-2148.

ELKIND, Y., EDWARDS, R., MAVANDAD, M., HEDRICK, S. A., RIBAK, O., DIXON, R. A. e LAMB, C. J. (1990) Abnormal plant development and down-regulation of phenylpropanoid biosynthesis in transgenic tobacco containing a heterologous phenylalanine ammonia-lyase gene. Proc Natl Acad Sci U S A. 87 (22): 9057-9061.

EVANS, D. M. e CARDON, L. R. (2005) A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. Am J Hum Genet. 76 (4): 681-687.

EWING, B., HILLIER, L., WENDL, M. C. e GREEN, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8 (3): 175-185.

EWING, B. e GREEN, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8 (3): 186-194.

FELSENSTEIN, J., (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 5: 164-166.

FERREIRA, M.E. e GRATTAPAGLIA, D. (1998) Introdução ao Uso de Marcadores Moleculares em Análise Genética. 3ª ed. Brasília: EMBRAPA-Cenargen. p. 220.

FLINT-GARCIA, S. A., THORNSBERRY, J. M. e Buckler, E. S. T. (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. 54: 357-374.

GARCIA-GIL, M. R., MIKKONEN, M. e SAVOLAINEN, O. (2003) Nucleotide diversity at two phytochrome loci along a latitudinal cline in Pinus sylvestris. Mol Ecol. 12 (5): 1195-1206.

GION, J., RECH P., GRIMA-PETTENATI, J., VERHAEGEN, D. e PLOMION, C. (2000) Mapping candidate genes in Eucalyptus with emphasis on lignification genes. Molecular Breeding. 6: 441-449.

GOICOECHEA, M., LACOMBE, E., LEGAY, S., MIHALJEVIC, S., RECH, P., JAUNEAU, A., LAPIERRE, C., POLLET, B., VERHAEGEN, D., CHAUBET-GIGOT, N. e GRIMA-PETTENATI, J. (2005) EgMYB2, a new transcriptional activator from Eucalyptus xylem, regulates secondary cell wall formation and lignin biosynthesis. Plant J. 43 (4): 553-567.

GRATTAPAGLIA, D.(2004) Genômica Florestal. Ed. Mir, L. Editora Atheneu, capítulo 46. pp. 917-934.

GREEN, E. D. (2001) Strategies for the systematic sequencing of complex genomes. Nat Rev Genet. 2 (8): 573-583.

GRIFFITHS, A.J.F., GELBART, W.M., MILLER, J.H., e LEWONTIN, R.C. Genética moderna. Rio de Janeiro: Guanabara Koogan, 2001. p. 589.

HALPIN, C., KNIGHT, M.E., PETTENATI, J.G., GOFFNER, D., BOUDET, A., e SCHUCH, A. (1992) purification and characterization of cinnamyl alcohol dehydrogenase from tobacco stems. Plant Physiol. 98: 12-16.

HALPIN, C., BARAKATE, A., ASKARI, B. M., ABBOTT, J. C. e RYAN, M. D. (2001) Enabling technologies for manipulating multiple genes on complex pathways. Plant Mol Biol. 47 (1-2): 295-310.

HEDRICK, P. W. (1987) Gametic disequilibrium measures: proceed with caution. Genetics. 117 (2): 331-341.

HIRSCHHORN, J. N. e DALY, M. J. (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 6 (2): 95-108.

HU, W. J., HARDING, S. A., LUNG, J., POPKO, J. L., RALPH, J., STOKKE, D. D., TSAI, C. J. e CHIANG, V. L. (1999) Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nat Biotechnol. 17 (8): 808-812.

HUANG, X. e MADAN, A. (1999) CAP3: A DNA sequence assembly program. Genome Res. 9 (9): 868-877.

HUTTLEY, G. A., SMITH, M. W., CARRINGTON, M. e O'BRIEN, S. J. (1999) A scan for linkage disequilibrium across the human genome. Genetics. 152 (4): 1711-1722.

INGVARSSON, P. K. (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (Populus tremula L., Salicaceae). Genetics. 169 (2): 945-953.

ITO, T., CHIKU, S., INOUE, E., TOMITA, M., MORISAKI, T., MORISAKI, H. e KAMATANI, N. (2003) Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. Am J Hum Genet. 72 (2): 384-398.

JANSSEN, P., AUDIT, B., CASES, I., DARZENTAS, N., GOLDOVSKY, L., KUNIN, V., LOPEZ-BIGAS, N., PEREGRIN-ALVAREZ, J. M., PEREIRA-LEAL, J. B., TSOKA, S. e OUZOUNIS, C. A. (2003) Beyond 100 genomes. Genome Biol. 4 (5): 402.

JARVINEN, P., LEMMETYINEN, J., SAVOLAINEN, O. e SOPANEN, T. (2003) DNA sequence variation in BpMADS2 gene in two populations of Betula pendula. Mol Ecol. 12 (2): 369-384.

JUNG, H. J. G. e NI, W. (1998) Lignification of plant cell walls: impact of genetic manipulation. Proc Natl Acad Sci U S A. 95 (22): 12742-12743.

KAO, Y. Y., HARDING, S. A. e TSAI, C. J. (2002) Differential expression of two distinct phenylalanine ammonia-lyase genes in condensed tannin-accumulating and lignifying cells of quaking aspen. Plant Physiol. 130 (2): 796-807.

KIM, S. J., KIM, M. R., BEDGAR, D. L., MOINUDDIN, S. G., CARDENAS, C. L., DAVIN, L. B., KANG, C. e LEWIS, N. G. (2004) Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in Arabidopsis. Proc Natl Acad Sci U S A. 101 (6): 1455-1460.

KIM, S. K., ZHANG, K. e SUN, F. (2004) A comparison of different strategies for computing confidence intervals of the linkage disequilibrium measure D'. Pac Symp Biocomput. 128-139.

KOSKI, L. B. e GOLDING, G. B. (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol. 52 (6): 540-542.

LAPIERRE, C., POLLET, B., PETIT-CONIL, M., TOVAL, G., ROMERO, J., PILATE, G., LEPLE, J. C., BOERJAN, W., FERRET, V. V., DE NADAI, V. e JOUANIN, L. (1999) Structural alterations of lignins in transgenic poplars with depressed cinnamyl alcohol dehydrogenase or caffeic acid O-methyltransferase activity have an opposite impact on the efficiency of industrial kraft pulping. Plant Physiol. 119 (1): 153-164.

LEE, D., MEYER, K., CHAPPLE, C. e DOUGLAS, C. J. (1997) Antisense suppression of 4-coumarate:coenzyme A ligase activity in Arabidopsis leads to altered lignin subunit composition. Plant Cell. 9 (11): 1985-1998.

LERCHER, M. J. e HURST, L. D. (2002) Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? Gene. 300 (1-2): 53-58.

LEWIS, R. (2002) SNPs as windows on evolution. The Scientist 16 (1): 16.

LEWONTIN, R. C. (1964) The Interaction of Selection and Linkage. Ii. Optimum Models. Genetics. 50: 757-782.

LEWONTIN, R. C. (1988) On measures of gametic disequilibrium. Genetics. 120 (3): 849-852.

LI, L., ZHOU, Y., CHENG, X., SUN, J., MARITA, J. M., RALPH, J. e CHIANG, V. L. (2003) Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. Proc Natl Acad Sci U S A. 100 (8): 4939-4944.

LI, Y., KAJITA, S., KAWAI, S., KATAYAMA, Y. e MOROHOSHI, N. (2003) Down-regulation of an anionic peroxidase in transgenic aspen and its effect on lignin characteristics. J Plant Res. 116 (3): 175-182.

LONG, J. R., ZHAO, L. J., LIU, P. Y., LU, Y., DVORNYK, V., SHEN, H., LIU, Y. J., ZHANG, Y. Y., XIONG, D. H., XIAO, P. e DENG, H. W. (2004) Patterns of linkage disequilibrium and haplotype distribution in disease candidate genes. BMC Genet. 5 (1): 11.

LONJOU, C., ZHANG, W., COLLINS, A., TAPPER, W. J., ELAHI, E., MANIATIS, N. e MORTON, N. E. (2003) Linkage disequilibrium in human populations. Proc Natl Acad Sci U S A. 100 (10): 6069-6074.

MACKAY, T. F. (2001) The genetic architecture of quantitative traits. Annu Rev Genet. 35: 303-339.

MANIATIS, N., MORTON, N. E., GIBSON, J., XU, C. F., HOSKING, L. K. e COLLINS, A. (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. Hum Mol Genet. 14 (1): 145-153.

MAURICIO, R. (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nat Rev Genet. 2 (5): 370-381.

MIR, K. U. e SOUTHERN, E. M. (2000) Sequence variation in genes and genomic DNA: methods for large-scale analysis. Annu Rev Genomics Hum Genet. 1: 329-360.

MORA, A.L. e GARCIA, C.H. (2000). A Cultura do Eucalipto no Brasil. Sociedade Brasileira de Silvicultura. p.112.

MORGANTE, M. e SALAMINI, F. (2003) From plant genomics to breeding practice. Curr Opin Biotechnol. 14 (2): 214-219.

NORDBORG, M. e TAVARE, S. (2002) Linkage disequilibrium: what history has to tell us. Trends Genet. 18 (2): 83-90.

OHL, S., HEDRICK, S. A., CHORY, J. e LAMB, C. J. (1990) Functional properties of a phenylalanine ammonia-lyase promoter from Arabidopsis. Plant Cell. 2 (9): 837-848.

PERTSEMLIDIS, A. e FONDON, J. W., 3rd (2001) Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol. 2 (10): REVIEWS2002.

PILATE, G., GUINEY, E., HOLT, K., PETIT-CONIL, M., LAPIERRE, C., LEPLE, J. C., POLLET, B., MILA, I., WEBSTER, E. A., MARSTORP, H. G., HOPKINS, D. W., JOUANIN, L., BOERJAN, W., SCHUCH, W., CORNU, D. e HALPIN, C. (2002) Field and pulping performances of transgenic trees with altered lignification. Nat Biotechnol. 20 (6): 607-612.

POKE, F.S., VAILLANCOURT, R.E., ELLIOTT, R.C., & REID, J.B. (2003) Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (ccr) and cinnamyl alcohol dehydrogenase 2 (cad2). Molecular Breeding 12: 107-118.

PRITCHARD, J. K. e PRZEWORSKI, M. (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet. 69 (1): 1-14.

RAES, J., ROHDE, A., CHRISTENSEN, J. H., VAN DE PEER, Y. e BOERJAN, W. (2003) Genome-wide characterization of the lignification toolbox in Arabidopsis. Plant Physiol. 133 (3): 1051-1071.

RAFALSKI, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol. 5 (2): 94-100.

RALPH, J., HATFIELD, R. D., PIQUEMAL, J., YAHIAOUI, N., PEAN, M., LAPIERRE, C. e BOUDET, A. M. (1998) NMR characterization of altered lignins extracted from tobacco plants down-regulated for lignification enzymes cinnamylalcohol dehydrogenase and cinnamoyl-CoA reductase. Proc Natl Acad Sci U S A. 95 (22): 12803-12808.

REMINGTON, D. L., THORNSBERRY, J. M., MATSUOKA, Y., WILSON, L. M., WHITT, S. R., DOEBLEY, J., KRESOVICH, S., GOODMAN, M. M. e Buckler, E. S. T. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A. 98 (20): 11479-11484.

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

ROZAS, J. e ROZAS, R. (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics. 15 (2): 174-175.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M. A. e BARRELL, B. (2000) Artemis: sequence visualization and annotation. Bioinformatics. 16 (10): 944-945.

SAIER, M. H., JR. (1998) Genome sequencing and informatics: new tools for biochemical discoveries. Plant Physiol. 117 (4): 1129-1133.

SANGER, F., NICKLEN, S. e COULSON, A. R. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74 (12): 5463-5467.

SEDEROFF, R. (1999) Building better trees with antisense. Nat Biotechnol. 17 (8): 750-751.

SIBOUT, R., EUDES, A., POLLET, B., GOUJON, T., MILA, I., GRANIER, F., SEGUIN, A., LAPIERRE, C. e JOUANIN, L. (2003) Expression pattern of two paralogs encoding cinnamyl alcohol dehydrogenases in Arabidopsis. Isolation and characterization of the corresponding mutants. Plant Physiol. 132 (2): 848-860.

STEPHENS, M., SMITH, N. J. e DONNELLY, P. (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68 (4): 978-989.

STEPHENS, M. e DONNELLY, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet. 73 (5): 1162-1169.

STERKY, F., REGAN, S., KARLSSON, J., HERTZBERG, M., ROHDE, A., HOLMBERG, A., AMINI, B., BHALERAO, R., LARSSON, M., VILLARROEL, R., VAN MONTAGU, M., SANDBERG, G., OLSSON, O., TEERI, T. T., BOERJAN, W., GUSTAFSSON, P., UHLEN, M., SUNDBERG, B. e LUNDEBERG, J. (1998) Gene

discovery in the wood-forming tissues of poplar: analysis of 5, 692 expressed sequence tags. Proc Natl Acad Sci U S A. 95 (22): 13330-13335.

SUBRAMANIAM, R., REINOLD, S., MOLITOR, E. K. e DOUGLAS, C. J. (1993) Structure, inheritance, and expression of hybrid poplar (Populus trichocarpa x Populus deltoides) phenylalanine ammonia-lyase genes. Plant Physiol. 102 (1): 71-83.

TARAZONA-SANTOS, E. e TISHKOFF, S. A. (2005) Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus. Genes Immun. 6 (1): 53-65.

THAMARUS, K. A., GROOM, K., MURRELL, J., BYRNE, M. e MORAN, G. F. (2002) A genetic linkage map for Eucalyptus globulus with candidate loci for wood, fibre, and floral traits. Theor Appl Genet. 104 (2-3): 379-387.

THE INTERNATIONAL HAPMAP PROJECT (2003) The International HapMap Project. Nature. 426 (6968): 789-796.

THORNSBERRY, J. M., GOODMAN, M. M., DOEBLEY, J., KRESOVICH, S., NIELSEN, D. e Buckler, E. S. T. (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet. 28 (3): 286-289.

THUMMA, B. R., NOLAN, M. F., EVANS, R. e MORAN, G. F. (2005) Polymorphisms in Cinnamoyl CoA Reductase (CCR) are Associated with Variation in Microfibril Angle in Eucalyptus spp. Genetics.

WALL, J. D. e PRITCHARD, J. K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet. 4 (8): 587-597.

WEISS, K. M. e CLARK, A. G. (2002) Linkage disequilibrium and the mapping of complex human traits. Trends Genet. 18 (1): 19-24.

YOSHIDA, K., KAMIYA, T., KAWABE, A. e MIYASHITA, N. T. (2003) DNA polymorphism at the ACAULIS5 locus of the wild plant Arabidopsis thaliana. Genes Genet Syst. 78 (1): 11-21.

YU, N., JENSEN-SEAMAN, M. I., CHEMNICK, L., RYDER, O. e LI, W. H. (2004) Nucleotide diversity in gorillas. Genetics. 166 (3): 1375-1383.

ZONDERVAN, K. T. e CARDON, L. R. (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet. 5 (2): 89-100.

ZUBIETA, C., KOTA, P., FERRER, J. L., DIXON, R. A. e NOEL, J. P. (2002) Structural basis for the modulation of lignin monomer methylation by caffeic acid/5-hydroxyferulic acid 3/5-O-methyltransferase. Plant Cell. 14 (6): 1265-1277.

Aracruz Celulose. Disponível em: http://www.aracruz.com.br/port/euca_historia.htm Acessado em 14.09.2004

Banco de Dados de Biomassa no Brasil. Disponível em: http://infoener.iee.usp.br/scripts/biomassa/br_carvao.asp Acessado em 29.07.2005.

Banco de Dados de Biomassa no Brasil. Disponível em: http://infoener.iee.usp.br/scripts/biomassa/br_lenha.asp Acessado em 29.07.2005.

EMBRAPA – Cenargen – Plataforma de seqüenciamento de DNA. Disponível em: http://www.cenargen.embrapa.br/laboratorios/psd/psd.html Acessado em 08.07.2004 e 02.08.2005.

Invitrogen: Disponível em: http://www.invitrogen.com/content.cfm?pageid=94 Acessado em 17.09.2004.

Rozen S., & Skaletsky H.J. (1996) Primer3: A software component for picking PCR primers. Código fonte disponível em: http://www-genome.wi.mit.edu/genome_software/other/primer3.html Programa online disponível em: http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi

9 APÊNDICE

9.1 Sequência completa do gene *CAD2* em *Eucalyptus grandis*.

>CAD GR Full Lenght

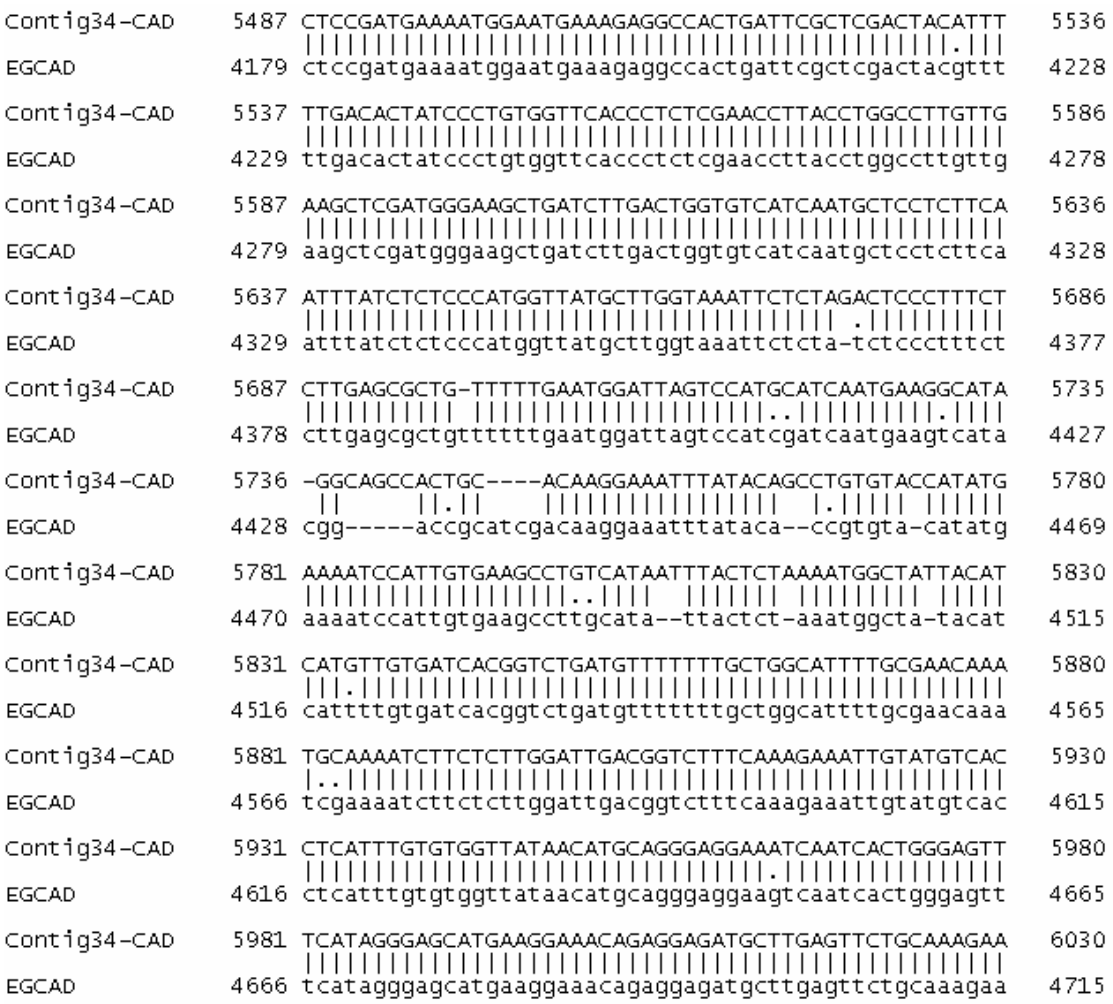
aaccctcactaaagggactagtcctgcaggtttaaacgaattcgcccttaaactacacaacacgaatagttcaaaattgacatga
attgagatgattttacttattctaataaaaaatatttgtaaacatgtttaaacattttgatgatggtacaagaaactgacaagaatacg
atgatgcgaaaaaaattgtcacccctacttccaattgctatcgctacttgtacatagaatatgtagtgtccaaagaaactttaagtct
atcctctctcctcccaaccccttatgtttctatcactagaagcataggaacagaaaacagactaaacatatcgacccatgtttttaa
ctgacctatttcaccctaaattcctcaatggtttcactttgcatagacactcaagaacatagtagaagaacaatgtgaacagagaa
cgatcacaacaaaagaagaatgatgagaatagaaaaggatcaaaacaaaggatttgttttaagagaaacctcacatgcgaga
atcaggatcttcacgaggatcatatcctttccttatccagaacttcacaaatggcccattattgaaatagtatgccacatttgtgaga
aggctacaagacctgacatattaaaggaaattatttgacgttgatctgacagaagaataatcaagattatcgcacactagaatttg
aagacacatctcgaaataagaagaatgctgactacctttgagcaaaagaactccaaagtccagacctttatctagcaatctttga
taaagagaattcttgggccatatgggccgctcctcaaacaatttagacacagccatttgcctctcccactcatctgatccttggatt
agtcgctcctcccaatttattatctttgggatctgtatggatgattaagtgaaacagatatagaacaaacataatctgcttgctgcgt
atgaccatgttaagtataactaagaaagaagcacaagtaatttggcatactgagatgggcaggtcatgttcaagactcattaaag
attacgagattgccacatcactcagtataccaagtctggaaaccatcttagtctacccaatcaccagcatcattcaactaaaaattt
aggatatctttgaccttgaaatcaagtgccagactgttctccacaaaaacctgcaaaagaaggaatgcacgctccttaagattgtt
gtacagatataattcttttacttactgtcttcccaaatgaatcttttcagatggtattagtgagattaggcctccaaagtaaacctcaa
catgctcagcagctgtatcctgtttccttttaaatttgaagtggccgggggacgacttaacctgaaaaaatggtgacaatgtttaa
aaaaggaggatggctcttgatgattttctacttatccatagaatataacgaatcattctcaaagtctaaattgtccctgtaagaattta
cagttcctcatttttgttattaattatcatttcttatcgccatttagtatgggccatccttcaaagcatgaactcatgctcccacatgtcc
ttgttgccctcttgtctcttatttctttcattacctcataatattaaccactttacttctattactgcacttcattcaggggcaatgacaac
cgtaatgcctacactggatgcgattccaactctatctaattggttttgaatgtcaaacattaagtggtttttccaactacacttatgatt
agatttaggcatcaattaaacccagtctacaacttcattaagcaaatgaattcttcttctgctatatttatgtgattatgtgtactttgct
atgatatttccagttgaattgatagaaagtaccactatcaatcactcaaaggcatgattcaaagagccaaaaaaagggataaatt
gagagccaattccaaatatagcatcataaagaagtaggtcaatcagtagatatgacagaaaaggagcagcgaatcgtacgcta
aagttgttgggacatccttggatgagaagagtgggggaacgatcatcattacatcctcactttcgacacttgtaagagcatccttt
gctatatagaccaccgtggaaaggaaaatgagaaaattatggatagtaatattaaaggaaaaaagttatcctattgacacgatag
gtgatattagaaacgtgagaaagattttaccaaagggaggcgcctccacttctccagttttttttcttttttccttcgagcattatcagc
atgaacagcaacaacatgttgataatcagccatccctttttcaaaatttgttggcaactactagttataaaatcatccatcaagtacc
ctagtctatataatagtttgtaaagaatgaattgacaaaaggatgcaaatgagtttcgattagcttctaaccttgaaaaacatatgcc
tcggagacacgagcagcgatatcagcacatagttgcgtttgatcctcttctttattggcaccagaagaactagtgcattgatctcc
ttccagactcattccagtattgctgacagcactatgatcttgagtaagtgaagggttagcacgtatcttgctccttttgatattctca
acaggaaattattgcatcggcgaagctccccaaaagtaggatggcaaaatgggtcttcaggccggaaacgaagttccattttgt
ttgaactggagctgcgtgcctgcaaagttcaagcttccagttgacatatatggccccagaaacctcaaaatgaccctactaaga
atgcaagtgattgaggcaacaccttcaaaatcccttcacttccgccaagagtctcgatagcccgtgaggcagaggaaggataa
cccggatagtggaccacaaatgtcttagtgcttggcaaacatccagaaatggctccatcctttatttctaccattgtttgctgagta
gagtaactaggaaaatagaatgtaataagagaagatatcagagagaaaaagagagaaactctacccatattccaactagctta
aaaatcaatctctaatttcagggactagagtagaggtttcagctgaagtatacaactgcacagacttacgcataggaattaattat
ctaaagcgaatcagcacctcgttgaggaatttcaccgaacaggtggtgggcgatagaggaagacagacaacgcacagaga
ggcaataagcatggaggaatcccaggcaactgacattttagtttttctcatacatgctgtatattaaattggaactttatcctgtacat
ctcaacatccattcatgccaacctacgtgaaaacaaaacaaaaaaaaggaataaaacccaatgccaagcagactgaaaccaa
tctcaaaggcgcaagcttttctgaactcacgatggttccagaaaggcgatgttttgctgacagagggagcgttcgatggagctt
ctccatcacttaatttgtcccttcaagatgaaaaaagtaagaggtccaccgtaccaaaacattcttccactcagaagaaaaccac
agtagctggagggagtcaagcatgtcagaagcacagaaactgggaatggctaaaaagcaagtcttgacccttaacccacccc

actggttcacctaccgcacctcgggttaggtattgcttgctgaggtgtcacttttcgccaaagtcacgtctctcttttggattcttcct
attggtccgtctcgtttcctcgttgctggttgctggtagcgtttttgtccatatatatatggcagtccatatgtatccccgtcactcctc
atctatgctcctacccggcaacttcccactacgataagcagcaagtctacggctctgtcgaatctctctccgagcaccactttga
aaaaagcttggatctttgagcaaaaatgggcagtcttgagaaggagaggactaccacgggttgggctgcaagggacccgtct
ggcgttctctctccttacacttatagcctcaggtagattcaagaacttgccttcttcaggattgataaagatagctaagaatctaagt
tttcgttgtgcttgtgatgtcgttctttaattcttgtttttgcttgttcgatcaattacgtattaatcaatattcgattgattagtctgagtttat
cggcaaaaaagatttgtctaagtcacttcccaacaaatgcagaaacacgggaccagaagatctttacatcaaggtgttgagctg
cggaatttgccacagtgacattcaccagatcaagaatgatcttggcatgtcccactaccctatggttcctgggtaggtcttttcttg
cattaatcatgactgattcttcctcttctgtgtctcttcctattctaattattctttcccctctttttcaggcatgaagtggtgggtgaggtt
ctggaggtgggatcagaggtgacaaagtacagagttggtgaccgagtgggaaccggtatagtggttgggtgctgcagaagc
tgtggcccttgcaattcggaccaggagcaatactgcaacaagaagatttggaattacaatgacgtgtacaccgatggcaagcc
cactcaaggtgggtttgctggtgagatagtggttggccaaaggtgagcacttgcatgcctctctctactctcttattctaagctgaa
atagcctacctaacgggggagagttgatgatccctcgaggagaccttgcctgtgtggaattttgccattggatggaaaagcaaa
atatgaaaaactgtccgggttgaggtctgttgtatttcaatccaatggtgaaagaatgcgagggcaccatgccccccatctactt
aaagcagatcgaaaagtgggcatcctttgtttgagcccttccattgttgattcattcaatttcagttaatttgtcttgggtcaatcgta
atggcccacaaaatcacattcaacttgatgaacaataatcagaaagaaacatgtctggtcttgtcaattttagctcttaagtacttc
gaatggcatgacttgatttgatggtgtattgtaaatttagtaaactagttggcgaacttgaaggtcctgggaattagttaagtggaa
tgatcataggaaagacacatgactagcctaagctaacttctacagcactgcatgtaaacccgtacaaagtgaaactgatggaag
atttgactggtccgattggggtctttgcccttcaggtttgtggtgaaaatcccagatgggttagagtcggaacaggcagcgccg
ctgatgtgcgctggtgtgaccgtgtacagccctctggtgcgctttgggctcaagcaaagtgggctgagaggagggatattggg
gcttggaggggttggccacatgggggtgaagatagccaaggccatgggacaccatgtgactgtgataagctcttctgataag
aagagaacggaggcattggagcacctgggtgccgatgcttacctagtgagctccgatgaaaatggaatgaaagaggccact
gattcgctcgactacatttttgacactatccctgtggttcaccctctcgaaccttacctggccttgttgaagctcgatgggaagctg
atcttgactggtgtcatcaatgctcctcttcaatttatctctcccatggttatgcttggtaaattctctagactccctttctcttgagcgc
tgtttttgaatggattagtccatgcatcaatgaaggcataggcagccactgcacaaggaaatttatacagcctgtgtaccatatga
aaatccattgtgaagcctgtcataatttactctaaaatggctattacatcatgttgtgatcacggtctgatgttttttgctggcattttg
cgaacaaatgcaaaatcttctcttggattgacggtctttcaaagaaattgtatgtcacctcatttgtgtggttataacatgcaggga
ggaaatcaatcactgggagtttcatagggagcatgaaggaaacagaggagatgcttgagttctgcaaagaaaaggattgac
ttcccagatcgaagtgatcaagatggattatgtcaacacggccctagagaggctcgagaagaatgatgtcaggtacaggttcg
tcgtggacgttgtgggaagcaagcttgattagttttgtcctttccccataattaaacaagaaatcgacgtgcttgtctctcaattcga
gttcctcatgccctctgttgtatcattgtttgttatgccgagagtattattttcttctgtcttcgtattgaaaccatagaccttctcgattgt
gtattcaatgatgaaggtgttaatgattttatcacttaagaaatttttgactatttggattctggaagcattttgaattgggttgtgctgt
gttttcaagaggggtgggtgagggttctctttttgacagtgacccaacaacaaactcggatgaataaaagtgacacaatgtgttt
cacaattcttggtggatacgactctttgatagggtgctcaccagcttttcctaagttagcaaaggtcaacatcgctttcacatggag
ccattatcccaagtgagcacaaagtaagataaaagagtagagggtggagactaggatgcttttctgcctaagtagcgtcgaaa
atgtcagcagacagattaacctgagtcgtcaaatatacgagccatcacgaagatggttttgggaaagatgggggatttatgagc
tcgctgagcttccgttcgatcaaaggaaacatggtgggtactccggagagatgcaacagcgaccgctcagctcgatctttctgt
ggtttccagctgatcctagaaacaagggcaaccactgtctcaatctttctgcggttcctaatcgatcccagcacaaatacgcgta
ggaagacacgaaccttgcccgcaatcgagacacggtttgatgcgcattctggaagcggctcttccgacacacgaggcggaa
gtcagaagagagggaaaatgatttttttcaggagagcgagacctcaagcaaaggatacgacgtgtggatgacttcaccccaac
gtttgctaatgtccgggtccacccgctggccaccctcgtcaactctctgccttccaaacatcctcggcatcatcctgcaccacca
aaacccacgactgtgacaaggcacgcccccgagactttcttgctctccctttatcggagagggaagaaaagtcgggcggctt
gtacgcggacctcatcaaacgaagtttgagttggactgcggtgactgcgaaattgaaggtcacgcccaactgaaagcgacct
ccttgaagacgacaaatggatctgctcggttcgtataccttcaaaatgaggtgtcatcttatcttttgagataagcaagcatgcc
gcttcatgtaacgacagacatcgcaccttccacccgggaataatcttgtgttcttcagggattatcatatgcaagtgaaatcgcc
ctgtttctcttctcgttctcgcccctcgaccgattattttactgattcctactcagaacatgcaagtgaaactgccaatttttcacaaaa
ggaatgccatgtatcgcatctcgcccttgttgag

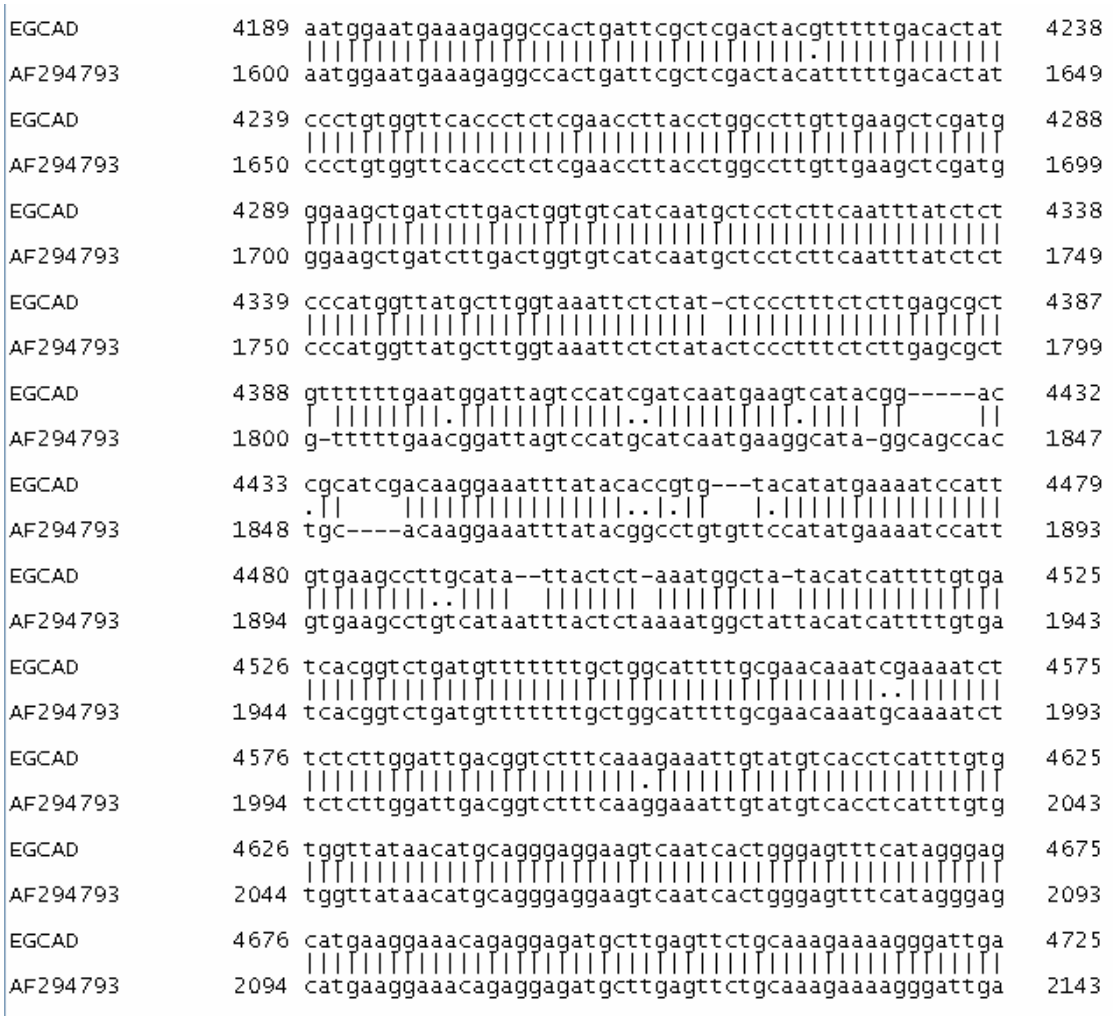
9.2 Alinhamento de *E. grandis* com *E. saligna* (*CAD2*) demonstranto os 99,2% de identidade.

```
Contig34-CAD    5397 CATGGGACACCATGTGACTGTGATAAGCTCTTCTGATAAGAAGAGAACGG   5446
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
AF294793        1500 catgggacaccatgtgactgtgataagctcttctgataagaagagaacgg   1549

Contig34-CAD    5447 AGGCATTGGAGCACCTGGGTGCCGATGCTTACCTAGTGAGCTCCGATGAA   5496
                     ||||||||||||||||||..||||||||||||||||||||||||||||||
AF294793        1550 aggcattggagcacctgggtcgcgatgcttacctagtgagctccgatgaa   1599

Contig34-CAD    5497 AATGGAATGAAAGAGGCCACTGATTCGCTCGACTACATTTTTGACACTAT   5546
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
AF294793        1600 aatggaatgaaagaggccactgattcgctcgactacatttttgacactat   1649

Contig34-CAD    5547 CCCTGTGGTTCACCCTCTCGAACCTTACCTGGCCTTGTTGAAGCTCGATG   5596
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
AF294793        1650 ccctgtggttcaccctctcgaaccttacctggccttgttgaagctcgatg   1699

Contig34-CAD    5597 GGAAGCTGATCTTGACTGGTGTCATCAATGCTCCTCTTCAATTTATCTCT   5646
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
AF294793        1700 ggaagctgatcttgactggtgtcatcaatgctcctcttcaatttatctct   1749

Contig34-CAD    5647 CCCATGGTTATGCTTGGTAAATTCTCTAGACTCCCTTTCTCTTGAGCGCT   5696
                     ||||||||||||||||||||||||||||.||||||||||||||||||||||
AF294793        1750 cccatggttatgcttggtaaattctctatactccctttctcttgagcgct   1799

Contig34-CAD    5697 GTTTTTGAATGGATTAGTCCATGCATCAATGAAGGCATAGGCAGCCACTG   5746
                     |||||||||.||||||||||||||||||||||||||||||||||||||||||
AF294793        1800 gtttttgaacggattagtccatgcatcaatgaaggcataggcagccactg   1849

Contig34-CAD    5747 CACAAGGAAATTTATACAGCCTGTGTACCATATGAAAATCCATTGTGAAG   5796
                     ||||||||||||||||.||||||||.|||||||||||||||||||||||||
AF294793        1850 cacaaggaaatttatacggcctgtgttccatatgaaaatccattgtgaag   1899

Contig34-CAD    5797 CCTGTCATAATTTACTCTAAAATGGCTATTACATCATGTTGTGATCACGG   5846
                     ||||||||||||||||||||||||||||||||||||.||||||||||||||
AF294793        1900 cctgtcataatttactctaaaatggctattacatcattttgtgatcacgg   1949

Contig34-CAD    5847 TCTGATGTTTTTTTGCTGGCATTTTGCGAACAAATGCAAAATCTTCTCTT   5896
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
AF294793        1950 tctgatgttttttgctggcattttgcgaacaaatgcaaaatcttctctt   1999

Contig34-CAD    5897 GGATTGACGGTCTTTCAAAGAAATTGTATGTCACCTCATTTGTGTGGTTA   5946
                     ||||||||||||||||||.|||||||||||||||||||||||||||||||||
AF294793        2000 ggattgacggtctttcaaggaaattgtatgtcacctcatttgtgtggtta   2049
```

9.3 Alinhamento entre *E. grandis* e *E. gunnii CAD1* identificando as indels.



9.4 Alinhamento entre *E. saligna* e *E. gunni* demonstrando novamente a presença das indels.



9.5 Análise de BLAST do clone genômico de *CAD2* contra o banco de contigs do Genolyptus

```
BLASTN 2.2.10 [Oct-19-2004]


   Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
   Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
   "Gapped BLAST and PSI-BLAST: a new generation of protein database search
   programs",  Nucleic Acids Res. 25:3389-3402.

   Query= Contig34-CAD
            (8106 letters)

   Database: genolyptus
              40,749 sequences; 24,131,485 total letters

                                                                      Score    E
   Sequences producing significant alignments:                      (bits) Value
   CL12Contig1&&109711987743227716958834582163                        870   0.0
   CL12Contig2&&109711987763510717070069379579                        823   0.0
   >CL12Contig1&&109711987743227716958834582163
             Length = 1533

    Score =  870 bits (439), Expect = 0.0
    Identities = 442/443 (99%)
    Strand = Plus / Plus


   Query: 5221 aggtttgtggtgaaaatcccagatgggttagagtcggaacaggcagcgccgctgatgtgc 5280
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Sbjct: 582  aggtttgtggtgaaaatcccagatgggttagagtcggaacaggcagcgccgctgatgtgc 641


   Query: 5281 gctggtgtgaccgtgtacagccctctggtgcgctttgggctcaagcaaagtgggctgaga 5340
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Sbjct: 642  gctggtgtgaccgtgtacagccctctggtgcgctttgggctcaagcaaagtgggctgaga 701


   Query: 5341 ggagggatattggggcttggaggggttggccacatgggggtgaagatagccaaggccatg 5400
               |||||||||||||||||||||||||||||||||||| ||||||||||||||||||||
   Sbjct: 702  ggagggatattggggcttggaggggttggccacatgggcgtgaagatagccaaggccatg 761


   Query: 5401 ggacaccatgtgactgtgataagctcttctgataagaagagaacggaggcattggagcac 5460
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Sbjct: 762  ggacaccatgtgactgtgataagctcttctgataagaagagaacggaggcattggagcac 821


   Query: 5461 ctgggtgccgatgcttacctagtgagctccgatgaaaatggaatgaaagaggccactgat 5520
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Sbjct: 822  ctgggtgccgatgcttacctagtgagctccgatgaaaatggaatgaaagaggccactgat 881


   Query: 5521 tcgctcgactacatttttgacactatccctgtggttcaccctctcgaaccttacctggcc 5580
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Sbjct: 882  tcgctcgactacatttttgacactatccctgtggttcaccctctcgaaccttacctggcc 941


   Query: 5581 ttgttgaagctcgatgggaagctgatcttgactggtgtcatcaatgctcctcttcaattt 5640
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Sbjct: 942  ttgttgaagctcgatgggaagctgatcttgactggtgtcatcaatgctcctcttcaattt 1001


   Query: 5641 atctctcccatggttatgcttgg 5663
               |||||||||||||||||||||||
   Sbjct: 1002 atctctcccatggttatgcttgg 1024
```